Equilibrium Sampling Approach to the Interpretation of Electron Density Maps

Andreas Vitalis,[1,*] and Amedeo Caflisch[1]

[1]*Department of Biochemistry*

University of Zurich

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

*: To whom correspondence should be addressed:

Andreas Vitalis: Tel: +41446355597, E-mail: a.vitalis@bioc.uzh.ch

RUNNING TITLE

Equilibrium Sampling with Density Restraints

SUMMARY

The derivation of molecular models from spatial density data generated by X-ray crystallography or electron microscopy is an active field of research. Here, we introduce and evaluate an approach relying on the equilibrium sampling of energy landscapes describing restraints to experimental input data. Our procedure combines density restraints with replica exchange methodologies in the parameter space of the restraints, and we demonstrate its applicability to both flexible polymers and the assembly of protein complexes from rigid components. For the most difficult system studied, we highlight the importance of advanced data analysis techniques in mining poorly converged data further. Successful and unbiased interpretation of input density maps is a prerequisite for using this approach as an auxiliary restraint term in molecular simulations. Because these simulations will also utilize physical interaction potentials, we hope that they will contribute to deriving families of structural models for input data that are ambiguous *per se*.

HIGHLIGHTS

- B-splines are used to describe restraints to density maps with implicit background
- Advanced sampling methods yield correct solutions for 3 examples
- Analysis of simulation data as statistical ensembles is possible
- Advanced data mining is helpful for poorly converged cases

INTRODUCTION

In structural biology, information at atomic resolution is routinely obtained by two major experimental techniques: X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. In addition, electron microscopy (EM) has seen major advances in data collection and processing in recent years, which have pushed its resolution down to near-atomic resolution (Cheng and Walz, 2009). Interactions of biological matter with electromagnetic or particle waves through absorption, scattering, or other phenomena produce a signal that is difficult to assign to specific system components. This is in contrast to NMR phenomena, which are very sensitive to nucleus type. Dramatically improved contrast in absorption or scattering experiments is obtained by incorporating or absorbing much larger nuclei, and this is used in such techniques as multiple isomorphous replacement (Green et al., 1954) in X-ray crystallography or immunogold labeling (Faulk and Taylor, 1971) in EM. Conversely, for unlabeled biomolecules, we can not expect to distinguish signals coming from, *e.g*., carbon *vs*. oxygen. This implies that even for high-resolution data, structural models are obtained indirectly, *i.e*., by minimizing deviations between observed signal and model-derived signals. The models invariably incorporate auxiliary information, which can be very basic (*e.g*., geometric parameters of chemical groups) or highly advanced (*e.g*., solved crystal structures of individual components when interpreting EM density maps of macromolecular assemblies, Wriggers et al., 1999).

Structural information is particularly difficult to obtain for systems exhibiting heterogeneity. The interpretation of data on these systems is often ambiguous (DePristo et al., 2004), and models relying on a combination of independent measurements and human intuition have to be constructed (Marsh et al., 2012; Petoukhov et al., 2002). The problem is challenging because conformational disorder may lead to indistinguishability of signals from background. For assemblies such as protein fibrils (Fändrich et al., 2009), filaments (Fujii et al., 2010), or large complexes (Cheng and Walz, 2009; Walzthoeni et al., 2013), any level of conformational or other polydispersity will render the signal at least partially

ambiguous. This can result in models of mixed resolution, which are similarly found in NMR ensembles of proteins structures. Structural data that are of poor or inconsistent quality directly motivate the use of these experimental signals in computer simulations based on physicochemical principles (Marsh et al., 2012; Robustelli et al., 2010; van Gunsteren et al., 2008). Ensembles generated *in silico,* which are physically sound and also explain the experimental data, can replace or add to those models relying partially on human or database input. It is important that this can highlight limitations to the interpretability of the experimental data alone (by quantifying model degeneracy), and that it can rank degenerate models by their physical feasibility rather than by human bias.

Crystal structure refinement generally utilizes least-squares deviations between model-derived and measured signals whether performed in real (Diamond, 1971) or in reciprocal space (Jack and Levitt, 1978). The input density may be subjected to a linear transform, or the normalized cross-correlation coefficient, which is scale- and translation-invariant, is maximized (Fabiola and Chapman, 2005). Due to a rigorous theoretical framework being unavailable, model-derived signals generally use Gaussian or similar functions of controllable width to represent atoms (Chapman, 1995; Tama et al., 2004). Refinement relies on optimization protocols to minimize a composite cost function incorporating the signal deviation along with stereochemical and excluded volume terms (Brünger et al., 1987). The underlying assumption is that the initial model is close enough to the optimal solution. Often, to aid with computational efficiency, the problem is also decomposed into piecewise optimization tasks for subsets of the system. Similar ideas have been developed for EM data. Early efforts focused on rigid-body optimization (Chacon and Wriggers, 2002; Roseman, 2000), but semi-flexible (Fabiola and Chapman, 2005; Tama et al., 2004; Topf et al., 2008) and fully flexible treatments (Trabuco et al., 2008; Vashisth et al., 2012) are now in use.

Here, we describe a versatile approach to compute lattice-based, physical property densities from particle coordinates. We also establish a protocol to quantitatively interpret an input map assuming a linear relation between signal and property density, thus providing the framework for a lattice-based

restraint potential to be used in molecular simulations. This article serves to demonstrate the validity of the approach by 3 proof-of-concept applications posing qualitatively different challenges. We utilize methods of broader relevance, specifically replica exchange (REX) sampling in the parameter space of the restraint potential, and a hybrid sampling protocol. The power of quantitative, statistical approaches to the mining of these types of simulation data is demonstrated. Points of departure from existing work are that the method is unbiased, *i.e.*, not a knowledge-based approach, that an implicit background can be accounted for, and that it deals naturally with data ambiguity.

EXPERIMENTAL PROCEDURES

RESTRAINING MOLECULAR SIMULATIONS TO A TARGET MASS DENSITY

In this section, we provide a brief outline of our approach. The goal is a quantitative interpretation of an input density map, which is approximated by a simulated density derived from molecular simulation. The potential is enumerated over lattice cells, and penalizes density deviations in either direction. As a consequence, simulations are expected to produce conformations that are largely free of volume overlap even in the absence of physical interaction potentials.

Lattice-Based Density Function from Atomic Positions

We need to transform a set of N atomic positions, $\left\{\vec{r}_n\right\}$, into a lattice-based property density, ρ. Here, we utilize cardinal B-splines in the same vein as in the particle-mesh Ewald method (Essmann et al., 1995), where the property would be atomic charge. Cardinal B-splines replace a point-like representation with a distribution that is polynomial in nature and has finite support:

$$\rho_{ijk} = V_{ijk}^{-1} \sum_{n=1}^{N} x_n \prod_{d=1}^{3} B_A(r_n^d - P_{ijk}^d) \qquad (1)$$

The lattice cell with indeces i, j, and k and volume V_{ijk} is associated with a reference point P_{ijk} corresponding to its center. For each atom, its property, x_n, is distributed across the lattice with a weight dependent on the distance from the reference point. Cardinal B-splines are factored for spatial

dimensions (d) and conveniently satisfy that the integral over all lattice cells recovers the total property value. Their order, A, determines the exact shape of the function with limiting values of $A = 1$ corresponding to a rectangular binning function and $A \rightarrow \infty$ being a Gaussian (Fig. 1a). All B-splines, $B_A(x)$, are assumed centered making them symmetric about $x = 0$. Because support is finite (A^3 lattice cells), the actual cost inherent in eq. 1 is linear in N and independent of lattice size.

Mass Density with Background

We choose mass density as the target quantity to restrain (atomic number could be used instead of mass). In general, the input density will have a background signal stemming not from vacuum, but often from aqueous solvent. To emulate this signal in simulations without an explicit representation of solvent, we use both lattice-based mass and volumes to arrive at:

$$\rho_{ijk} = \rho_{sol} + V_{ijk}^{-1} \sum_{n=1}^{N} \left[m_n - \gamma_n V_n \rho_{sol} \right] \prod_{d=1}^{3} B_A(r_n^d - P_{ijk}^d) \qquad (2)$$

Here, m_n are atomic masses, V_n are atomic volumes estimated from published radii (Vitalis and Pappu, 2009), γ_n is a factor correcting for volume overlaps between topologically connected atoms, and ρ_{sol} is the assumed physical background density. The values for γ_n are determined as $\gamma_n = V_n^{eff} / V_n$, where V_n^{eff} is obtained by subtracting from an atom's volume half of the overlap volume with each covalently bound partner atom as determined by linear approximations. If $A > 2$, ρ is a continuously differentiable function usable in gradient-based simulation or modeling techniques.

Processing of Experimental Input Densities

Experimental input densities are most likely derived from X-ray crystallography or EM. Sample heterogeneities and their impact on averaged signals, possible overlap of radiation diffraction and absorption processes, sample damage due to continuing exposure, and limitations in signal processing may all weaken the link between signal and its physical source. We therefore treat the input density as having arbitrary units, and will assume that the signal is linearly proportional to mass density as in related work (Trabuco et al., 2008).

The linear transform is meant to accomplish two things: 1) it aligns the background signal in the input with the chosen value for ρ_{sol}; 2) it allows scaling of the data so that there is control over the assumed contrast levels within physically reasonable bounds. The scaling is controlled by a parameter, ω_t:

$$\Xi_{ijk} \propto c_2(\omega_t)\omega_{ijk} \quad (3)$$

Here, ω is the input density, Ξ is the interpreted density, and ω_t is used to determine the scaling factor, c_2. Qualitatively, ω_t should correspond to an isocontour value for the input that would give a rough molecular envelope. Importantly, ω_t does not remove or distort information contained in the input map. Details are given in the Supplemental Information (SI). Note that it is common to consider auxiliary modifications of the input such as flattening or re-binning, which do alter the data qualitatively. The lack of assumptions we make means that the above treatment works equally well for other properties and other types of input data.

Lattice-Based Harmonic Restraint Potential

By virtue of having defined lattice-based mass densities both as input and as derived from simulations, constructing a restraint potential from their squared deviations is straightforward:

$$V_D = f_D \sum_{i,j,k} \left(\rho_{ijk} - \Xi_{ijk}\right)^2 \quad (4)$$

Here, f_D is a unitless scale factor. Evaluation of eq. 4 scales directly with the number of lattice cells. Forces require the partial derivative of ρ_{ijk} with respect to a given atomic position (compare eq. 2) and incur a cost that is *O(N)*. Eq. 4 could also have been written using an ensemble average for ρ_{ijk}, but this is not explored here. As written, each instantaneous conformation is penalized for deviations from the input map, and the underlying assumption is that a single conformation can explain the input density. This assumption may be incorrect. However, the order of B-splines, *A*, in conjunction with lattice dimensions defines an inherent averaging at the level of an instantaneous conformation, which can be matched to the contrast level set by ω_t. Larger values for *A* will “smear out” each atom's property, and

we use this in the generation of synthetic maps. Modifications to eqs. 2 and 4 may be required for input maps with highly heterogeneous contrast levels.

SAMPLING METHODOLOGY

Our general sampling engine is a hybrid scheme using both Monte Carlo (MC) and force-based integrators in internal coordinate space (IMD) as implemented in and documented for the software CAMPARI (http://campari.sourceforge.net). The degrees of freedom can but need not include all rigid-body coordinates, all freely rotatable dihedral angle degrees of freedom, and pucker degrees of freedom for the flexible rings in proline and sugars. The latter are peculiar in that they are only sampled by the MC segments (Radhakrishnan et al., 2012), but have to be frozen in IMD. All systems are represented with united atoms in the CHARMM19 convention (Brooks et al., 1983) and masses are adjusted accordingly. For all runs, steps alternated in segments of 6000 IMD steps followed by 600 MC steps (90.9 % IMD *vs*. 9.1 % MC). The velocity rescaling thermostat (Bussi et al., 2007) was used throughout in IMD to ensure sampling of proper NVT ensembles at 300 K.

The Hamiltonian generally consists of the density restraint potential (eq. 4) and few bonded potentials taken from appropriate force fields for polypeptides and polynucleotides, respectively (see SI). The potentials are required, for example, to keep peptide moieties planar, or to preserve the covalent geometry of flexible rings. The ubiquitin system (see below) employs an additional term, *viz*., a statistical potential biasing ϕ/ψ-angles based on local preferences (see SI and Fig. S1). The potential is included because of the challenging nature of this system, but, importantly, it employs no system-specific information for ubiquitin or any other protein. The lack of any excluded volume or stiff harmonic terms in the Hamiltonian means that the integration time step for the IMD portion can be large. Details regarding the simulation protocol are provided in the SI.

In terms of performance, a single replica for a system with 749 atoms, more than 400 degrees of freedom, 3.8×10^4 lattice cells, and $A = 5$ produced $\sim 2 \times 10^7$ steps per day on a single core of the Schrödinger supercomputer at the University of Zurich.

TEST SYSTEMS

Test systems of increasing difficulty are chosen to highlight the potential and versatility of our approach.

Actin-Related Protein 2/3 (Arp2/3) Protein Complex

The Arp2/3 complex from *Bos taurus* as resolved by crystallography (PDB ID: 1TYQ Nolen et al., 2004) is composed of 7 polypeptide chains. Cofactors and water were removed, and missing atoms in incomplete residues were rebuilt (all dihedral angles at 180º). Residues missing entirely were not reconstructed. The resultant structure contained 16502 atoms and was used to generate synthetic low-resolution maps. The system was defined to be a periodic lattice with cubic cells of 3 Å and dimensions of 36x40x58 in the *x*, *y*, and *z* directions, respectively. This comfortably accommodates the assembled complex. Using eq. 2 with $\rho_{sol} = 1.01\ \mathrm{g/cm^3}$ and A_g being 2, 8, or 17, synthetic maps of varying resolution were created (see Fig. 1a). The notation A_g is used to distinguish it from parameter A used in eq. 2 during the actual simulations.

These maps served as input to independent REX runs. Every such run utilized 16 replicas that differed in their values for f_D ranging from 0.01 to 0.16 in steps of 0.01. All other parameters were constant between runs and replicas, *viz.*, $\rho_{sol} = 1.01\ \mathrm{g/cm^3}$, $A = 3$, $M_M = 191.7\ \mathrm{kD}$, and $\omega_t = 1.55\ \mathrm{g/cm^3}$. Each simulation consisted of 1.15×10^7 total steps per replica. The degrees of freedom were just the rigid-body coordinates, which allows for a very large time step of 5.0 ps and a net simulation time per replica of ~52 μs. Further details are given in the SI.

RNA Stem-Loop

The NMR ensemble (8 structures) of a 17-residue RNA stem-loop (PDB ID: 2LBL Chang and Nikonowicz, 2012) was used to construct average synthetic density maps consistent with the NMR ensemble. The 408 united atoms were used as input to eq. 2 assuming a periodic lattice with cubic cells of 1.4 Å side length and dimensions of 21x35x21. The B-spline order in map generation, A_g, was either

2 or 8. All structures in 2LBL were given equal weight.

The two maps served as input to independent REX runs each using 16 replicas with values for f_D ranging from 0.02 to 0.17 in increments of 0.01. The threshold parameter was $\omega_t = 1.1\ \mathrm{g/cm^3}$ for the higher and $\omega_t = 1.5\ \mathrm{g/cm^3}$ for the lower resolution. Varying ω_t is generally necessary for high-resolution data in order to ensure comparable average molecular densities for the interpreted maps (here, $1.86\ \mathrm{g/cm^3}$). The remaining parameters were constant between runs and replicas, *viz*., $\rho_{sol} = 1.01\ \mathrm{g/cm^3}$, $M_M = 5.35$ kD, and $A = 5$. Each simulation consisted of 1.6×10^8 total steps per replica. The degrees of freedom included rigid-body coordinates (6), nucleic acid backbone and side chain torsions (117), and the pucker angles of the ribose moieties (17 sets of highly coupled degrees of freedom). The time step was 10 fs for a total simulation time of ~1.5 μs per replica. MC moves are described in the SI and included a dedicated move type for sugar pucker angles.

Ubiqutin

We obtained the crystallographic electron density of a unit cell for the 76-residue protein ubiquitin (PDB ID: 1UBQ Vijay-Kumar et al., 1987) from the Uppsala Electron-Density Server (Kleywegt et al., 2004). The formal resolution of 1.8 Å is sufficient to isolate a single protein molecule from the unit cell with the help of UCSF Chimera's volume viewer (Pettersen et al., 2004). The density describing a single molecule of ubiquitin was surrounded by a flat background signal with numerical value -1.0, and the resolution was reduced by re-binning it by roughly a factor of 2.0. This is for three reasons: 1) our method is not a crystallographic refinement or solution tool meant to operate on data at these resolutions; 2) performance is improved by the more tractable number of lattice cells; 3) as will become clear below, very high resolution input data is likely to result in trapping, which slows or prevents convergence. Tests performed at the original resolution produced data of no statistical significance. After re-binning, the lattice has 32x33x36 roughly cubic cells with side lengths of 1.1915, 1.2058, and 1.2063 Å, respectively. This size is comparable to the stem-loop example (1.4 Å).

We next set up 4 identical, but independent REX runs each using 48 replicas that differed in their values for f_D and ω_t. Replicas were arranged such that either, but never both parameters vary minimally between neighboring replicas. We chose a very compact schedule in terms of both values to ensure that the REX technology remained effective (see SI for details). The remaining parameters were constant between runs and replicas, *viz.*, $\rho_{sol} = 1.01\ \mathrm{g/cm^3}$, $M_M = 8.56\ \mathrm{kD}$, and $A = 5$. The sampled degrees of freedom are rigid-body coordinates (6), polypeptide ω, ϕ, ψ, and χ torsions (386), and the pucker angles in 3 proline residues. Some rotatable dihedral angles involving either only hydrogen atoms or symmetric substituents at a planar site are frozen (*e.g.*, guanidino groups). Each simulation consisted of 1.72×10^8 steps per replica for a total simulation time of ~1.6 μs per replica (10 fs time step). MC moves are described in the SI and consisted of various move types including dedicated moves for proline pucker angles (Radhakrishnan et al., 2012).

RESULTS

OUTLINE OF PROBLEM AND SOLUTION STRATEGY

As outlined in the Introduction, we ultimately want to push the limits of unbiased interpretability of input density data, *e.g.*, those coming from high-resolution EM experiments. This approach will in general have to employ a combination of physical potentials and experimental restraints. However, we first need to demonstrate that the restraints on the experimental density data work as expected. Therefore, this manuscript describes the performance of the density restraint potential in isolation. For this test to be meaningful, we use an equilibrium sampling protocol that is equally capable of supporting physical potentials. We note that the overall approach is general for spatial distributions in that it is not tailored toward a specific biopolymer type, toward a specific set of degrees of freedom, or even toward a specific type of input, *i.e.*, in eqs. 1 and 2, we could use any property of any spherical particle.

The most common way of dealing with ambiguous data is to incorporate auxiliary information into the

model determination, *e.g.*, the use of crystal or independently modeled structures of individual components in EM. We use synthetic data on the Arp2/3 complex as an example of this type as it has been employed in comparable work (Lasker et al., 2009). For fully flexible polymers, rather than using general physical potentials, it is common to employ database-derived information in several highly sophisticated and successful methodologies, most importantly the Rosetta-derived approaches (Adams et al., 2013). Conversely, we intentionally avoid a knowledge-based approach of this type. This in turn means that our flexible test systems necessarily have to be small to keep the complexity manageable.

The complexity of sampling flexible polymers in the presence of high-resolution density restraints is twofold: first, despite the absence of excluded volume terms, the potential energy surface is rugged, which means that trapping is likely to occur. Second, the search space is vast. We note that the challenges are comparable to those of simulations of reversible protein folding. Using both the RNA stem-loop and ubiquitin, we demonstrate that this challenge can be overcome for data at appropriate resolution. We show that convergence becomes increasingly difficult with increasing numbers of degrees of freedom and increasing resolution of the input data. Search efficiency is aided by advanced sampling methodologies, *i.e.*, a hybrid sampling protocol in conjunction with the REX method (Sugita and Okamoto, 1999) in the parameter space of the restraint potential (see SI). Neither technique limits our ability to interpret the generated trajectories as proper NVT ensembles.

Before presenting the results on the 3 systems, which are arranged by increasing difficulty, we want to emphasize that size and scope of test systems necessarily differs from that in recent refinement (Haddadian et al., 2011) or template-based remodeling efforts (Terwilliger et al., 2012). Instead, we hope to highlight the versatility of the approach by including data and systems at different scales and by using different biopolymers.

ARP2/3

The Arp2/3 complex is comprised of 7 polypeptide chains, and has been used a model system for fitting components of macromolecular assemblies into density data (Lasker et al., 2009). Low resolution,

synthetic density maps are created from the crystal structure (1TYQ) as describe above, and are shown in Fig. 1a. For starting conformations of the assembly simulations, we randomized the rigid-body coordinates of all 7 chains within the confines of the unit cell (see Fig. S2a). Individual components are kept rigid during simulations meaning that there are 42 degrees of freedom to consider.

The REX runs used variable restraint strengths, f_D. For high enough f_D, for all input resolutions, the native complex is found with very high statistical weight. This is asserted by clustering the data with a recent tree-based algorithm that is appropriate for this task because it can handle large data sets and produces tight, overlap-free clusters, whose centroids tend toward regions of high data density (Vitalis and Caflisch, 2012). The metric for clustering is the RMSD of 3 selected atoms for each domain (SI for details). This information is enough to describe the assembly entirely. We stress that structures are never subjected to alignment because the density restraints provide an absolute reference in space. The snapshots best approximating the centroids of the respective top clusters from the replicas with largest f_D are shown in Fig. 1b in comparison to the crystal structure. It is obvious that the differences are minor. This is despite the fact that these are snapshots from equilibrium sampling that have not undergone any kind of minimization. Fig. 1c-e corroborates this quantitatively. For each resolution, we plot as a function of f_D the weight of the native-like cluster (it is always the largest one) and the all-atom RMSD of its centroid snapshot to the crystal structure. These plots are illustrated by the corresponding centroids at the largest f_D values (same as in b) along with the input density.

In summary, the correct assembly is predicted unanimously and independent of resolution (within the range studied). The impact of lowering resolution is apparent by lower cluster weights and increasing RMSD values (Fig. 1c-e). This is what would be expected intuitively. Sampling does not appear to be an issue. We inspected the clustering results for all trajectories and found minor clusters with much larger RMSD values. These differ only in the rotation state of individual, spherical domains (see Fig. S2b).

RNA STEM-LOOP

The NMR ensemble of a 17-residue RNA stem-loop (2LBL) served as the input data for generating synthetic density data at high enough resolution for a flexible treatment of the RNA to be successful in the absence of auxiliary potentials. The inclusion of multiple conformations in map generation is an important departure from the Arp2/3 test case, because it allows for differences in contrast levels to appear for different parts of the molecule. The degrees of freedom are dominated by torsion angles in the polynucleotide backbone, but also include side chain torsions, rigid-body coordinates and pucker angles. Initial structures (Fig. S3a) are generated by randomizing all freely rotatable nucleic acid backbone torsions with all other sampled angles left at a fixed default value, usually 180º (pucker states are all initially in C2'-endo).

Utilizing REX runs with variable restraint strength, f_D, we find the native conformation with high statistical weight for both input resolutions (Fig. 2a) and for most values of f_D. This is based on a clustering using the RMSD of all heavy atoms (SI for details). The centroid snapshots reproduce the first structure in the NMR ensemble to well below 1 Å RMSD, and the same is true for all other members of the NMR ensemble (not shown). Interestingly, the weight of the native cluster is higher for lower resolution when f_D is large. Fig. 2b points out that this is a direct result of sampling convergence by plotting block-averaged restraint energies and RMSD values as a function of simulation progress. Clearly, convergence is more rapid for low resolution and does not seem to involve significant trapping. The enthalpic gap between misfolded and correctly folded structures also seems to be lowered considerably. Fig. 2c shows a direct comparison of the NMR ensemble to ensembles created from the centroid snapshots of the native-like clusters found for trajectories corresponding to individual replicas with sufficiently high f_D. Clearly, higher resolution produces a tighter ensemble that is visually difficult to distinguish from the NMR ensemble. An analysis of possible correlations between sequence-specific heterogeneities is found in Fig. S3b.

Examples of conformational traps hindering convergence are shown in Fig. 2d. These are all extracted as cluster centroids from the trajectory with the largest f_D within the high resolution run. The weights of

the clusters in question are in the range of 2 to 5 % each. Two of the traps have misfolded parts in the 5' end of the stem-loop, whereas the third one (rightmost structure) has a backbone registry that is shifted by one nucleotide. The nature of these traps indirectly highlights why the system is tractable as a fully flexible polymer, and this is because density patterns created by side chains and backbone, respectively, are very characteristic and easy to distinguish from one another. Lastly, Fig. 2e-f shows two-dimensional (2D) histograms for individual trajectories (29600 snapshots) at the highest f_D. These histograms (plotted logarithmically) make the point that there is a clear enthalpic gap between folded and misfolded conformations, and that the landscape is indeed rugged, more so for the case of high resolution input.

UBIQUITIN

Ubiquitin is a 76-residue protein with a reasonably complex α/β fold. Its structure has been determined by X-ray crystallography and the electron density is available as input (see Fig. S4 for a representation of both original and interpreted (eqs. 3 and S3) input densities). There are 3 major differences to the RNA stem-loop: 1) the polymer is much larger (roughly a factor 3 in the effective number of degrees of freedom); 2) the density is derived from an experimental measurement; 3) it is a polypeptide rather than a polynucleotide implying that backbone and side chain densities are harder to tell apart.

Starting structures are obtained by randomizing ϕ/ψ angles while leaving other dihedral angles at default values, usually 180º. Initial tests highlighted the difficulty in sampling this system, and as a result there are 4 identical REX simulations each using variable f_D and ω_t values. Moreover, the Hamiltonian is extended to include a weakly residue-specific, statistical potential applied to ϕ/ψ-angles (Fig. S1). This potential is meant to limit the search space and improve convergence rates, but no stringent test of its efficacy for this system could be performed. The impact is expected to be weak in that for those replicas included in the analysis the total energy correlates very strongly with the density restraint term, but not with any other term. Poor convergence was apparent upon visual inspection of

trajectories, and this prompted us to combine data from different replicas for analysis (details in SI).

Fig. 3a shows the logarithmic plot of a 2D histogram of RMSD values to the crystal structure and recomputed restraint energies for the combined trajectories. The overall histogram created from 28/48 replicas of every REX run highlights that there is good correlation between RMSD and the restraint energies, which were recomputed for all 1.12×10^6 snapshots to make data coming from different replicas comparable. To assess convergence quantitatively, it is best to compare data sets that are completely independent. This is why we next clustered the 4 REX runs separately using a set of atoms describing the entire chain (see SI). The largest clusters from each run span a certain range of RMSD and energy values, and these ranges (representing 90% of each cluster) are indicated as diamonds with solid lines. The same is done for the tightest cluster representing at least 0.25 % of the data if it is not the largest one (dashed diamonds). The circles highlight minimum energy structures from each run, and these structures are shown on the left of Fig. 3a. Taken together, these results suggest strongly that only the REX run in green samples the crystal structure.

Quantitative evidence for an overall lack of convergence is found in Fig. 3b where we plot time-dependent, block-averaged quantities (similar to Fig. 2b). These reveal two important results. First, over the simulation length considered here, stable plateau values are not reached for any of the REX runs. Second, there is a dramatic difference between the run sampling the crystal structure *vs*. those that do not in terms of the restraint energy. More so than Fig. 3a, this result emphasizes that the crystal structure corresponds to a deep, enthalpic minimum, which is expected given the resolution of the input data. However, the energy surface is so rugged that persistent trapping occurs as seen for the various, partial misfolds in Fig. 3a. The latter are analyzed in detail in Fig. 3c. Here, we plot subset RMSD values for 4-residue segments along the sequence for all snapshots highlighted in the histogram in Fig. 3a (clusters are represented by the snapshot best approximating the cluster centroid). As for the other systems, all snapshots are directly taken from the simulations (no minimization or refinement performed). Fig. 3c makes the point that all these snapshots are folded correctly in parts. Misfolding

appears to occur everywhere along the chain, but is generally more prominent toward the C-terminus. With the exception of the run shown in cyan, minimum energy snapshots appear similar to at least one of the cluster centroids shown.

There are at least two questions to ask. First, do varying contrast levels in the input density contribute to convergence issues? Second, what can be extracted from these data if the run shown in green is discarded? Fig. 4a shows a comparison of the crystallographic B-factors averaged over residues to root-mean square fluctuation (RMSF) data across a pseudo-ensemble created by gathering all the 9998 snapshots in the native-like cluster (green diamond in Fig. 3a). While the RMSF values in the N-terminal portion of ubiquitin are generally low, there appears to be good correlation for the C-terminal half. In particular the high variability of residues 61—64 and 71—76 appears to be present in both ensembles. This indicates that the poorer quality of the input data for these residues translates into more ambiguity in the simulations. This is confirmed by Fig. 4b where we plot cumulative histograms of RMSD values for individual 4-residue segments on data pooled from all runs except the one in green. These data highlight that the aforementioned segments are generally less likely to be placed correctly, presumably due to the higher contrast present in other parts of the density.

For the second question, we performed additional cluster analyses on coordinate subsets corresponding to the same 4-residue segments utilized in Figs. 3c and 4b. This is an example of a problem decomposition strategy at the data analysis level, *i.e.*, in theory we can also construct models by combining information from different snapshots. Fig. 4c shows the weights of the largest clusters for all 19 segments when analyzing the 3 relevant REX runs separately. These data are juxtaposed with RMSD values to the crystal structure for the corresponding segments and confirm the notion that the N-terminal side of the protein is more reliably folded than the C-terminal side. Importantly, there is a clear anticorrelation between statistical weight of the cluster and its RMSD. For example, for residues 45—56 the run shown in blue yields lower cluster weights than the other runs and is also the only one predicting incorrect placement of these residues. From the clusters for individual segments, we can also

directly assess their mutual compatibility. Here, we define a consensus assignment if all 3 centroid snapshots of the largest clusters for a given segment are within 1.5 Å of each other. The resulting partial structure contains 12 4-residue segments, all of which are placed correctly. This is illustrated in Fig. 4d by comparison to the crystal structure. Note that the fragment-based structure shown contains information from 12 *different* simulation snapshots taken directly from equilibrium sampling, and yet the covalent geometry appears quite reasonable throughout (even at the putative chain breaks). The bottom row of Fig. 4d shows that the majority of side chains are also placed correctly. Models of this type could be used as an incomplete template for partial remodeling, etc., but this is not explored here.

DISCUSSION AND CONCLUSIONS

The results presented in the prior section demonstrate that the density restraints combined with advanced sampling methodologies can solve problems of rather different nature, which is satisfactory as a proof-of-concept study. However, there are obvious limitations to the complexity of problems that will reliably yield solutions. Complexity is a result of the number of degrees of freedom to explore and the ruggedness incurred by high resolution input data. In this sense, despite its size, the Arp2/3 system is by far the simplest one considered here. The counterproductive nature of high resolution input is most clearly illustrated for the RNA stem-loop (Fig. 2b). This result appears intuitive if we place our simulations in the context of molecular simulations at equilibrium in the presence of physical interaction potentials, which face challenges that are at least as considerable (Smith et al., 2002).

As mentioned in the Introduction, our methodology is meant to eventually work as an *additional* term to the potential energy for systems that exhibit strong heterogeneity, cannot rely on database guidance, and/or offer input densities at insufficient resolution given the set of degrees of freedom to consider. Conversely, if the input data are of near-atomic resolution and/or on systems that can easily exploit knowledge-based approaches, then there are vastly superior methods available that rely on innovations

in search, optimization, and database utilization (Adams et al., 2013; Zhang et al., 2011). These make it possible to solve crystal structures by using, for example, strategies such as knowledge bias, automated pattern recognition, hierarchical (iterative) assignments, *etc*. The work of Perrakis and colleagues (Perrakis et al., 1997) provides a good example. We emphasize that we do not wish to present our method as a competitive alternative to the different families of highly elaborate, efficient, and well-established approaches to solving crystal structures of biomolecules. This is also consistent with the fact that the phase problem is not considered to any extent here.

While there are no fundamental departures in restraining real-space density information (Diamond, 1971), we believe that our approach offers novelty in dealing with an implicit background and in the translation of the input into physical quantities with contrast levels that are controllable by ω_t. One of the favorable properties of this approach is that the restraints yield densities that are physically meaningful. This is manifested, for example, by the low steric overlap seen for the Arp2/3 assembly (Fig. 1c). Here, the number of interchain clashes between heavy atoms with distances that are more than 0.5 Å too short (Hooft et al., 1996) is only 61 despite the complete absence of excluded volume terms. Disseminated within popular simulation software, the simulation approach of Trabuco *et al*. has been incorporated into other protocols that aim to refine crystal (Haddadian et al., 2011) or EM structures (Vashisth et al., 2012) with different advanced sampling/modeling approaches. However, the underlying restraint potential requires the presence of both physical interactions and additional bias terms, which is why it is not used in the generation of *ab initio* models. The power of the methodology we propose and explore here is illustrated by Movie S1, which shows how the REX technology in the parameter space of the restraint potential resolves trapping problems for the RNA example.

In conclusion, we have shown that both restraints and sampling protocols work as intended. We reemphasize that the starting structures are completely uninformed at the level of the set of degrees of freedom being sampled, distinguishing the method from refinement approaches. Simulation snapshots corresponding to cluster centroids (or combinations thereof) are of surprisingly high quality given that

no minimization or refinement are performed, neither at the level of physical potentials, nor for the density fit itself. We anticipate our approach to be useful in interpreting low-resolution density data, for which dedicated, knowledge-based approaches are unavailable, *e.g.*, high-resolution EM data on amyloid fibrils. Consequently, current research is concerned with using the approach in conjunction with physical potentials and with the formalization of consensus data analysis schemes as used in Fig. 4.

ACKNOWLEDGMENT

The authors want to thank Drs Marcus Fändrich, Nikolaus Grigorieff and Matthias Schmidt for many insightful conversations and the inspiration to undertake this project. We are grateful to two anonymous reviewers for valuable comments. A. V. acknowledges support from the “Holcim Stiftung Wissen”. This work was partially funded by a grant from the Swiss National Science Foundation to A. C.

REFERENCES

Adams, P.D., Baker, D., Brünger, A.T., Das, R., DiMaio, F., Read, R.J., Richardson, D.C., Richardson, J.S., and Terwilliger, T.C. (2013). Advances, interactions, and future developments in the CNS, Phenix, and Rosetta structural biology software systems. Annu. Rev. Biophys. *42*, 265–287.

Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R., and Schneider, B. (1992). The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J. *63*, 751–759.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. *4*, 187–217.

Brünger, A.T., Kuriyan, J., and Karplus, M. (1987). Crystallographic R factor refinement by molecular dynamics. Science *235*, 458–460.

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. J. Chem. Phys. *126,* 14101.

Chacon, P., and Wriggers, W. (2002). Multi-resolution contour-based fitting of macromolecular structures. J. Mol. Biol. *317,* 375–384.

Chang, A.T., and Nikonowicz, E.P. (2012). Solution nuclear magnetic resonance analyses of the anticodon arms of proteinogenic and nonproteinogenic tRNA(Gly). Biochemistry *51,* 3662–3674.

Chapman, M.S. (1995). Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron-density function. Acta Cryst. *A51,* 69–80.

Cheng, Y., and Walz, T. (2009). The advent of near-atomic resolution in single-particle electron microscopy. Annu. Rev. Biochem. *78,* 723–742.

DePristo, M.A., Bakker, P.I.W. de, and Blundell, T.L. (2004). Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure *12,* 831–838.

Diamond, R. (1971). A real-space refinement procedure for proteins. Acta Cryst. *A27,* 436–452.

Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., and Pedersen, L.G. (1995). A smooth particle mesh Ewald method. J. Chem. Phys. *103,* 8577–8593.

Fabiola, F., and Chapman, M.S. (2005). Fitting of high-resolution structures into electron microscopy reconstruction images. Structure *13,* 389–400.

Fändrich, M., Meinhardt, J., and Grigorieff, N. (2009). Structural polymorphism of Alzheimer Aβ and other amyloid fibrils. Prion *3,* 89–93.

Faulk, W.P., and Taylor, G.M. (1971). An immunocolloid method for the electron microscope. Immunochemistry *8,* 1081–1083.

Fujii, T., Iwane, A.H., Yanagida, T., and Namba, K. (2010). Direct visualization of secondary structures

of F-actin by electron cryomicroscopy. Nature *467,* 724–728.

Green, D.W., Ingram, V.M., and Perutz, M.F. (1954). The structure of haemoglobin IV. Sign determination by the isomorphous replacement method. Proc. R. Soc. Lond. A *225,* 287–307.

Haddadian, E.J., Gong, H., Jha, A.K., Yang, X., DeBartolo, J., Hinshaw, J.R., Rice, P.A., Sosnick, T.R., and Freed, K.F. (2011). Automated real-space refinement of protein structures using a realistic backbone move set. Biophys. J. *101,* 899–909.

Hooft, R.W.W., Vriend, G., Sander, C., and Abola, E.E. (1996). Errors in protein structures. Nature *381,* 272.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD - Visual molecular dynamics. J. Molec. Graphics *14*, 33–38.

Jack, A., and Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy and *R* factor. Acta Cryst. *A34,* 931–935.

Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wählby, A., and Jones, T.A. (2004). The Uppsala electron-density server. Acta Cryst. *D60,* 2240–2249.

Lasker, K., Topf, M., Sali, A., and Wolfson, H.J. (2009). Inferential optimization for simultaneous fitting of multiple components into a cryoEM map of their assembly. J. Mol. Biol. *388,* 180–194.

Marsh, J.A., Teichmann, S.A., and Forman-Kay, J.D. (2012). Probing the diverse landscape of protein flexibility and binding. Curr. Opin. Struct. Biol. *22,* 643–650.

Nolen, B.J., Littlefield, R.S., and Pollard, T.D. (2004). Crystal structures of actin-related protein 2/3 complex with bound ATP or ADP. Proc. Natl. Acad. Sci. USA *101,* 15627–15632.

Perrakis, A., Sixma, T.K., Wilson, K.S., and Lamzin, V.S. (1997). wARP: Improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. Acta

Cryst. *D53*, 448–455.

Petoukhov, M.V., Eady, N.A., Brown, K.A., and Svergun, D.I. (2002). Addition of missing loops and domains to protein models by X-ray solution scattering. Biophys. J. *83*, 3113–3125.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Radhakrishnan, A., Vitalis, A., Mao, A.H., Steffen, A.T., and Pappu, R.V. (2012). Improved atomistic Monte Carlo simulations demonstrate that poly-L-proline adopts heterogeneous ensembles of conformations of semi-rigid segments interrupted by kinks. J. Phys. Chem. B *116*, 6862–6871.

Robustelli, P., Kohlhoff, K.J., Cavalli, A., and Vendruscolo, M. (2010). Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. Structure *18*, 923–933.

Roseman, A.M. (2000). Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Cryst. *D56*, 1332–1340.

Smith, L.J., Daura, X., and van Gunsteren, W.F. (2002). Assessing equilibration and convergence in biomolecular simulations. Prot. Struct. Func. Bioinf. *48*, 487–496.

Sugita, Y., and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. *314*, 141–151.

Tama, F., Miyashita, O., and Brooks III, C.L. (2004). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. J. Mol. Biol. *337*, 985–999.

Terwilliger, T.C., DiMaio, F., Read, R.J., Baker, D., Bunkóczi, G., Adams, P.D., Grosse-Kunstleve, R.W., Afonine, P.V., and Echols, N. (2012). phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. J. Struct. Funct. Genomics *13*, 81–90.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryoEM density. Structure *16,* 295–307.

Trabuco, L., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure *16,* 673–683.

van Gunsteren, W.F., Dolenc, J., and Mark, A.E. (2008). Molecular simulation as an aid to experimentalists. Curr. Opin. Struct. Biol. *18,* 149–153.

Vashisth, H., Skiniotis, G., and Brooks III, C.L. (2012). Using enhanced sampling and structural restraints to refine atomic structures into low-resolution electron microscopy maps. Structure *20,* 1453–1462.

Vijay-Kumar, S., Bugg, C.E., and Cook, W.J. (1987). Structure of ubiquitin refined at 1.8 A resolution. J. Mol. Biol. *194,* 531–544.

Vitalis, A., and Caflisch, A. (2012). Efficient construction of mesostate networks from molecular dynamics trajectories. J. Chem. Theory Comput. *8,* 1108–1120.

Vitalis, A., and Pappu, R.V. (2009). ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. J. Comput. Chem. *30,* 673 -699.

Walzthoeni, T., Leitner, A., Stengel, F., and Aebersold, R. (2013). Mass spectrometry supported determination of protein complex structure. Curr. Opin. Struct. Biol. *23,* in press.

Wriggers, W., Milligan, R.A., and McCammon, J.A. (1999). Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. J. Struct. Biol. *125,* 185–195.

Zhang, J., Liang, Y., and Zhang, Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure *19,* 1784–1795.

FIGURE CAPTIONS

FIGURE 1: Results for the assembly of the Arp2/3 complex. **a**. Input maps were generated at different resolutions by using the atomic representations as shown (lattice with cubic cells of 3.0 Å side length). The resultant resolutions are illustrated by showing corresponding surfaces (with different threshold values) that enclose similar volumes (scale and orientation are identical for all 3 images). **b**. Centroids of the native-like clusters for runs at the highest restraint strength ($f_D = 0.16$) for all 3 resolutions are shown together with the crystal structure of Arp2/3 (PDB: 1TYQ). The latter is emphasized by using darker colors. General color coding is by chain. **c**. For the highest resolution, the centroid of the native-like cluster taken from the run with $f_D = 0.16$ is shown along with the input density at an enclosing surface value of $1.5\,\mathrm{g/cm^3}$ (transparent). In the bottom half of the panel, we plot the weight (number of snapshots in cluster divided by total number) of the native-like cluster and its all-atom RMSD from the crystal structure as a function of restraint strength. For restraint strengths lower than the ones shown, no clusters describing at least 1% of the data (10 snapshots) are found, and these trajectories are omitted in the plot. **d**. The same as **c** for intermediate resolution. **e**. The same as **c** for the lowest resolution. All illustrations were rendered with UCSF Chimera (Pettersen et al., 2004).

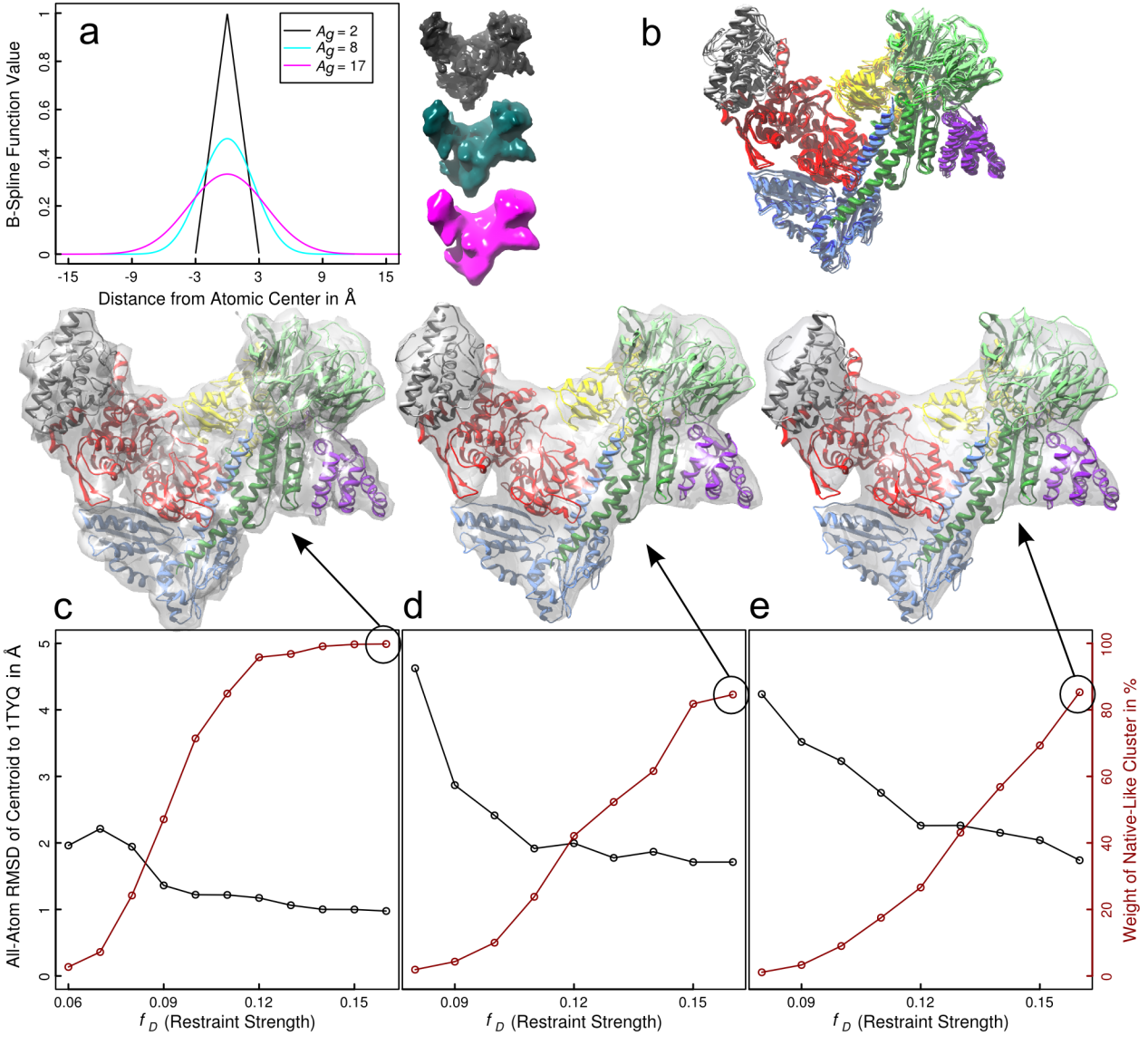
FIGURE 2: Results for the RNA stem-loop (PDB: 2LBL). **a**. For two different resolutions of the input map (distinguished by line type) and several restraint values, the weight of the native-like cluster and the heavy atom RMSD of its centroid to the first model in 2LBL are shown. **b**. For both resolutions of the input map, the trajectories for the strongest restraint condition ($f_D = 0.17$) are partitioned into 100 blocks, and time-dependent averages over individual blocks are plotted for both V_D and the heavy atom RMSD to 2LBL (#1). **c**. NMR and derived ensembles for the stem-loop are shown colored according to the Nucleic Acid Database Atlas convention (Berman et al., 1992). The left and right images show the 10 centroids of the native-like cluster for trajectories with values of f_D from 0.08 to 0.17 for low and high resolution of the input map, respectively. The corresponding input map at an enclosing surface

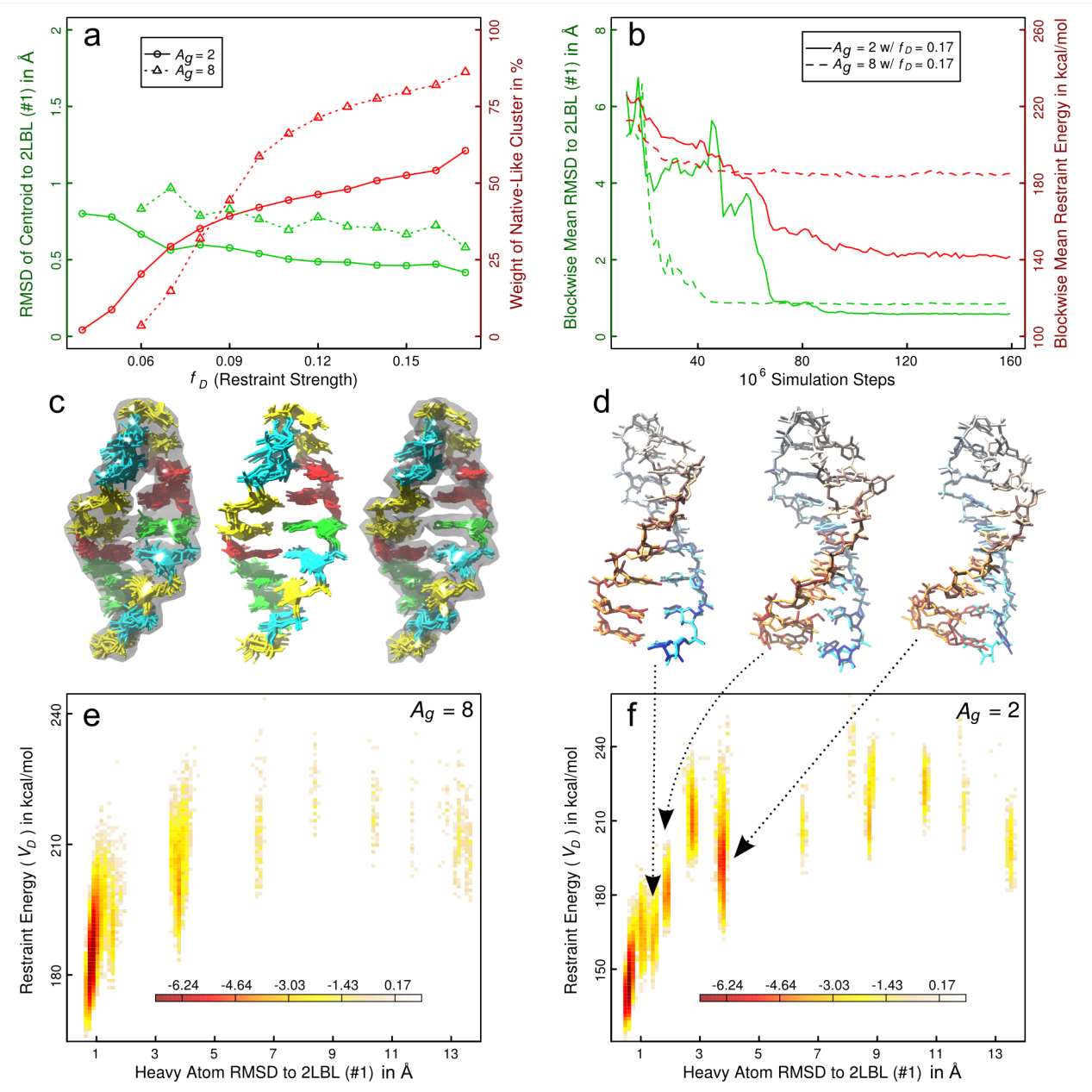
value of 1.71 g/cm^3 is overlaid as a transparent envelope. The NMR ensemble is depicted in the middle. **d**. Cluster centroids corresponding to traps encountered in the simulation with $f_D = 0.17$ at high input resolution are shown in comparison to the native-like cluster centroid. The latter uses lighter colors, and coloring proceeds from red (5' end) to blue (3' end). **e**. The negative logarithm of a 2D histogram of restraint energy and heavy atom RMSD for the trajectory with $f_D = 0.17$ at low input resolution. Points with no counts appear in white. The histogram is discontinuous because it is derived from a REX trajectory at a single condition. **f**. The same as **e** for higher input resolution. Arrows locate the traps depicted in **d** in the plot. Images in **c**-**d** were generated with UCSF Chimera.

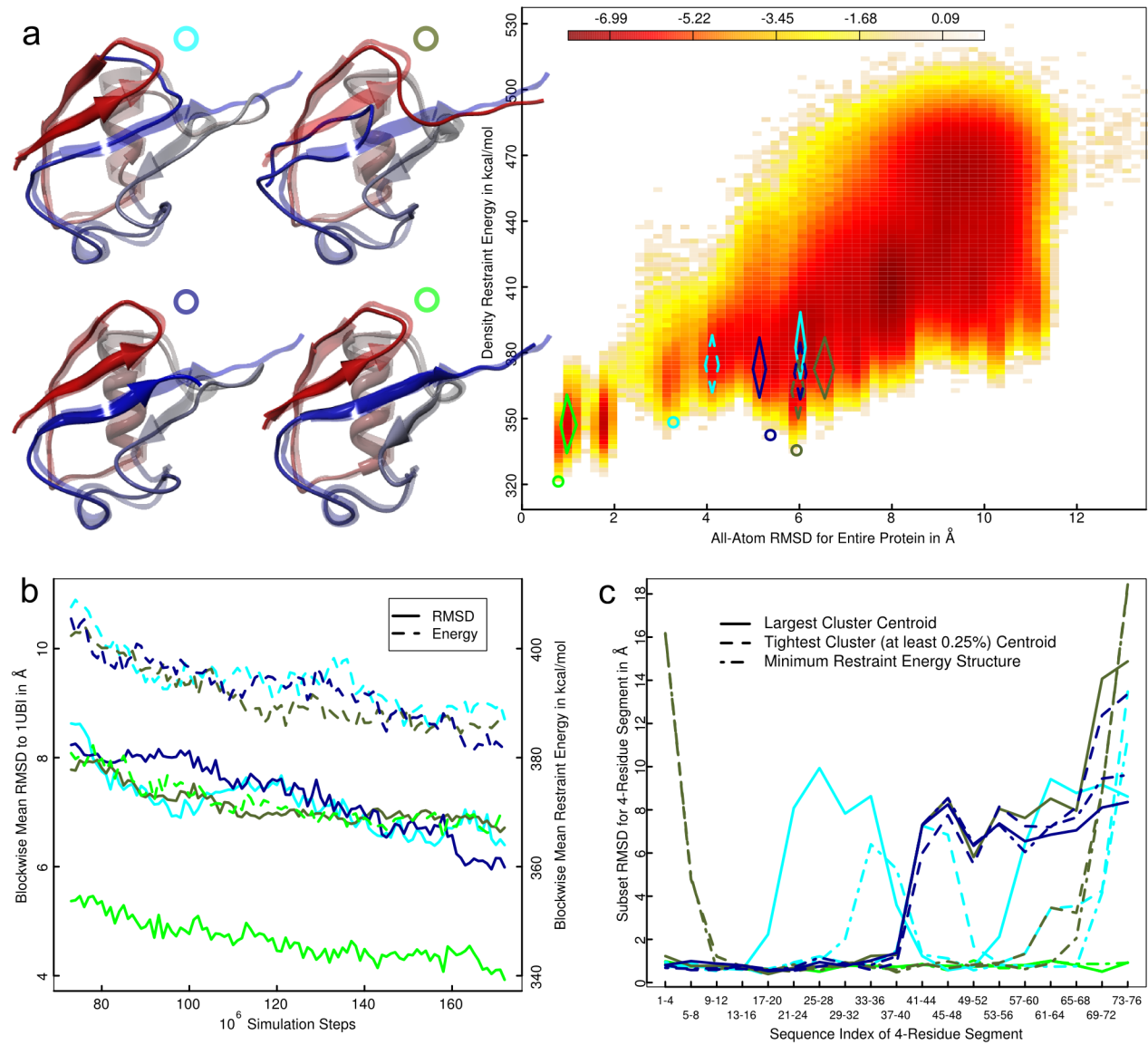
FIGURE 3: Results for ubiquitin (PDB: 1UBI). **a**. The negative logarithm of a combined 2D histogram of restraint energy and RMSD is plotted. Energy values are recomputed to make data from different replicas comparable ($f_D = 0.08$ and $\omega_t = -0.21$). RMSD computations use 620 of 746 atoms with those atoms creating artefactual deviations being excluded (see SI). Data are combined from all REX runs and all replicas with $f_D = 0.059$ or larger. Points with no counts appear in white. Circles highlight minimum restraint energy structures from each REX run (color-coded). Solid and dashed diamonds indicate the spread (90 %) of largest and tightest clusters (at least 0.25 % of data represented), respectively. The minimum energy structures are illustrated on the left along with the crystal structure (transparent). Color code is red to blue from N- to C-terminus, and images were generated with UCSF Chimera **b**. Average RMSD and recomputed restraint energies are shown as a function of simulation steps. Trajectories from individual replicas are partitioned into 100 blocks to give block averages. Blocks are then averaged across conditions for each REX run separately. The color code for the individual runs is the same as in **a**. **c**. For cluster centroid and minimum energy snapshots from all runs, subset RMSD values are plotted for 4-residue segments. The color code is the same as in **a**.

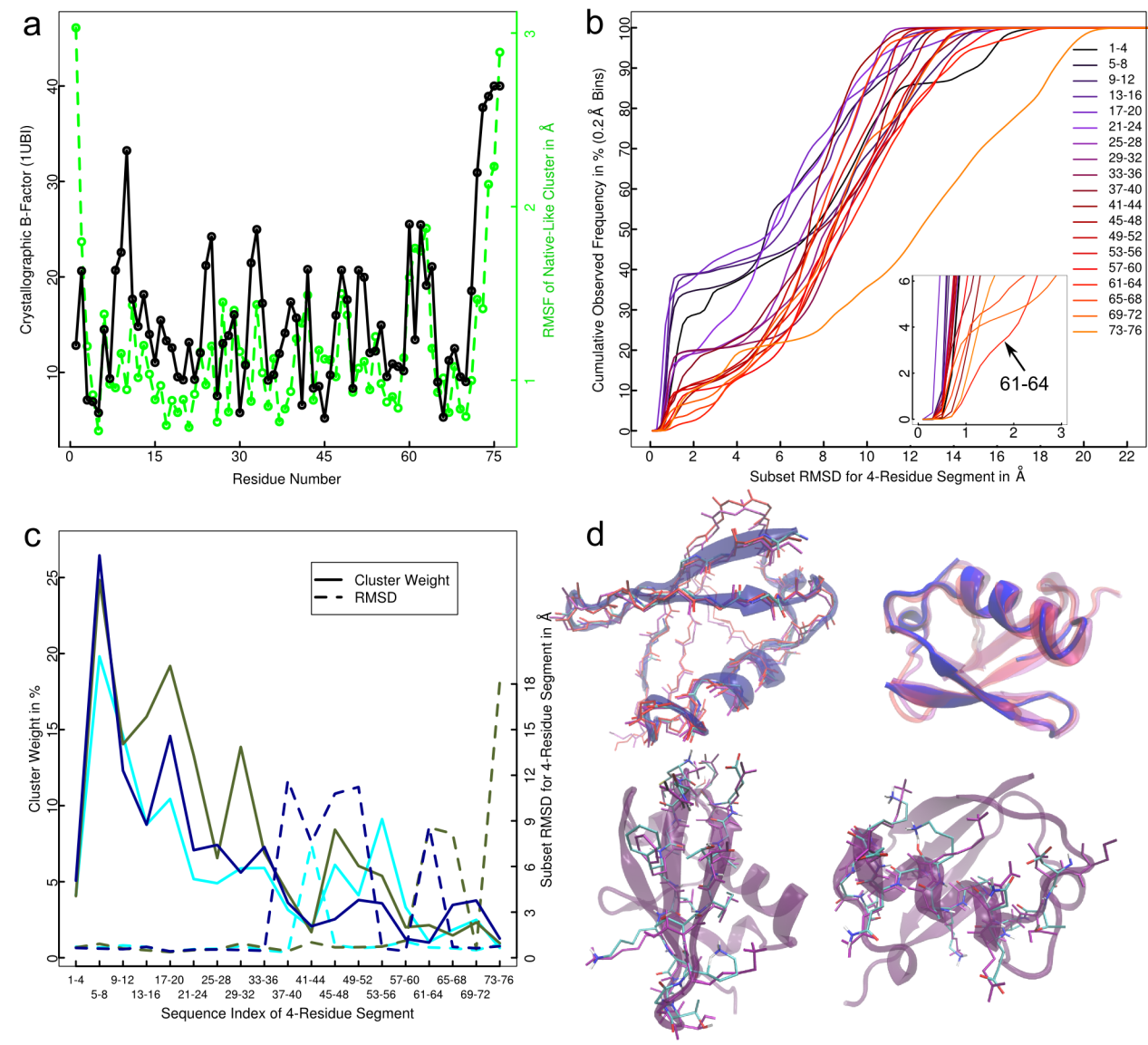
FIGURE 4: Sequence specificity for ubiquitin. **a**. Crystallographic B-factors averaged over residues are compared to RMSF values for a pseudo-ensemble of the snapshots constituting the native state. These all come from the run shown in green in Fig. 3. The data exclude ambiguous atoms (see SI). **b**.

Cumulative histograms of RMSD values for 4-residues segments are shown. Data from 3 runs were combined and all 840000 snapshots contribute to the histograms. All segments accumulate significant density in the 0.0 to 1.5 Å regime, and this is shown more clearly in the inset. **c**. The statistical weights of the largest clusters from an analysis of 4-residue segments are plotted along with the RMSD of the centroid snapshot to the corresponding segment in the crystal structure. The color code is the same as in Fig. 3a. **d**. The top left shows backbone stick representations for crystal structure (purple), the global minimum energy structure (red, see Fig. 3a), and a hybrid model constructed from consensus fragments (see **c**). A cartoon representation is added for the fragment model. The top right shows the same from a different angle just as cartoons. The bottom row highlights the N-terminal hairpin (left) and main helix (right) by showing a comparison of the crystal structure (purple) to the fragment model (colored by type) for all heavy atoms. A cartoon representation for the crystal structure is overlaid for orientation. Images were generated with VMD (Humphrey et al., 1996).


a
Ag = 2
Ag = 8
Ag = 17
B-Spline Function Value
Distance from Atomic Center in Å
b
c
d
e
All-Atom RMSD of Centroid to 1TYQ in Å
Weight of Native-Like Cluster in %
f D (Restraint Strength)


a
A_g = 2
A_g = 8
RMSD of Centroid to 2LBL (#1) in Å
Weight of Native-Like Cluster in %
f_D (Restraint Strength)
b
A_g = 2 w/ f_D = 0.17
A_g = 8 w/ f_D = 0.17
Blockwise Mean RMSD to 2LBL (#1) in Å
Blockwise Mean Restraint Energy in kcal/mol
10^6 Simulation Steps
c
d
e
A_g = 8
Restraint Energy (V_D) in kcal/mol
-6.24 -4.64 -3.03 -1.43 0.17
Heavy Atom RMSD to 2LBL (#1) in Å
f
A_g = 2
Restraint Energy (V_D) in kcal/mol
-6.24 -4.64 -3.03 -1.43 0.17
Heavy Atom RMSD to 2LBL (#1) in Å


a
-6.99
-5.22
-3.45
-1.68
0.09
Density Restraint Energy in kcal/mol
All-Atom RMSD for Entire Protein in Å
b
RMSD
Energy
Blockwise Mean RMSD to 1UBI in Å
Blockwise Mean Restraint Energy in kcal/mol
10^6 Simulation Steps
c
Largest Cluster Centroid
Tightest Cluster (at least 0.25%) Centroid
Minimum Restraint Energy Structure
Subset RMSD for 4-Residue Segment in Å
Sequence Index of 4-Residue Segment


a
Crystallographic B-Factor (1UBI)
RMSF of Native-Like Cluster in Å
Residue Number
b
Cumulative Observed Frequency in % (0.2 Å Bins)
Subset RMSD for 4-Residue Segment in Å
1-4
5-8
9-12
13-16
17-20
21-24
25-28
29-32
33-36
37-40
41-44
45-48
49-52
53-56
57-60
61-64
65-68
69-72
73-76
61-64
c
Cluster Weight
RMSD
Cluster Weight in %
Subset RMSD for 4-Residue Segment in Å
Sequence Index of 4-Residue Segment
d