Nat Methods. Author manuscript; available in PMC 2014 October 01.

Published in final edited form as:
Nat Methods. 2014 April ; 11(4): 403–406. doi:10.1038/nmeth.2841.

A mass spectrometry-based hybrid method for structural modelling of protein complexes

Argyris Politis[#1,4], **Florian Stengel**[#2], **Zoe Hall**[1], **Helena Hernández**[1], **Alexander Leitner**[2], **Thomas Walzthoeni**[2], **Carol V. Robinson**[1,6], and **Ruedi Aebersold**[2,3,6]

[1]Department of Chemistry, University of Oxford, South Parks Road, Oxford, United Kingdom. [2]Department of Biology, Institute of Molecular Systems Biology, Eidgenossische Technische Hochschule (ETH) Zurich, Zurich, Switzerland. [3]Faculty of Science, University of Zurich, Zurich, Switzerland

[#] These authors contributed equally to this work.

Abstract

We describe a method that integrates data derived from different mass spectrometric (MS) techniques with a modelling strategy for structural characterization of protein assemblies. We encoded structural data derived from native MS, bottom-up proteomics, ion mobility-MS and chemical cross-linking MS into modelling restraints to compute the most likely structure of a protein assembly. We used the method to generate near-native models for three known structures and characterized an assembly intermediate of the proteasomal base.

Cells contain macromolecular assemblies, which are composed of physically interacting proteins[1]. Elucidating the structure and dynamics of these assemblies are primary goals of structural biology.

Recently, analysis of protein complexes using hybrid methods has garnered great interest[1,2], enabling insights for systems which remain refractory to structure determination by a single method[3]. Among the methods that contribute to structural analyses, structural mass spectrometry (MS) is generally applicable and requires only small sample amounts. Different types of MS measurements can provide multiple and orthogonal datasets for a specific protein complex. Label free, quantitative bottom up analyses by liquid chromatography (LC) MS/MS defines the composition and relative abundance of the

[6]Correspondence should be addressed to: Ruedi Aebersold: aebersold@imsb.biol.ethz.ch and Carol V. Robinson: carol.robinson@chem.ox.ac.uk.

Author Contributions F.S. and A.P. conceived the study; A.P., F.S., C.V.R. and R.A. designed the research; A.P. performed all modelling and developed the software; F.S. carried out the experiments; Z.H and H.H. performed part of the IM-MS and native MS experiments. A.L. and T.W. supported CX-MS experiments and analysis; F.S. and A.P. analysed the data; A.P., F.S, C.V.R. and R.A. wrote the paper; All authors commented on and edited the final version of the paper.

[4]Current address: Department of Life and Health Sciences, School of Biomedical Sciences, University of Ulster, Londonderry, United Kingdom

Competing Financial Interests: The authors declare no competing financial interests

Europe PMC Funders Author Manuscripts

complex subunits. Native MS of intact protein complexes and their subcomplexes provides information on the overall stoichiometry and protein-protein interactions. MS coupled with ion mobility (IM), IM-MS elucidates protein architectures and dynamics by measuring their collisional cross sections (CCSs)[4,5]. Chemical cross-linking coupled with MS (CX-MS) technology identifies protein subunit interfaces[6]. While the utility of the individual techniques has been documented, combining information from all four MS-based approaches with modelling, has not been reported to our knowlege.

Here we describe a generic hybrid structural biology method that integrates orthogonal datasets for the same protein complex, generated by native MS, label free quantification (LFQ) by LC-MS/MS, (IM)-MS and CX-MS. The method differs from other approaches by its ability to generate orthogonal datasets and to computationally integrate diverse MS datasets with different levels of resolution and information content from the same sample. Overall, the method enables accurate prediction of multiprotein and heterogeneous complexes when high-resolution information of the individual subunits is used, and it consists of experimental techniques that require only low microgram sample amounts and that exhibit high measuring speed and tolerance for heterogeneous sample environments[8].

The method involves four steps: (i) protein purification and data collection by the respective MS technique. Aliquots of the purified protein complex are first analysed by (LFQ) and CX-MS experiments; and, after buffer exchange, IM-MS and native MS (Online Methods) ii) encoding MS data into restraints (iii) structure prediction by iterative sampling and scoring of models and, (iv) ensemble analysis to generate most likely structure(s) (Fig. 1a and Online Methods).

We developed and benchmarked the method using three well-characterised complexes, exhibiting distinct topologies, methane monooxygenase hydroxylase from *Methylococcus capsulatus* (MMOH), toluene/oxylene monooxygenase hydroxylase (ToMOH) from *Pseudomonas stutzeri*, and urease from *Klebsiella aerogenes* urease (Online Methods, Supplementary Note 1 and Supplementary Fig. 1). Native MS to allowed us to determine the stoichiometry of the complexes and their subunit connectivities[4,8] (Supplementary Fig. 2). IM-MS added orientationally averaged CCSs[9], and CX-MS allowed us to identify high-confidence inter- and intra-protein interactions[10-12]. Using these MS-based restraints allowed sampling of complex models. Next, we refined the models using an optimization step and ranked the models with a weighted scoring function. We selected representative structures from the pool of highly ranked models upon pairwise clustering of their alpha-carbon root-mean-square deviations (C-α RMSDs). A refinement step ensured physical interactions between subunits (Online Methods). For all complexes we found good agreement (RMSD <12 Å) of the best-scored models with their native structures (Fig. 1b,c and Supplementary Figs. 3-7).

To evaluate contributions of each restraint for predicting near-native structures we carried out statistical tests using Receiver Operating Characteristics (ROC) (Supplementary Note 2). Plotting the ROC curves and their predictive values shows that combining restraints from IM-MS and CX-MS increases (~10%) predictabilty (Fig. 1d and Supplementary Figs. 8-12). Next, we assessed the impact on predictabilty when partial or no high-resolution

structures are available. The results showed a decrease in predictability (~10%) when only homology models were used (Supplementary Table 1). If no high-resolution subunit information is available or can be computed, predictability will be substantially reduced. However, combining restraints still increased the predictive power of the method (Fig. 1e and Supplementary Table 2)).

We further assessed the individual contribution of CX-MS and IM-MS restraints to the scoring function by weighting their impact in a training set of complexes. To optimise weighting, we calculated true positives for varied degrees of input data (Online Methods). We defined a "true positive" as a model with RMSD < 12 Å from the native structure. We calculated optimal weightings of 0.64 and 0.36 (s.d. ± 0.05) for CX-MS and IM-MS restraints, respectively (Fig. 1f). We henceforth used these values for complexes with unknown structures.

Next we applied our method to a biologically important assembly, the proteasome. Our structural knowledge of the intact complex is derived from two EM maps, containing all but the smallest lid subunit (Sem1)[3,13]. By isolating the proteasomal lid using pull-downs of tagged lid subunits and subjecting aliquots to the various MS methods, we confirmed successful enrichment of the lid subunits with LFQ (Supplementary Fig. 13). Exemplary mass spectra of the intact lid and its subcomplexes are shown (Fig. 2a,b), together with corresponding CCSs derived from IM-MS (Supplementary Figs. 14-15, Supplementary Tables 2-4). We identified a total of 170 inter-links, between non-identical subunits, (28 non-redundant) within the lid[10,14] (Supplementary Tables 5-9).

Native and CX-MS data defined two distinct modules in the lid (Rpn5/6/8/9/11 and Rpn3/Rpn11/Sem1) (Fig. 2c and Supplementary Figs. 16-17). It is interesting to speculate, in light of a recently published study[15], that these two modules may function as intermediates subcomplexes en route to the formation of the lid. Using our hybrid method, we predicted models of the lid in good agreement with the corresponding EM maps[3,13] (Fig. 2d, Supplementary Fig. 18 and Supplementary Table 10). We showed a marked similarity for the best-scoring ensemble of models of the Rpn5/8/9/12 module using hierarchical clustering (Supplementary Figs. 19-20), and by overlaying them onto the corresponding density map (Fig. 2e)[13]. Interestingly, in our model we placed Sem1 in the density cleft formed between subunits Rpn3 and Rpn7 (Fig. 2c), consistent with recent studies using deletion strains of Sem1/Rpn15, EM and MS[16].

Next, we attempted to characterize assembly intermediates, which are notorious challenging targets for classical structural biology methods. Molecular and biochemical studies have shown that the proteasomal base is assembled via a multistep process where precursors are transiently associated with proteasome-dedicated chaperones or proteasome interacting proteins (PIPs). Despite some successes on smaller complexes[14,17], efforts to uncover high-resolution structures of intact assembly intermediates have failed, presumably due to their heterogeneous and transient nature[15].

The combined LFQ data from lid affinity pull-downs (Supplementary Tables 6-9 and Supplementary Fig. 21) indicated that in addition to all known 19S subunits, we detected the

(PIPs), Hsm3, Rpn14, Nas2, Ubp6 and Nas6 (PSD10) that assist assembly of the base[18,19]. To probe these PIP containing complexes, we used pull-downs from Rpn14 and Nas6 tagged cells. LFQ confirmed that the base subunits are the main interacting partners of these PIPs (Online Methods and Supplementary Fig. 22). Native MS revealed the intact Nas6/Rpt3/Rpt6/Rpn14 precursor as well as multiple stable subcomplexes thereof (Fig. 3, Supplementary Figs. 23-24). IM yielded the CCS of the Rpt3/Rpt6/Nas6 trimer, while CX-MS confirmed 4 unique high-confidence PIP-based inter cross-links (Supplementary Table 11). These data together with crystallographic information on the Nas6/Rpt3 interface[14] allowed us to confidently predict a structural ensemble of the intact Nas6/Rpt3/Rpt6/Rpn14 precursor (Fig. 3 and Supplementary Table 12).

We also detected multiple high-quality interlinks for the base ATPase hexamer (Rpt1-6), all in agreement with the proposed subunit order of subunits (Rpt1/Rpt2/Rpt6/Rpt3/Rpt4/Rpt5; Supplementary Results and Supplementary Figs. 25-26)[19]. Together with the known composition and stoichiometry of the precursors[20], this allowed us to propose a structural model for early steps in base assembly (Figure 3). We further proposed, based on LFQ and CX-MS data, the structural organization of other known intermediate precursors (Nas2/Rpt4/Rpt5, Hsm3/Rpt1/Rpt2/Rpn1 and Rpn2/Rpn13 modules) that act as building blocks for the formation of the base[15] (Fig. 3, Supplementary Figs. 27-30).

Overall, we developed, validated and applied a generic method consisting of complementary MS-based approaches and computational data integration for structural analysis of protein complexes. The computational data integration is available as Supplementary Software and its Python package documentation is described in Supplementary Note 3. Since this hybrid method can be coupled to any purification protocol, provided expression levels are μM, we anticipate it will be very useful for probing heterogeneous assemblies, especially in the 50-300 kDa range that is challenging for current EM approaches.

Online Methods

Overall Workflow

First, the protein complex of interest is purified, either by a recombinant expression system or by affinity purification and, if needed, subsequently enriched by centrifugal concentration. Then the sample is split and used for LFQ and CX-MS and after buffer exchange, for ion mobility and native MS experiments. LFQ generates a list of subunits and their relative abundance present in the sample. Native MS of the intact complexes yields the composition and stoichiometry of protein complexes while further information is attained from gas-phase dissociation techniques such as collision-induced dissociation (CID), which reveales subunit interaction networks[8]. Ion mobility coupled with MS provids us with topological information in the form of an orientationally averaged CCS[9]. Furthermore, the CCSs of stable subcomplexes can be used to reveal the structures of the building blocks of a complex. We identified multiple high-confidence inter- and intra-protein interactions by applying isotopic labelled cross-linkers and searching and validating the identified cross linked peptides against a database generated from the LFQ experiment using the xQuest and xProphet pipeline[10,11]. We used the identified cross-links as upper-bound distance restraints (35 Å) for structural modelling.

With the data encoded into spatial restraints in hand, we applied our computational strategy for structure determination of protein complexes. We first selected an appropriate representation scheme that best reflects the resolution of the available data. In order to be able to generate pseudo-atomic models, we used high-resolution information of the individual subunits. These can be X-ray crystals, NMR structures or high-confidence homology models given available templates.

We used the subunit list from LFQ to generate the structural input for the various subunits of the proteasomal assembly. For full exploitation of the cross-linking information (residue level), high-resolution structures should be available for the individual subunits within the complexes. We therefore generated homology models for all subunits for which no high-resolution structures are available. Sequence Id for the test case proteins was between 20% and 100% (Supplementary Table 1) and between 19% and 56% for the lid proteins (Supplementary Table 10), respectively. Next we set out to build a large number of structural models of protein complexes from their building blocks. A critical part of sampling is to accurately determine the stoichiometry and copy number of subunits and subcomplexes within the intact assembly. We acquired this information by combining LFQ data with the native MS data of the intact complexes and additional subcomplexes identified by CID that allowed us to build structural models consistent with the experiments. We generated model structures that satisfy the input data using a Monte Carlo search step and subsequently optimized through a conjugate gradient optimization algorithm. Next, we scored the candidate models using a weighted scoring function, which encodes the three types of restraints. We selected the representative structures from the pool of highly ranked models upon pairwise clustering (described below in detail). Finally, a flexibility step using energy minimization/molecular dynamics (MD) simulations allowed us to search for energetically favourable structures and eliminate potential steric clashes.

Protein Purification

We used a training set of three well-characterised complexes exhibiting distinct topologies to develop and optimise our method. The complexes are i) toluene/o-xylene monooxygenase hydroxylase from *Pseudomonas stutzeri*[21] (ToMOH, PDB ID 2inc, 212 kDa) an $\alpha_2\beta_2\gamma_2$ globular heterohexamer; ii) methane monooxygenase hydroxylase from *Methylococcus capsulatus*[21] (MMOH, PDB ID 1mty, 251kDa), a rectangular-shaped $\alpha_2\beta_2\chi_2$ complex and iii) urease from *Klebsiella aereogenes*[22] (PDB ID 1kra, 249 kDa for the apo enzyme) an $\alpha_3\beta_3\chi_3$ triangular-shaped assembly (Supplementary Fig.1).

We purified the proteasome lid and its subcomplexes from *RPNX-3XFLAG* strains (*MATa rpnX::RPNX-3XFLAG-HIS3*) essentially as described before[21]. Additionally, we performed control pull-downs for the proteasome-interacting proteins (PIPs) using the commercially available Tap-Tagged library[22].

Briefly, *RPNX-3xFLAG* cells were cultured, lysed and pulled down with anti-FLAG M2 agarose beads. We then subjected affinity-purified proteasomes to anion exchange chromatography after treatment with high salt to promote dissociation of the 26S proteasome and prior to elution with FLAG peptide. For enrichment of each sub-complex, we subjected the eluted samples to a 15 - 40% sucrose gradient, followed by fractionation

and SDS-PAGE. Prior to MS analysis, we pooled and concentrated lower fractions using Vivaspin centrifugal concentrators (10K MWCO, Sartorius) followed by buffer exchange using micro biospin 6 columns (BioRad) into ammonium acetate, pH 7.5 for the MS of intact assemblies and ion mobility analysis.

We lysed and pulled down Tap-Tag strains with IgG beads (Sigma I5006) coupled to Dynabeads (M-270 Epoxy, 143.01, Invitrogen). We then washed the proteins bound to beads after IP three times with 50 mM HEPES pH 7.1, 100 mM NaCl, 10 mM MgCl plus protease inhibitors (Roche), followed by cross-linking and tryptic digestion.

Cross linking coupled to Mass Spectrometry (CXMS)

For cross linking experiments, equimolar amounts of light and heavy isotopically labelled crosslinkers disuccinimidyl suberate (DSS)-d0/DSS-d12 (Creative Molecules) dissolved in dimethylformamide (DMF, Thermo Scientific) at a stock concentration of 25 mM were used. We added cross-linkers to the proteins at a final concentration of 1 mM and incubated the sample for 30 min at 37 °C with slight shaking before the cross linking reaction was quenched with ammonium bicarbonate at a final concentration of 50 mM for 10 minutes at 37 °C. We then reduced (alkylated) and digested the proteins with trypsin using standard protocols followed by a SEC enrichment protocol[12] prior to LC-MS/MS measurement on a Thermo LTQ Orbitrap XL or Thermo Orbitrap Elite mass spectrometer (LIT-orbitrap, linear ion trap-orbitrap) equipped with a standard nanoelectrospray source. We loaded the peptides onto a 75 micronID analytical column, packed in house with Michrom Magic C18 material (3 μm particle size, 200 Å pore size). We separated the peptides at a flow rate of 300 nL min^{-1} ramping a gradient from 5% to 35% mobile phase B (water/acetonitrile/formic acid; 3:97:0.1). We set the ion source and transmission parameters of the mass spectrometer to spray voltage 2 kV, capillary temperature at 200 °C, capillary voltage at 60 V and tube lens voltage at 135 V. We operated the mass spectrometer in data-dependent mode, selecting up to five precursors from a MS1 scan (resolution = 60,000) in the range of m/z 350-1,600 for collision-induced dissociation (CID). We rejected singly and doubly charged precursor ions and precursors of unknown charge states. CID was performed for 30 ms using 35% normalized collision energy and an activation *q* of 0.25. We activated the dynamic exclusion with a repeat count of 1, exclusion duration of 30 s, list size of 300 and a mass window of ±50 ppm. Ion target values were 1,000,000 (or maximum 500 ms fill time) for full scans and 10,000 (or maximum 200 ms fill time) for MS/MS scans, respectively.

We analysed cross-linked peptides using the xQuest[11] and xProphet[10] software platforms. Except where indicated otherwise, we considered only cross-links that scored a FDR of < 0.05 after xProphet analysis. For some of the reciprocal PIP pull-down and some of the recombinant “test-case” protein samples a valid FDR could not be calculated, as not enough decoy matches could be generated. In those cases, we considered as cut-off the absolute Id threshold of Id 25 (PIPs) or Id18 (recombinant test cases) and a deltaScore of < 0.95. We further analyzed all spectra by visual inspection in order to ensure good matches of ion series on both cross-linked peptide chains for the most abundant peaks.

Label Free Quantification (LFQ)

We preformed Label Free Quantification using *Progenesis* 4.0 (Nonlinear Dynamics) by automatic alignment of total ion chromatograms of raw files, using imported pep.xml files from *X!Tandem* searches against the yeast UniProtKB/Swiss-Prot protein database. We then calculated protein abundances by taking the sum of MS1 raw abundances over all biological replicates and samples and corrected for the number of amino acids of each protein. We used the resulting identifications to generate the library for subsequent cross-linking searches and identification of subcomplexes in native MS experiments.

Nano-electrospray mass spectrometry of intact complexes

We obtained mass spectra for MS and tandem MS of intact assemblies on a Q-ToF 2 (Waters/Micromass UK Ltd.) modified for high-mass operation[23], using a previously described protocol to preserve noncovalent interactions[24], with the following instrumental parameters: nanoelectrospray capillary 1600 V, sample cone 40 V, extractor cone 0 V, ion transfer stage pressure 9.5×10^{-3} bar and up to 35 μbar of argon in the collision cell. Voltage in the collision cell was at 25 V for MS and up to 200 V for tandem MS experiments. We externally calibrated spectra using a 33 mg/mL aqueous solution of cesium iodide (Sigma, St. Louis, MO). We processed the acquired data with MassLynx software (Waters, Milford MA, USA). The data is shown with minimal smoothing.

NanoES ion mobility analysis (absolute measurements)

We collected mass spectra and drift time (DT) profiles for absolute CCS measurements on a quadrupole-IM-ToF mass spectrometer in positive ion mode (Synapt G1 HDMS, Waters, Manchester, UK) with a custom made 18 cm ion mobility cell that has a radial RF ion confinement (radio frequency of 2.7 MHz and peak-to-peak amplitude of 200 V) and a linear voltage gradient to direct ions along the axis of transmission to the time-of-flight mass analyzer[25]. We acquired the measurements at 20 C° and at 0.994 Torr using helium in the mobility cell and monitored the pressures with a calibrated absolute pressure transducer (MKS Baratron model 626A, Wilmington, MA) connected directly to the ion mobility cell. We kept the cone voltage at 60 V (or 15 V for a second series of experiments), extraction cone at 1 V, trap at 10 V (5 V) and bias at 20 V. Source pressure was ~ 5.7 mbar, trap and IMS at 4.9×10^{-2} mbar and 1.4×10^{-0} mbar, respectively and ToF analyzer pressure at 2.3×10^{-6} mbar. We determined the Ω values directly from the slopes of drift time versus reciprocal drift voltage plots[26,27], using drift voltages ranging from 50 to 200 V, where the difference in potentials between the entrance and exit electrodes denotes the drift voltage.

Spatial restraints

With the experimental data from the different MS approaches, we converted them into restraints for subsequent modelling analysis. LFQ data used to define all potential members of the proteasomal assembly and the various native MS measurements to define overall stoichiometries of the intact protein complex and its various subcomplexes. From all MS data, we built an experimental tree of the proteasomal assembly (Supplementary Fig. 16). We subsequently used this tree to sample and score the predicted models. In addition, we constructed an interaction map of all subunits within the complex by integrating native MS

with identified binary interactions from CX-MS (Supplementary Figs. 15, 14). We also used the CCSs derived from IM as restraints, implemented as a harmonic function, to measure the closeness-of-fit between experiments and calculated CCSs for models. Finally, we used the confirmed high quality cross-links (using the FDR estimation) as upper bound distance restraints between the residues in proteins. We further segregated the cross-links into inter-protein cross-links which specify distance restraints between the cross-linked residues in interacting subunits and intra-protein cross-links which can be used to examine the consistency of atomic coordinates (crystal structures or homology models) with the identified cross-links.

Sampling/ Optimization

Generating an adequate number of models is a critical step of our approach. Here, we built models of the subcomplexes observed in our experiments in a stepwise manner starting from the smallest subcomplex identified in our MS-based experiments (usually a dimer) and building up to the oligomeric state of the intact complex (e.g., 6mer for MMOH and ToMOH and 9-mer for urease). In order to adequately sample the conformational space of proteins, we utilised a Monte Carlo sampling approach guided by the connectivity restraints derived from MS-based experiments. We incorporated the MS Connectivity restraint for use during sampling (http://salilab.org/imp/nightly/doc/). This restraint ensured that all subunits remained connected, as well as enabling evaluation of the ensemble of generated structures by their deviation to the experimental tree derived from MS and CX-MS data. Furthermore, the sampling explored only positions consistent with the overall stoichiometry (number of subunit copies and inter-subunit connectivities) of the respective complex under investigation. This step followed by a conjugate gradient optimization step as implemented in Integrative Modelling Platform (IMP; http://salilab.org/imp)[28]. Overall at each step we generated 10,000-20,000 model structures at atomic level, depending on the size, shape and composition of the complex. Next, we subjected these models to further analysis by measuring their closeness to the experimental data.

Scoring Function

The scoring function captured the encoded information from the raw data and used to score the candidate model structures. Along with the imposed optimization process, the restraints ensure consistency of the models generated with the experimentally available data. In the cases studied here, we first filtered our structures using the interaction maps constructed from native MS and LFQ data. Next, we evaluated the structures consistent with the input data by penalising the violation of restraints provided by the various types of structural information, namely CX-MS and CCS. We gave a penalty of a unit score to model structures for each violation of an identified residue-specific inter-subunit cross-link. We implemented the CCS restraint as a harmonic function, where perfect agreement between model and experimental CCS would take a value of 0 and violations of restraint would result in higher values[4]. Therefore, we used the CCS restraint as shown in the equation (1):

$$S_{CCS} = \left(\frac{CCS' - CCS}{\sigma'} \right)^2 \quad (1)$$

where the S_{CCS} score computed by the closeness of fit between the experimental (CCS′) and calculated (CCS) values. σ' denotes the experimental error in the data. In our experiments, using a linear drift tube, the measured CCS accuracy is estimated to be <3%. Here, in order to ensure realistic errors, we used σ' of ±6%.

We expressed the scoring function as a probability density function of the Cartesian coordinates of the assembly proteins (*C*) given information (*I*) on a restrained feature, P_f^2.

$$p\left(C/I\right) = \prod f p_f\left(C/I_f\right) \quad (2)$$

We can then write the overall scoring function as the logarithm of the probability density function:

$$S\left(C\right) = -\ ln \prod\nolimits_f p_f\left(C/I_f\right) = \sum\nolimits_f r_f\left(C\right) \quad (3)$$

Practically, we calculated the scoring function, *S(C)* by summing individual restraints, *r* with weights *w*.

$$S\left(C\right) = \sum\nolimits_f w_f r_f \quad (4)$$

We used the weighting scoring scheme, which integrates information from CX-MS and IMMS to evaluate all structural models that satisfy the input restraints derived from LFQ and native MS. Adequate sampling is critical, in order to exhaustively search the conformational space of structures fitting the data. For example, IMP makes use of Monte Carlo sampling algorithms to generate tens of thousands of random configurations. We then optimised these structures by simultaneously minimising violations of input restraints. We achieved this using conjugate gradients, and simulated annealing molecular dynamics, which refine the position of particles[3,29,30]. Ideally, the global optimum corresponds to the native assembly structure.

Weighting

As discussed in the main text, we optimized the scoring function using the training set of complexes. Bringing together data from varied sources into a single scoring function introduces heterogeneities and inconsistencies, which can be tackled by weighting the impact of the different datasets. Moreover, each of these datasets has different error features associated with both the experimental methods and the computational approaches. Here, we calculated the impact of each individual source of data as in equation (5), where $P_{(TP/y)}$ denotes the probability of identifying true positives from a certain type of data and the sum of probabilities of all types is described.

$$W_y = \frac{P_{(TP/y)}}{\sum f P_{(TP/f)}} \quad (5)$$

The probability to identify true positives from a certain type of data is given by equation (6) where *TP*/*y* denotes the true positives of a certain type and $\Sigma_f TP/f$ is the sum of true positives of all types:

$$P_{(TP/y)} = \frac{TP/y}{\sum f\, TP/f} \quad (6)$$

Such an approach allowed us to estimate the weights for the complete datasets from both types as well as for various levels of incomplete data for CX-MS. Therefore, using the values derived for the theoretical cross-links, we weighted the impact of our data from CX-MS experiments in the training set of complexes.

To estimate the impact of each individual experiment when incomplete datasets are available, we calculated the individual weights using various percentages of data available from each type. We estimated the weights for complete IM-MS and CX-MS datasets using equations (3) and (4) yielding the values of $W_{IM\text{-}MS}$=0.361 and $W_{CX\text{-}MS}$ =0.639 ± 0.05 (s.d.) for MMOH, ToMOH and Urease. Thus, as protein complexes with very different shapes and stoichiometries assigned with very similar weighting scores, we are able to use this as a generic setting for our subsequent predictions of complexes with unknown high resolution structures.

Clustering Analysis

We judged the uniqueness of the candidate models by performing clustering analysis. As such, we clustered the best-scoring models into distinct subsets based on their structural similarities, using a hierarchical tree approach[31]. Here, we hierarchically clustered the 1% of best-scoring models based on their pairwise RMSDs and represented each identified cluster by the model with the best score.

Flexibilty

In a final step, to account for flexibility we subjected the best-scoring models to dynamical analysis using NAMD (Supplementary Data)[32]. Thus, we refined the atomic positions of the subunits within the subcomplexes by performing energy minimization (Supplementary Data). We performed such an analysis at all intermediate steps needed to build the assembly. This allowed us to not only eliminate any steric clashes in the final models but also to search for the most energetically favourable conformation(s).

Rigid docking on density map

To confirm the validity of our models we fitted the model structures assembled for all complexes and subcomplexes of the proteasomeal lid into the corresponding density map[13] using the UCSF Chimera package (version 16.2)[35]. Briefly, we first manually placed the model structure into the map and then rigidly docked using the automated docking tool as implemented in UCSF chimera. We quantitatively assessed the quality-of-fit of the best-scoring structures of the intact lid complex and subcomplexes, to the density map, using the cross-correlation coefficient (CCC)

Homology modelling

We performed homology modelling for MMOH, ToMOH and Urease benchmark cases (Supplementary Table 1), the proteasomal lid (Supplementary Table 6) and the base subcomplexes (Supplementary Table 9) using MODELLER (version 9.11). We selected the final structures upon satisfaction of spatial restraints and the Discrete Optimised Protein Energy (DOPE) assessment scores[33] as implemented in MODELLER[34]. Finally, we verified the predicted structures using the PROCHEK validation program[35].

Software

Software documentation for the method is described in Supplementary Note 3 and the software is available as Supplementary Software and can be found at http://github.com/integrativemodeling/hybrid_ms_method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

MMOH and ToMOH were a generous gift of. J. Lippard, Massachusetts Institute of Technology, Cambridge, MA, USA. Urease from *Klebsiella aerogenes* was a generous gift from R.P. Hausinger (Michigan State University, East Lansing, MI, USA). This work was supported by funding PROSPECTS (Proteomics Specification in Space and Time Grant HEALTH-F4-2008-201648) within the European Union 7th Framework Program (AP, CVR, RA), ERC advanced grants *Proteomics v3.0* (233226) and *IMPRESS* (268851) to RA and CVR. HH is funded by an MRC programme grant (G1000819). FS is a Sir Henry Wellcome Fellow funded by the Wellcome Trust (Grant 09595) and CVR is funded by the Royal Society.

References

1. Robinson CV, Sali A, Baumeister W. Nature. 2007; 450:973–982. [PubMed: 18075576]
2. Alber F, et al. Nature. 2007; 450:683–694. [PubMed: 18046405]
3. Lasker K, et al. Proc Natl Acad Sci U S A. 2012; 109:1380–1387. [PubMed: 22307589]
4. Hall Z, Politis A, Robinson VV. Structure. 2012; 20:1596–1609. [PubMed: 22841294]
5. Politis A, et al. PLoS One. 2010; 5:e12080. [PubMed: 20711472]
6. Leitner A, et al. Structure. 2012; 20:814–825. [PubMed: 22503819]
7. Walzthoeni T, Leitner A, Stengel F, Aebersold R. Curr Opin Struct Biol. 2013; 23:252–260. [PubMed: 23522702]
8. Heck AJ. Nat Methods. 2008; 5:927–933. [PubMed: 18974734]
9. Ruotolo BT, Benesch JL, Sandercock AM, Hyung SJ, Robinson CV. Nat Protoc. 2008; 3:1139–1152. [PubMed: 18600219]
10. Walzthoeni T, et al. Nat Methods. 2012; 9:901–903. [PubMed: 22772729]
11. Rinner O, et al. Nat Methods. 2008; 5:315–318. [PubMed: 18327264]
12. Leitner A, et al. Mol Cell Proteomics. 2010; 11:1634–1649. [PubMed: 20360032]
13. Lander GC, et al. Nature. 2012; 482:186–191. [PubMed: 22237024]
14. Nakamura Y, et al. Biochem Biophys Res Commun. 2007; 359:503–509. [PubMed: 17555716]
15. Saeki Y, Tanaka K. Methods Mol Biol. 2012; 832:315–337. [PubMed: 22350895]
16. Bohn S, et al. Biochem Biophys Res Commun. 2013; 435:250–254. [PubMed: 23643786]
17. Barrault MB, et al. Proc Natl Acad Sci U S A. 2012; 109:E1001–10. [PubMed: 22460800]
18. Roelofs J, et al. Nature. 2009; 459:861–865. [PubMed: 19412159]

19. Tomko RJ Jr. Funakoshi M, Schneider K, Wang J, Hochstrasser M. Mol Cell. 2010; 38:393–403. [PubMed: 20471945]
20. Saeki Y, Toh EA, Kudo T, Kawamura H, Tanaka K. Cell. 2009; 137:900–913. [PubMed: 19446323]
21. Sakata E, et al. Mol Cell. 2011; 42:637–649. [PubMed: 21658604]
22. Ghaemmaghami S, et al. Nature. 2003; 425:737–741. [PubMed: 14562106]
23. Sobott F, Hernandez H, McCammon MG, Tito MA, Robinson CV. Anal Chem. 2002; 74:1402–1407. [PubMed: 11922310]
24. Hernández H, Robinson CV. Nat Protoc. 2007; 2:715–726. [PubMed: 17406634]
25. Pringle SD, et al. Int. J. Mass.Spectrom. 2007; 261:1–12.
26. Kemper PR, Dupuis NF, Bowers MT. Int. J. Mass.Spectrom. 2009; 287:46–57.
27. Bush MF, et al. Anal. Chem. 2010; 82:9557–9665. [PubMed: 20979392]
28. Russel D, et al. PLoS Biology. 2012; 10:e1001244. [PubMed: 22272186]
29. Alber F, Kim MF, Sali A. Structure. 2005; 13:435–445. [PubMed: 15766545]
30. Alber F, Forster F, Korkin D, Topf M, Sali A. Annu Rev Biochem. 2008; 77:443–477. [PubMed: 18318657]
31. Johnson S. Psychometrika. 1967; 32:241–254. [PubMed: 5234703]
32. Phillips JC, et al. J Comput Chem. 2005; 26:1781–1802. [PubMed: 16222654]
33. Shen MY, Sali A. Protein Sci. 2006; 15
34. Sali A, Blundell TL. Journal of Molecular Biology. 1993; 234:779–815. [PubMed: 8254673]
35. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. J. Appl. Cryst. 1993; 26:283–291.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

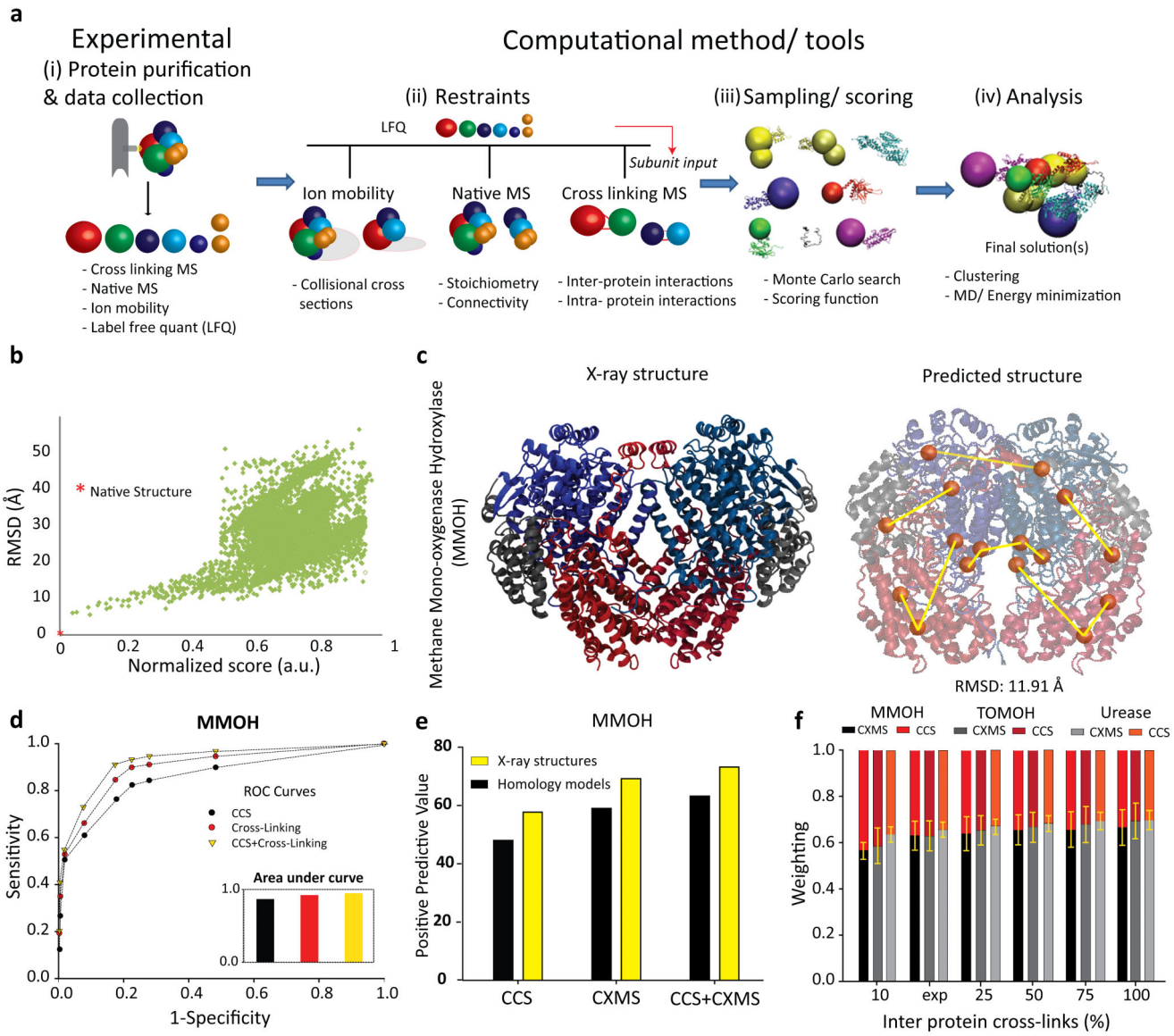

Figure 1. Workflow and benchmark of hybrid method for structure determination of protein assemblies using complementary MS data

(**a**) The workflow is composed of four steps: (i) The complex of interest is purified, either by a recombinant expression system or by affinity purification and analysed by four complementary MS-based approaches, bottom-up proteomics (LFQ), native MS, IM-MS and CX-MS, (ii) the acquired data are translated into restraints, which provide information about the overall shape of subunits and subcomplexes (IM-MS), their stoichiometry and connectivity (native MS, LFQ) and inter-protein proximities (CX-MS). (iii) Models generated by sampling the conformational space using a Monte Carlo search (>10,000 models) followed by a refinement step and **evaluation**, (iv) clustering of the best-scoring models determines the final solution(s). (**b**) The structural similarity of the models to the native structure is evaluated using their pairwise RMSD. The native structure is indicated by

the red star. (**c**) A representative structure of the best-scored ensemble of structures for MMOH oligomer (6-mer) reveals good agreement with an X-ray structure. (**d**) ROC curves assessed the accuracy and confidence levels of all restraints, individually and combined. Sensitivity, (TP/(TP+FN)), specificity, (TN/(TN+FP)), TP, true positive, FP, false positive, FN, false negative and TN, true negative). (**e)** Positive predictive values (TP/(TP+FP)), were calculated for all restraints, individually and combined, for the benchmarked complexes. (**f**) Weighting of the scoring function that accounts for both IM-MS and CX-MS restraints. The probability of identifying TPs plotted for each restraint against the percentage of inter-protein cross links available. Errors bars indicate standard deviations.

Europe PMC Funders Author Manuscripts

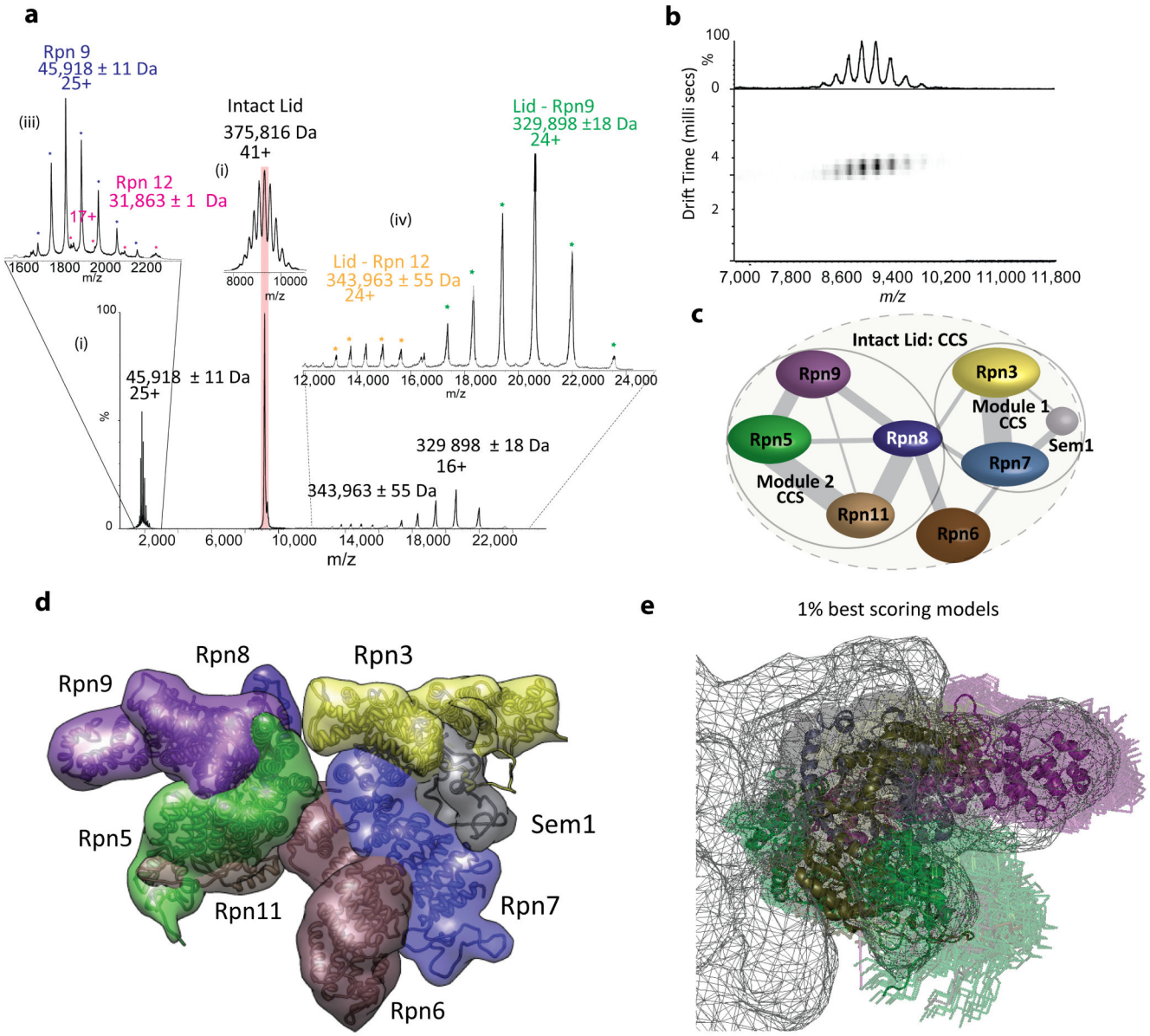

Figure 2. Structural models of the intact proteasomal lid and two distinct submodules

(**a**) Mass spectra of the intact proteasomal lid and two of its subcomplexes as observed by native MS. Insets, assigned spectra of peripheral subunits Rpn9 and Rpn12 and of the remaining "stripped" subcomplexes. (**b**) IM data plotted as drift time versus m/z, (**c**) Connectivity map of the proteasome lid generated by integrating subcomplex information from native MS with pairwise subunit contacts identified by CX-MS (**d**) A three-dimensional model of the lid predicted by integrating all MS-derived restraints. The individual subunits are depicted as simulated density maps, generated by the UCSF Chimera package (**e**) We overlaid the 1% best scoring ensemble of structures (~100 conformations) of the Rpn5/8/9/12 module and subsequently docked them into a high resolution EM

density. All models exhibited a marked similarity (RMSD<10 Å) to each other. The representative, best-scored model is shown in cartoon representation.

Europe PMC Funders Author Manuscripts

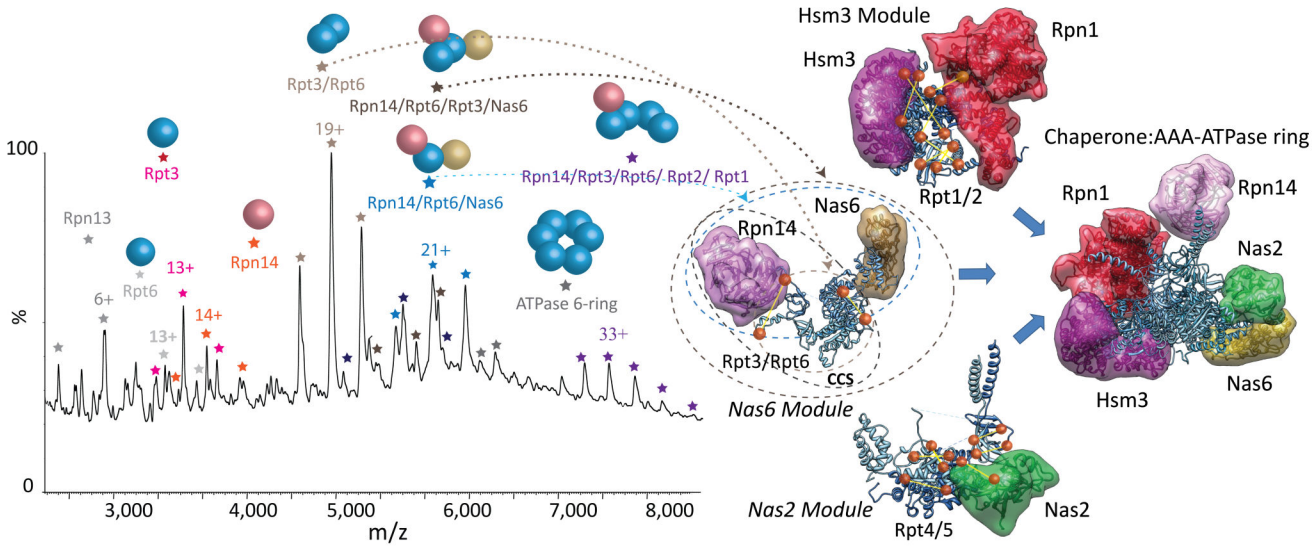

Figure 3. Structural models of chaperone:base assembly intermediates involved in the formation of the proteasomal base complex

We generated homology models and collected X-ray crystal structures of all individual subunits (base subcomplex and associated PIPs) for downstream analysis using the MS-restrained modelling strategy. Native MS spectrum from an Rpn14 pull-down showing the intact Rpn14/Rpt6/Rpt3/Nas6 and subcomplexes thereof (shaded region in the spectrum). We built a structural model for the Rpn14/Rpt6/Rpt3/Nas6 module (best-scoring model of an ensemble of structures) combining native MS, IM-MS and CX-MS. Finally, we proposed a structural model of the assembly pathway of the proteasomal base consistent with the MS-derived datasets. Experimentally identified cross-links, subcomplexes and CCS measurements are indicated. Base-dedicated chaperones with their simulated density map envelopes are shown.

Europe PMC Funders Author Manuscripts