

LECTURE 1 - THE MOLECULAR SOCIOLOGY OF THE CELL

C. V. Robinson et al. *The molecular sociology of the cell*, Nature **450**, 973, 2007.
B. Andrew et al., *Integrative structural biology*, Science, 339, 913, 2013.

The molecular sociology of the cell

Carol V. Robinson¹, Andrej Sali² & Wolfgang Baumeister³

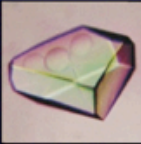
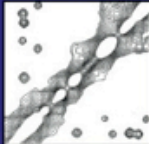

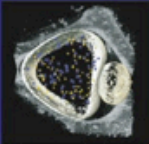
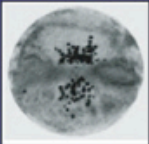


Proteomic studies have yielded detailed lists of the proteins present in a cell. Comparatively little is known, however, about how these proteins interact and are spatially arranged within the 'functional modules' of the cell: that is, the 'molecular sociology' of the cell. This gap is now being bridged by using emerging experimental techniques, such as mass spectrometry of complexes and single-particle cryo-electron microscopy, to complement traditional biochemical and biophysical methods. With the development of integrative computational methods to exploit the data obtained, such hybrid approaches will uncover the molecular architectures, and perhaps even atomic models, of many protein complexes. With these structures in hand, researchers will be poised to use cryo-electron tomography to view protein complexes in action within cells, providing unprecedented insights into protein-interaction networks.

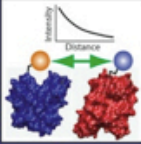
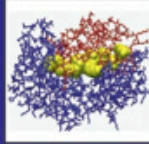




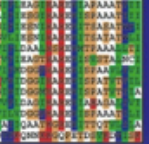
This of molecular sociology is an example of INTEGRATIVE approach, complementing those of more classical molecular biophysics (i.e. physical biochemistry, based on single molecules, dilute environments, controlled rarefied interactions)

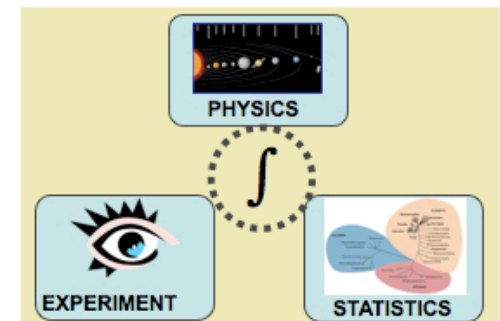
Mindset

for maximizing accuracy, resolution, completeness, and efficiency of structure determination

Use structural information from any
 source: measurement, first principles, rules;
 resolution: low or high resolution
 to obtain the set of all models that are consistent with it.

						
X-ray crystallography	NMR spectroscopy	2D & single particle electron microscopy	electron tomography	immuno-electron microscopy	chemical cross-linking	affinity purification mass spectroscopy
subunit structure	subunit structure	subunit shape	subunit shape		subunit structure	
subunit shape	subunit shape	subunit-subunit contact	subunit-subunit contact		subunit-subunit contact	
subunit-subunit contact	subunit-subunit contact	subunit proximity	subunit proximity		subunit-subunit contact	subunit-subunit contact
subunit proximity	subunit proximity	subunit stoichiometry	subunit proximity	subunit proximity	subunit proximity	subunit proximity
assembly symmetry	assembly symmetry	assembly symmetry	assembly symmetry	assembly symmetry		
assembly shape	assembly shape	assembly shape	assembly shape			
assembly structure	assembly structure					

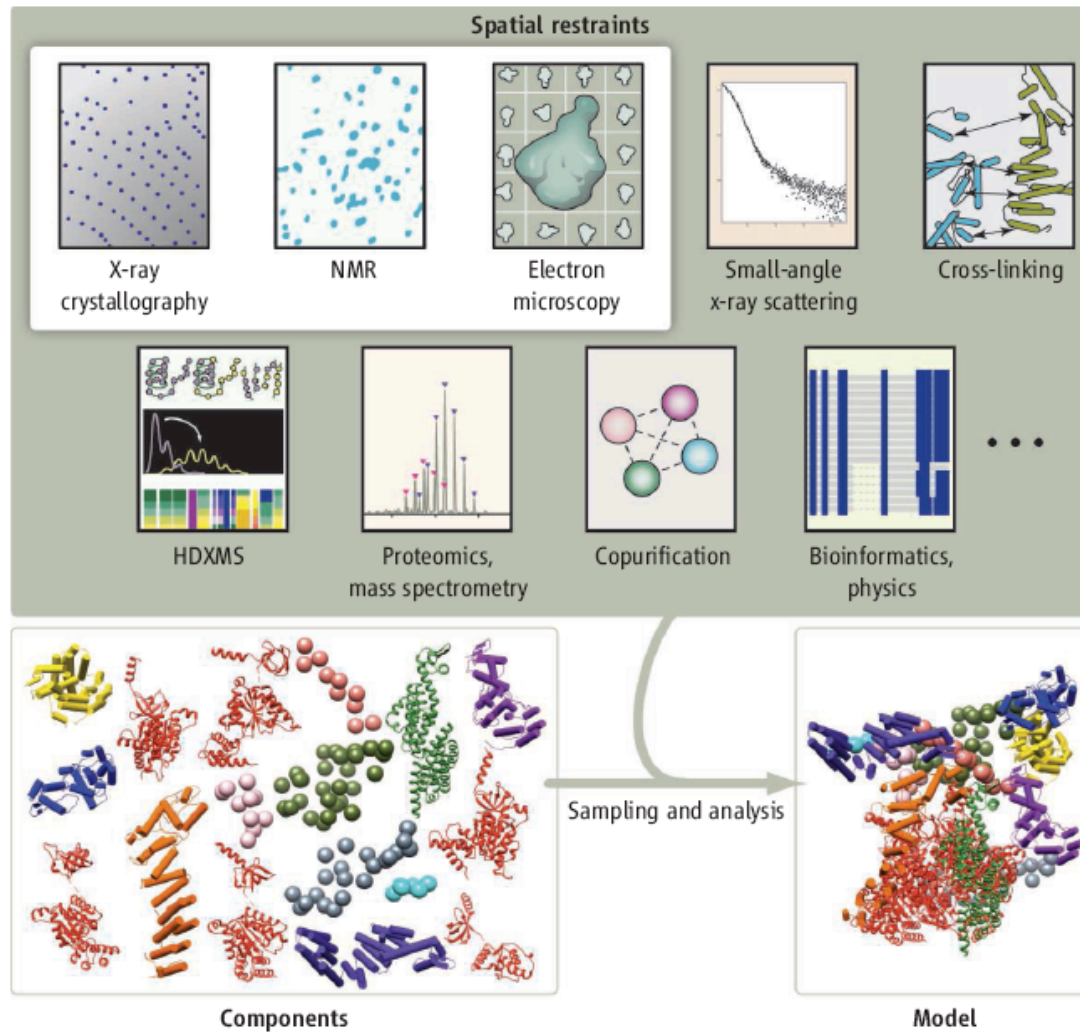
						
FRET	site-directed mutagenesis	yeast two-hybrid system	gene/protein arrays	protein structure prediction	computational docking	bioinformatics
subunit-subunit contact				subunit structure		
subunit proximity	subunit-subunit contact	subunit-subunit contact	subunit-subunit contact	subunit shape	subunit-subunit contact	subunit-subunit contact
		subunit proximity	subunit proximity			



Sali, Earnest, Glaeser, Baumeister. From words to literature in structural proteomics. *Nature* 422, 216-225, 2003.

From Andrej Sali Lab.

INTEGRATIVE STRUCTURAL BIOLOGY



Complex structure solutions. Models of macromolecules and their complexes can be constructed by combining different types of information generated by various experimental and theoretical techniques (gray box). The data are converted into spatial restraints, which are combined into a scoring function that guides sampling algorithms to obtain a detailed structural model.

A LIST OF TECHNIQUES FOR INTEGRATIVE STRUCTURAL BIOLOGY OF THE CELL

- X-ray crystallography
- NMR Spectroscopy
- SAXS (SANS)
- Cryo-electron microscopy
- FRET spectroscopy

- Sequence comparison (Evolutive pressure on structures. Paradigm: sequence > structure > function)
- Co-purification
- Hydrogen-deuterium exchange mass spectrometry (HDXMS)
- Single molecule fluorescence
- Atomic force spectroscopy
- Light scattering
- Electron paramagnetic resonance
- Double electron-electron resonance
- Chemical cross-linking
- Mutagenesis

Let us build individual lexicons (any idea about free software?)

Examples of integrative structure determinations (from Robinson2007)

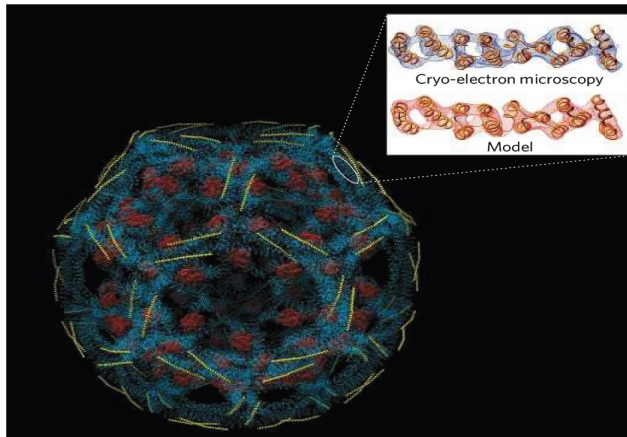


Figure 1 | A polypeptide-chain model for a clathrin D6 barrel. An α -carbon trace of the clathrin heavy (blue) and light (yellow) chains, derived by fitting atomic homology-based models into the density map from an 8 Å-resolution cryo-electron-microscopy reconstruction¹⁶. The position of a bound auxilin fragment (residues 547–910; red) was determined from a 12 Å-resolution cryo-electron-microscopy difference map. The inset zooms in to illustrate how closely the α -carbon coordinates of part of the heavy chain, as shown in the main figure (inset, lower), fit within the cryo-electron-microscopy density map (inset, upper). (Image reproduced, with permission, from ref. 16.)

Alpha-clathrin D6 barrel

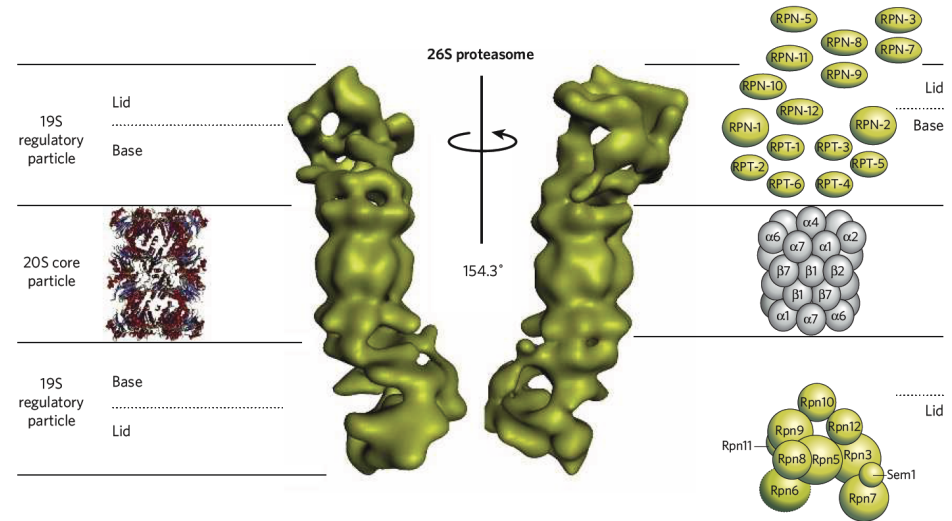


Figure 3 | The molecular architecture of the 26S proteasome. The 26S proteasome consists of 19S regulatory particles associated with the ends of a barrel-shaped 20S core particle. The part of each 19S regulatory subunit that is closest to the core is known as the base, and the part that is farthest away is known as the lid. Crystal structures have been obtained for archaeal, bacterial and eukaryotic 20S core particles^{63,77–79} (left, α -helices in red, and β -sheets in blue). For the eukaryotic 26S holocomplex, only a low-resolution structure, obtained by cryo-electron microscopy⁶⁷, is available (centre; two orientations, rotated by 154.3°). Topological models

of the regulatory particle have been deduced from yeast two-hybrid screens of *Caenorhabditis elegans* proteins⁶⁸ (upper right) and from mass spectrometry of yeast proteins⁴⁵ (lower right). These models agree reasonably well, albeit not completely. A topological model of the 20S core (centre right) that corresponds to the crystal structure (left) is also shown. No attempt has yet been made to obtain the molecular architecture of the entire 26S proteasome by integrating these topological models with the cryo-electron-microscopy map. RPN, non-ATPase subunit; RPT, ATPase subunit. (Central image reproduced, with permission, from ref. 65.)

26S proteasome

To understand what 26S means see for example par 12.4.5 of PBC (physical Biology of the cell).

In a nutshell: S (Svedberg). In a centrifuge: $V_D = m g_c / \gamma$; $S = m / \gamma \times 10^{-13}$ is the sedimentation coefficient)

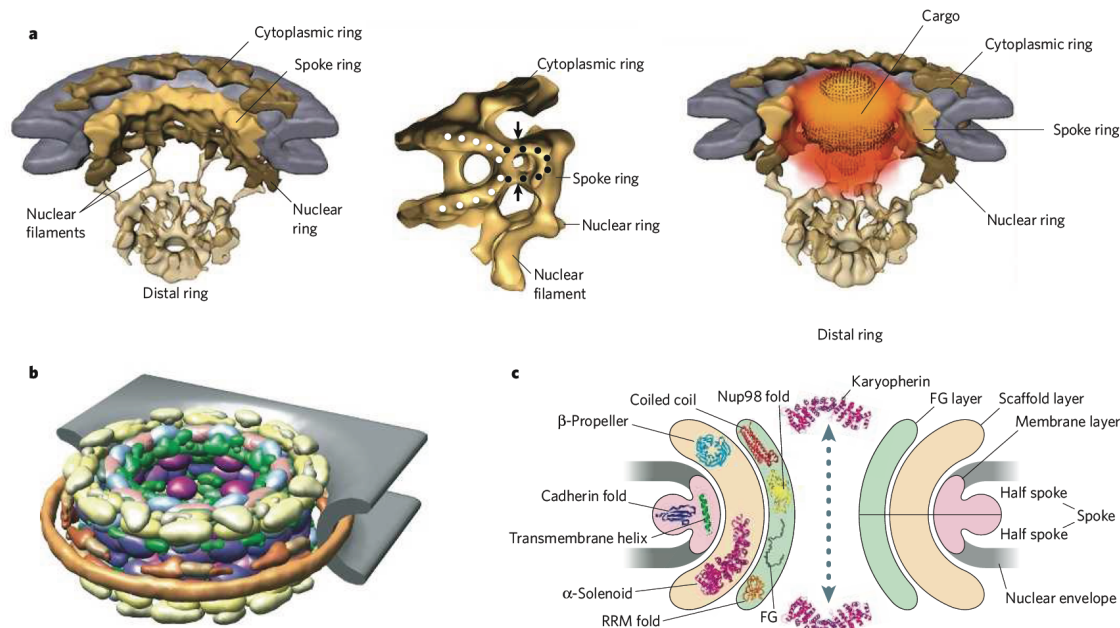


Figure 4 | The molecular architecture of the NPC. By using a variety of techniques, different aspects of the NPC structure have been revealed. **a**, Using cryo-electron tomography, a density map of the *Dictyostelium discoideum* NPC at 5.8 nm resolution was generated, allowing single molecules to be observed during nuclear import²⁰. A cutaway view of the structure of rejoined asymmetrical units is shown (left), with subjective segmentation for the cytoplasmic ring, spoke ring and nuclear ring (brown and yellow), and the inner nuclear membrane and outer nuclear membrane (that is, the nuclear envelope; grey). For clarity, the central plug (that is, the transporter) has been omitted, and the basket with nuclear filaments and distal ring was rendered transparent. A cutaway view of a protomer is shown (centre). The fused inner nuclear membrane and outer nuclear membrane (white circles), as well as the clamp-shaped spoke structure (black circles), are indicated; arrows mark the entry and exit of what seems to be a channel. A cutaway view of the NPC structure with a three-dimensional probability distribution of import cargo is shown (right). The classical import cargo NLS-2GFP (Asn-Leu-Ser with two green fluorescent protein molecules

attached) was labelled with gold, and the probability distribution for the cargo (orange; brightness indicates higher probability) is superimposed onto the central plug (brown dots). **b**, Various experimental data were integrated⁷, revealing the configuration of the 456 core proteins (excluding FG (Phe-Gly) repeats in FG nucleoporins and the basket) that form the yeast NPC²¹. The inner and outer nuclear membranes (grey) are shown. The NPC proteins are coloured according to their assignment to various NPC modules: membrane rings (brown), outer rings (yellow), inner rings (purple, light and dark shades), linker nucleoporins (blue and pink, light shades) and FG nucleoporins (green). (Panel adapted, with permission, from ref. 7.) **c**, Structural folds were assigned to the domains of the NPC proteins, by comparing their sequences to those of known protein structures, revealing a simple fold composition and modular architecture for the NPC²². The architecture of the NPC ring, viewed as a transverse section, is segregated into three layers: membrane (pale pink), scaffold (pale yellow) and FG (pale green). The arrow denotes the direction of cargo transport. RRM, RNA-recognition motif.

Nuclear pore complex NPC A.
From Robinson2007

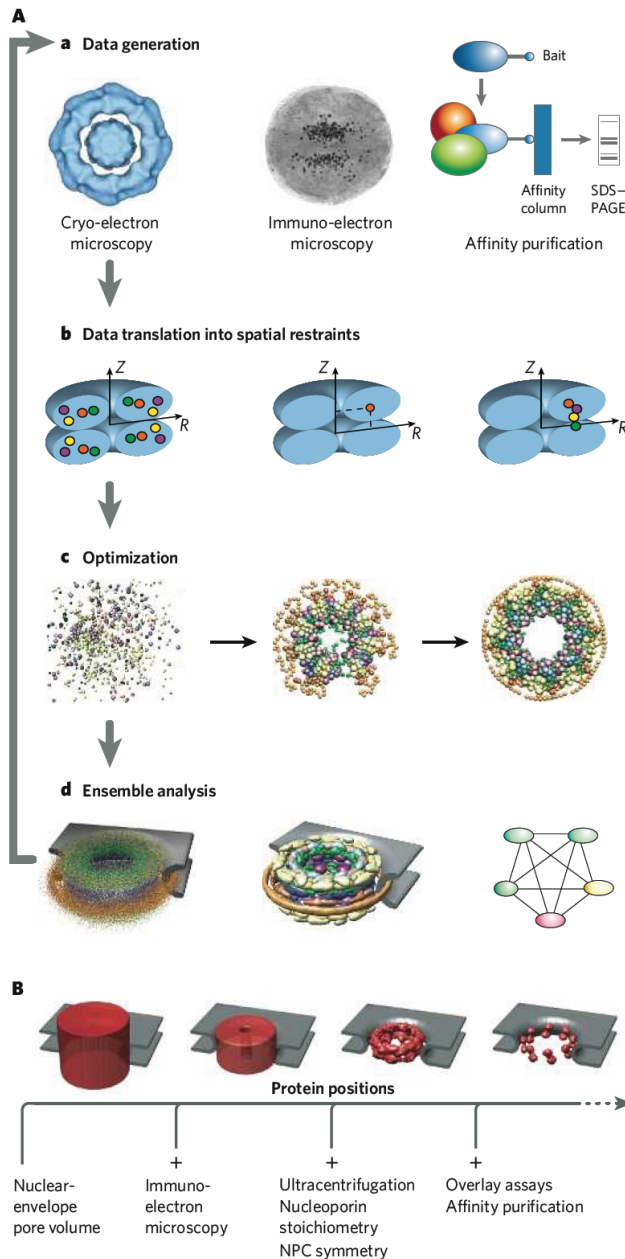


Figure 5 | Integrative structure determination. A, Using the NPC as an example⁷, the four steps to determine a structure by integrating varied data are illustrated. These steps are data generation (a), data translation into spatial restraints (b), optimization (c) and ensemble analysis (d). **a**, First, structural data are generated by experiments, such as cryo-electron microscopy (left), immuno-electron microscopy (centre) and affinity purification of subcomplexes (right). Many other types of information can also be included. **b**, Second, the data and theoretical considerations are expressed as spatial restraints that ensure the observed symmetry and shape of the assembly (from cryo-electron microscopy, left), the positions of constituent gold-labelled proteins (from immuno-electron microscopy, centre) and the proximities of the constituent proteins (from affinity purification, right). The assembly is indicated in blue, and constituent proteins are indicated as coloured circles. **c**, Third, an ensemble of structural solutions that satisfy the data is obtained by minimizing the violations of the spatial restraints (from left to right). **d**, Fourth, the ensemble is clustered into sets of distinct solutions (left), and analysed in different representations, such as protein positions (centre) and protein-protein contacts (right). The integrative approach to structure determination has several advantages. First, synergy among the input data minimizes the drawbacks of incomplete, inaccurate and/or imprecise data sets. Each individual restraint contains little structural information, but by concurrently satisfying all restraints derived from independent experiments, the degeneracy of structural solutions can be markedly reduced. Second, this approach has the potential to produce all structures that are consistent with the data, not just one structure. Third, the variation between the structures that are consistent with the data allows an assessment of whether there are sufficient data and how precise the representative structure is. Last, this approach can make the process of structure determination more efficient, by indicating which measurements would be the most informative. **B**, When applying the process described in **A**, the position of each protein is specified with increasing accuracy and precision as each type of synergistic experimental information is added⁷. Each panel illustrates the localization volume (red) of 16 copies of nucleoporin 192 (Nup192) in the ensemble of NPC structures that satisfy the spatial restraints corresponding to the experimental data sets indicated. The smaller the volume, the better the proteins are localized. Further experiments could localize the proteins to a greater degree, as indicated by the dashed arrow. Therefore, the NPC structure is, in essence, 'moulded' into shape by the large quantity of diverse experimental data. (Panel reproduced with permission from ref. 7.)

Nuclear pore complex NPC B.
From Robinson2007

Summary: integrative structural approach (scalable, applicable from single proteins to complexes)

- 1.Data generation
- 2.Data translation into spatial restraints
- 3.Optimization (computational fitting, score functions)
- 4.Ensemble analysis (Bayesian inference ?)

Important examples

Rosetta (www.rosettacommons.org (David Backer))

Integrative modeling (www.integrativemodeling.org)

Inferential Structure Determination

W. Rieping et al. see www.isd.bio.cam.ac.uk

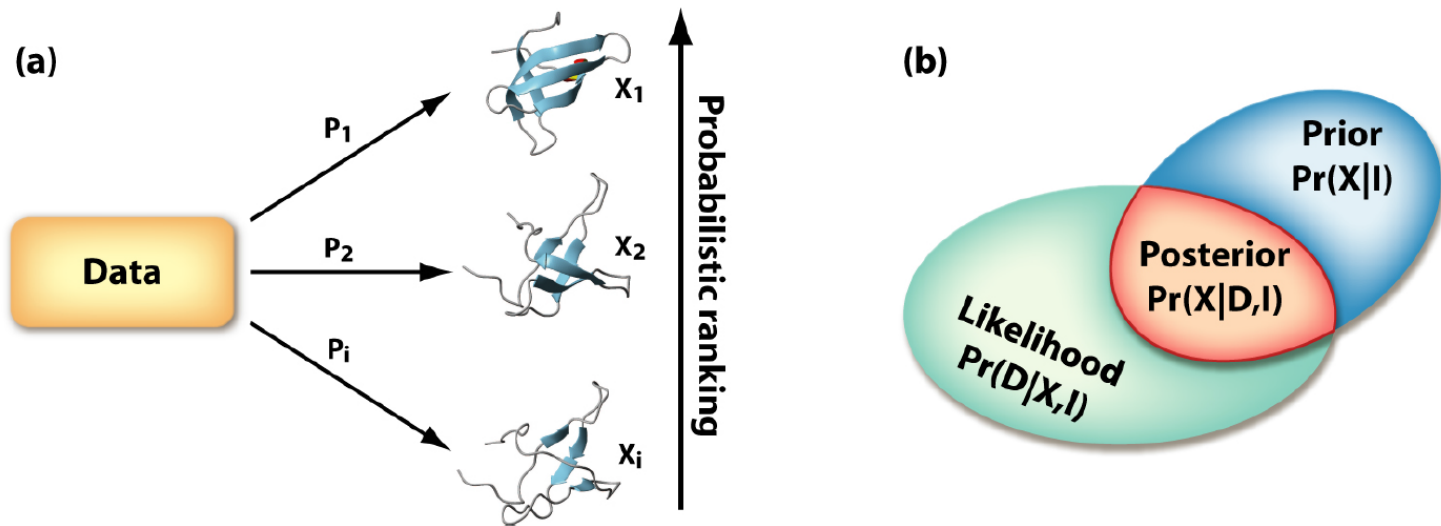


Figure 1: Probabilistic ranking and Bayes' theorem. The experimental data are used to rank every conformation of a protein in terms of a probability (a), i.e. we do not derive geometrical constraints that would completely rule out structures. If of two conformations one has higher probability, then it is more supported by the data. The spread of the probability distribution reflects how well we can determine a structure from the available information. If only a single conformation has non-zero probability, the data uniquely determine the structure. If the probabilities are constant, the available data is uninformative with respect to the structure. Realistic cases lie somewhere in between. Bayes' theorem (b) combines prior information with experimental evidence, represented in terms of a likelihood function, in a consistent way. The posterior distribution represents everything that can be said about the molecular structure given the data and our prior knowledge.

Bayesian structure determination

The aforementioned problems have a common source: Structure determination requires reasoning from incomplete information which is why protein structures necessarily remain uncertain to some degree. Existing methods, however, are based on the concept of structural constraints, and are therefore incapable of taking this uncertainty into account. In essence, ISD relies on Bayesian probabilistic inference that represents any uncertainty through probabilities which are then combined according to the rules of probability calculus. The application of this approach is computationally demanding, and has become feasible only recently due to the development of efficient stochastic sampling algorithms (Markov chain Monte Carlo methods) and increased computational power provided by computer clusters.

Where is the problem...

The principal difficulty in structure determination by NMR is the lack of information required to unambiguously reconstruct a protein structure. Conventional methods view structure determination as a minimisation problem: A so-called “hybrid energy” function combines a pseudo energy term that incorporates the experimental constraints with a force field describing the physical interactions between the atoms. Minimising the hybrid energy is then assumed to answer what the “true” structure of a molecule is. This rule, however, implicitly assumes that there is a unique answer. Repeating the minimisation procedure multiple times, as is standard practice in conventional approaches, does not adequately represent the ambiguity and makes it difficult to judge the validity and precision of NMR structures in an objective way.

$$\Pr(D|X, \alpha, \sigma, I) = \prod_i \frac{1}{\sqrt{2\pi}\sigma I_i} \exp \left\{ -\frac{1}{2\sigma^2} \left[\log I_i - \log(\alpha d_i^{-6}) \right]^2 \right\}$$

$$\Pr(X|I) \propto \exp \left\{ -\beta E(X) \right\}$$

$$\Pr(X, \alpha, \sigma|D, I) \propto \alpha^{-1} \sigma^{-(n+1)} \exp \left\{ -\beta E(X) - \frac{1}{2\sigma^2} \sum_i [\log I_i - \log(\alpha d_i^{-6})]^2 \right\}.$$

Numerical sampling of the posterior distribution: Gibbs sampler

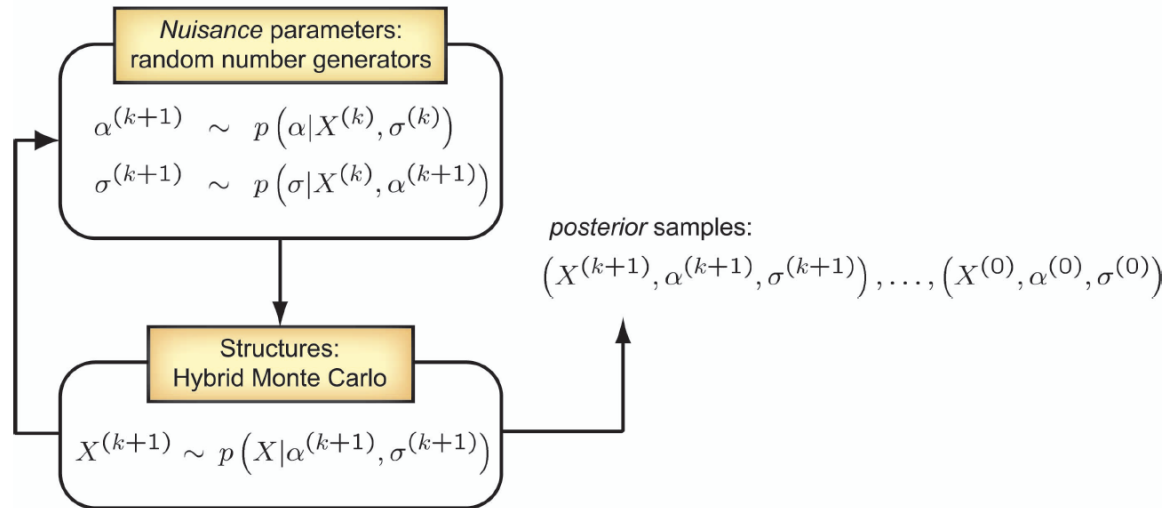


Figure 2: Gibbs sampling scheme used to generate samples from the posterior probability for protein conformation X and nuisance parameters α and σ . Gibbs sampling is an iterative scheme that, upon convergence, produces samples from the full posterior distribution. The nuisance parameters can directly be drawn from their posterior probabilities. To update the conformational degrees of freedom we employ the HMC algorithm. This algorithm uses molecular dynamics [14] to generate a candidate conformation which is accepted according to the Metropolis criterion [15]. The molecular dynamics is defined by the negative log-posterior probability with fixed nuisance parameters.

Drawback: getting trapped into metastable states

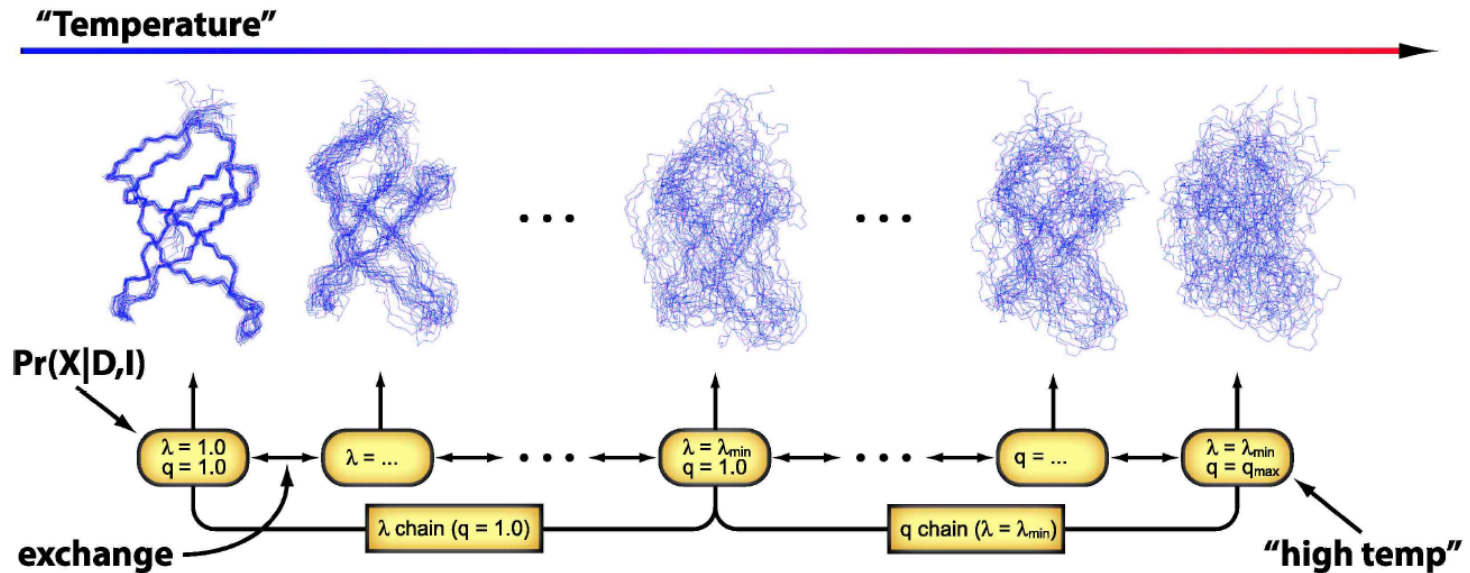


Figure 3: Replica-exchange Monte Carlo algorithm. We embed the Gibbs sampler (figure 2) in a Replica-exchange Monte Carlo scheme which simulates a sequence of “heated” replicas of the system. Two generalized temperatures, λ and q , control the shape of the likelihood function and of the prior distribution, respectively. For $\lambda = 1$ the data are switched on, for $\lambda = 0$ they are switched off. For $q = 1$, the canonical ensemble is restored as prior probability [cf. Eq. (2)]. For $q > 1$ physical interactions are gradually switched off and the prior probability approaches a flat distribution over conformation space. We arrange the replicas in such a way that first the data are switched off (by gradually decreasing λ). In the other half of the arrangement, we additionally switch off the physical interactions by increasing q .

Avoiding metastabilities: Replica exchange + Tsallis ensemble