Modeling Structural Heterogeneity in Proteins From X-Ray Data

Ankur Dhanik[1], Henry van den Bedem[2], Ashley Deacon[2], and Jean Claude Latombe[1]

[1] Computer Science Department, Stanford University {ankurd}@stanford.edu
[2] Joint Center for Structural Genomics, SLAC {vdbedem}@slac.stanford.edu

Abstract: In a crystallographic experiment, a protein is precipitated to obtain a crystalline sample (crystal) containing many copies of the molecule. An electron density map (EDM) is calculated from diffraction images obtained from focusing X-rays through the sample at different angles. This involves iterative phase determination and density calculation. The protein conformation is modeled by placing the atoms in 3-D space to best match the electron density. In practice, the copies of a protein in a crystal are not exactly in the same conformation. Consequently the obtained EDM, which corresponds to the cumulative distribution of atomic positions over all conformations, is blurred. Existing modeling methods compute an "average" protein conformation by maximizing its fit with the EDM and explain structural heterogeneity in the crystal with a harmonic distribution of the position of each atom. However, proteins undergo coordinated conformational variations leading to substantial correlated changes in atomic positions. These variations are biologically important. This paper presents a sample-select approach to model structural heterogeneity by computing an ensemble of conformations (along with occupancies) that, collectively, provide a near-optimal explanation of the EDM. The focus is on deformable protein fragments, mainly loops and side-chains. Tests were successfully conducted on simulated and experimental EDMs.

1 Introduction

Proteins are not rigid molecules [12, 19]. Each atom is subject to small, temperature-dependent high-frequency vibrations about its equilibrium position. In addition, in its native state, a protein may also undergo coordinated lower-frequency conformational variations leading to correlated changes in atomic coordinates. Such diffusive motions are of vital interest in the study of the protein's biological functions [29]. Accurately capturing such low-frequency protein dynamics from X-ray crystallography data has remained a challenge.

In a crystallographic experiment, a protein of known sequence is precipitated to obtain a crystalline sample (hereafter called a crystal) containing

many copies of the molecule. A three-dimensional electron density map (EDM) is calculated from a set of diffraction images, obtained from focusing X-rays through the sample at different angles. This EDM is an array of voxels, each encoding an electron density. The protein conformation is then modeled by placing the atoms in 3-D space to best match the electron density [11].

In an ideal crystal, all copies of the precipitated protein would have the same conformation. In practice, this is not the case, and corresponding atoms in different cells of a crystal do not occupy exactly the same position. The resulting EDM corresponds to the cumulative distribution of atomic positions over all conformations in the crystal. For instance, an EDM may appear locally blurred when a fragment of the main-chain or a side-chain adopts two or more neatly distinct conformational states (also called *conformers*). To illustrate, Figure 1 shows an isosurface of an EDM corresponding to a fragment (residues 104-112) of the protein with Protein Data Bank (PDB, [3]) ID 2R4I that occurs in two conformers. Extracting conformers from a locally disordered EDM is then akin to gleaning structure from a 3D image blurred by motion of the articulated subject.

Uncertainty in atomic positions is usually modeled with an *isotropic* Gaussian distribution. This model, further parameterized by the *temperature factor*, accounts for small vibrations about each atom's equilibrium position. Fitting an anisotropic (trivariate) Gaussian function requires estimating 9 parameters per atom, which for the complete model typically exceeds the amount of data in the EDM [28]. A sparser parameterization involves partitioning the protein into rigid bodies undergoing independent equilibrium displacements [22]. Owing to their "equilibrium-displacement" nature, these models are unable to accurately describe distinct conformational substates, such as those caused by the low-frequency diffusive motion of the protein [16, 29].

The presence of distinct conformers in a crystal has been observed on many occasions [4, 27, 29] and the importance of accurately representing structural heterogeneity by an ensemble of conformers has long been recognized [2, 13, 29]. However, while several programs are available for automatically building a structural model into an EDM to a high degree of accuracy [9, 10, 15, 20, 23], these have been engineered towards building a single conformer at unit occupancy. They often leave ambiguous electron density due to correlated changes in atomic coordinates uninterpreted. Building a heterogeneous protein model then requires substantial manual effort by skilled crystallographers using interactive graphics program. In [7, 17, 25] single-conformer, approximate starting models are perturbed to generate a multi-conformer ensemble. However, each one of these conformers can be seen as a possible interpretation of the EDM. Together, they do not provide a *collective* interpretation of the EDM. Automatically *building* a heterogeneous model into an EDM is a formidable challenge, and any progress could have a major impact on the way protein models are stored in the PDB.

It is imperative to accurately represent the data from the earliest stages of model building [7, 16]. In a crystallography experiment, the phase angle of a

diffracted beam is lost. Only magnitudes are measured on the sensitive surface of the detector [11]. Phases are estimated and improved by building and interpreting successive EDMs using Maximum Likelihood (ML) algorithms [20, 24]. However, disregarding structural heterogeneity in the successive EDMs or omitting fragments from a model altogether bias the phases in this procedure. Providing an ensemble of atom coordinates as initial values to ML algorithms could lead to improving the EDM more quickly.

In this paper, we present a new approach to automatically and accurately model heterogeneity in an EDM. Our main contribution lies in abandoning the single-conformer model in favor of a multi-conformer model where appropriate, and providing an estimate of the relative frequency of occurrence (called *occupancy*) in the crystal for each of the conformers. We focus on protein fragments, mostly loops, which are often the most deformable substructures in a protein [7, 25, 26]. Our method computes an occupancy-weighted *ensemble* of conformations that *collectively* best represents the input EDM. To this end, an idealized protein fragment is modeled as a kinematic linkage, with fixed groups of atoms as links and rotatable bonds as joints. Our method is based on a sample-select protocol, which adaptively alternates *sampling* and *selection* steps. Each sampling step generates a very large set of {conformation, temperature factors} samples. A subsequent selection step applies an efficient linear-programming algorithm to concurrently fit this set of samples to the input EDM and compute the occupancy of each sample. Samples with small occupancies (less than 0.1) are then discarded. As the sampled space has very high dimensionality, the successive sampling steps consider portions of the protein fragment of increasing lengths. The overall sampling process is guided by the results obtained at previous selection steps. It should be emphasized that the algorithm *infers* the ensemble size from the data; it has no prior knowledge about the number of conformers.

This paper is divided into two main sections. In Section 2 we present and discuss results obtained with our method on both simulated and experimental EDMs. This section allows us to characterize more precisely the type of problem addressed in the paper. In Section 3 we describe in detail our sample-select method to model heterogeneity in an EDM.

2 Results and Discussion

Validation tests against simulated EDMs in Section 2.1 demonstrate that our method extracts the correct ensemble for a variety of fragment lengths over a range of resolution levels, noise levels, occupancies and temperature factors. They furthermore show that the algorithm correctly identifies and models side-chains in multiple conformations.

We also tested our method against experimental EDMs. In Section 2.2, we show results obtained with the 398-residue Flavoprotein TM0755 (PDB ID 1VME). The main chain for residues A316-A325 is bi-modally disordered [27].

Our method models the two conformations to within 0.6Å RMSD[3]. In Section 2.3 we give additional results on experimental EDMs for side-chains in multiple conformations.

Depending on the length of the fragment to compute and the resolution of the EDM, our partly parallelized implementation takes 2-4 hours to complete.

2.1 Algorithm Validation with Simulated Data

Given a protein structure, the simulated EDM corresponding to its distribution of atoms can easily be calculated at different resolution levels while controlling the temperature factors and occupancy of individual atoms. Such a simulated EDM allows us to test our method, and understand the effects of experimental noise and discrete sampling with idealized geometry.

Table 1. Single conformer results from validation tests using simulated data. Each row lists PDB ID, map resolution (Res, in Å), anchors and size of loop, average temperature factor of loop atoms in the PDB structure ($\bar{B}$obs, in Å^2), RMSD of calculated conformation to PDB conformation, and average temperature factor of calculated conformation ($\bar{B}$calc, in Å^2). All calculated occupancies sum to 1.0. The final column identifies a side-chain in dual conformation at 0.5/0.5 occupancy.

PDB ID	Res	Loop(size)	$\bar{B}$obs	RMSD	$\bar{B}$calc	side-chain
1AAJ	1.9	82-85(4)	7.9	0.19 0.22	12.9	GLU(84)
1BGC	2.3	40-43(4)	29.5	0.32 0.33 0.41	27.1	LYS(41)
1HFC	1.9	142-149(8)	11.7	0.29 0.31 0.37 0.42 0.44 0.71	10.1	LEU(147)
1Z8H	2.3	71-78(8)	40.9	0.35 0.39 0.56 0.61 0.69 0.69 0.77	36.0	HIS(73)
1TML	1.9	243-254(12)	11.5	0.41 0.41 0.43 0.44 0.55 0.95	12.7	THR(247)
1CTM	2.3	142-149(12)	38.8	0.36 0.44 0.44 0.52 0.68 0.68	34.8	ARG(18)

Single Conformer

We first validated our algorithm on simulated data corresponding to single-conformer fragments, computed at various resolution levels. Six fragments varying in length from 4 to 12 residues at various temperature factors were selected from the PDB. The algorithm consistently identified conformers in the simulated EDM within 0.4Å RMSD of the true conformers (see Table 1). In each case, the returned ensemble contains more than one conformation. However, all conformations in an ensemble are pairwise very close and could easily be merged in a post-processing step by a clustering algorithm. Multiple

[3] Unless otherwise noted, RMSD denotes the square root of the averaged squared distances between corresponding N, C_α, C_β, C, and O atoms.

conformations are returned due to finite resolution in our sampling scheme as described in Section 3. To confirm that, we ran the same test again, but this time we added the true conformers to the sample set. Then for each EDM, our method returned only the true conformer.

The algorithm furthermore returns temperature factors to within a 10.0Å^2 interval of the average, true temperature factors. These temperature factors and coordinate errors are well within the radius of convergence of standard crystallographic refinement packages.

Side-chains commonly occur in multiple rotameric conformations in protein structures determined from X-ray data. To test if the algorithm correctly models side-chains in multiple rotameric conformations while the backbone is best represented by a single conformer, a second rotamer was added to a selected side-chain of each main-chain at 0.5/0.5 occupancy (see the last column of Table 1). The main-chain RMSDs differed not meaningfully from those found earlier, while all dual rotamers were identified at the correct occupancy and within 0.5Å RMSD.

Dual Conformers

The 123-residue protein with PDB ID 2R4I, a NTF-3 like protein, was solved by the Joint Center for Structural Genomics (JCSG) at a resolution of 1.6Å. The asymmetric unit contained four, nearly identical copies of the molecule, distinguished by chain identifiers A-D in the PDB file. In each of the four chains the fragment spanning the residues 104-112 crystallized in slightly different conformations. We added the atoms from residues 104-112 from chain A to the corresponding residues from chain B (Figure 1). Indeed, the fragment can presumably adopt both of these states. The conformers are closely intertwined, separated by only 1.4Å RMSD.

Simulated electron density data for the dual conformer was generated at different resolutions and at various occupancies. Gaussian noise with a standard deviation of 10% of the magnitude of the calculated data was added to simulate experimental errors. The temperature factors of the individual PDB structures were retained, averaging 19.0Å^2.

The algorithm returns an ensemble in excellent agreement with the actual conformations, with a good estimate of the true occupancy values and average temperature factors (see Table 2). Again the finite discretization of our sampling scheme results in ensembles that contain more than two conformations. But every returned ensemble contains two groups of very similar conformations that could be merged by a clustering algorithm.

We ran the same test again, but this time we added the true conformers to the sample set. The results presented in Table 3 show that in most cases our method returns the true conformers. In some cases, it produced more than two conformers and in all cases occupancies and temperature factors are slightly inexact. These small discrepancies seem to be caused by the Gaussian noise added to the EDM. The greater discrepancy in the results presented in Table 2

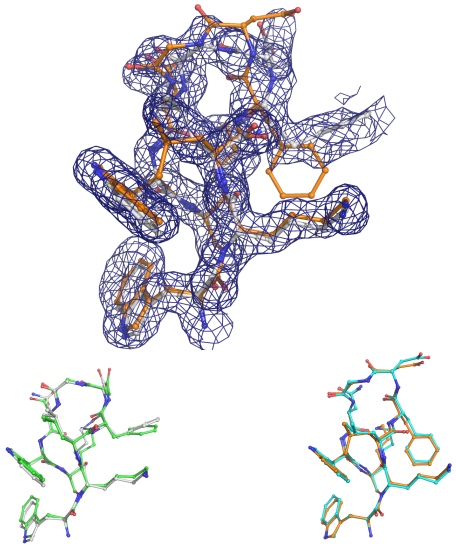


Fig. 1. Residues 104-112 of 2R4I. Top panel: Conformations from chain A and B in the EDM at 0.7/0.3 occupancy. At high contour levels, atoms from the chain at lower occupancy are no longer contained within the iso-surface. Lower panel: PDB fragment from chain A (left) in green and PDB fragment from chain B (right) in cyan together with the calculated conformers.

are, thus, manifestations of both discretization errors and errors due to added noise.

Furthermore, coordinate error is larger for the lower occupancy conformer. It should be noted that at an occcupancy of 0.3, a Carbon atom only scatters at about twice the magnitude of a Hydrogen atom. The signal of a Hydrogen atom is distinguished from the background level only at resolution levels better than 1.3 Å(i.e. $<$ 1.3 Å). At resolution levels considered here, Hydrogens are not explicitly included in PDB files.

Ensemble of Conformations

Solvent exposed fragments may have only weakly preferred substates. This common situation is often characterized by a blurring of the main-chain EDM and ambiguous or weak side-chain density. To emulate this situation, a collection of 20 conformations of the 8-residue loop 142-149 of 1HFC (Table 1) was generated along a coordinated motion of the loop (as shown in Figure 2(a)). The start and finish conformations are 2.7Å apart in RMSD. A 1.9Å EDM was calculated at equal occupancy (0.05) for the members of the collection.

The algorithm returned a 7-conformer ensemble, with occupancies ranging from 0.10 to 0.23. Since the algorithm only retains conformations with calculated occupancy greater than or equal to 0.1, it could not return the

Table 2. Details of calculated dual conformers for loop 104-112 of 2R4I. Each row lists occupancies for the conformers (Occ), map resolution (Res, in Å), RMSD of calculated conformers to PDB conformers, the cumulative calculated occupancies for the conformers (Calc Occ), and average temperature factor of calculated conformers ($\bar{B}$calc, in Å^2). Average, observed temperature factors are 19.0Å^2.

Occ	Res	RMSD	Calc Occ	$\bar{B}$calc
0.5/0.5	1.3	0.26 0.34	0.29	24.3
		0.38 0.64 0.77 1.29	0.71	
0.5/0.5	1.5	0.32 0.64	0.36	27.4
		0.38 0.49 0.52 0.69	0.64	
0.5/0.5	1.7	0.29 0.42 0.42 0.43	0.50	25.5
		0.23 0.23 0.34	0.50	
0.6/0.4	1.3	0.29 0.40 0.64	0.53	25.5
		0.31 0.35 0.64	0.47	
0.6/0.4	1.5	0.30 0.44 0.54 0.58	0.61	25.7
		0.33 0.62	0.39	
0.6/0.4	1.7	0.23 0.24 0.35 0.60	0.58	22.0
		0.23 0.62	0.41	
0.7/0.3	1.3	0.27 0.34 0.34 0.34 0.64	0.70	24.1
		0.48	0.30	
0.7/0.3	1.5	0.33 0.33 0.40	0.64	29.0
		0.61 0.76	0.36	
0.7/0.3	1.7	0.31 0.37 0.42 0.47	0.65	21.7
		0.44 0.62	0.35	

full collection of 20 conformations. Nevertheless, it successfully extracted the range of motion from the data (see Figure 2(b)).

We furthermore applied the loop-fitting option of RESOLVE (v2.10) [23], a widely-used crystallographic model-building algorithm, to this EDM. RESOLVE, unable to assign residue identities and side-chains, modeled a single poly-alanine loop into the EDM, shown in yellow in Figure 2(b). Analysis of the results reveals that occupancy-weighted main-chain EDM correlation coefficients of the ensemble range from 0.84 to 0.96 per residue versus 0.64 to 0.87 for the single poly-alanine conformer. Moreover, the average, occupancy-weighted temperature factor of the ensemble (15.0Å^2) is closer to the average of the 20 conformations (11.7Å^2) than the single, poly-alanine chain (35.7Å^2). Thus, our 7-conformer ensemble is a significantly improved interpretation of the data, both quantitatively and qualitatively (range of motion), over a single, averaged conformer.

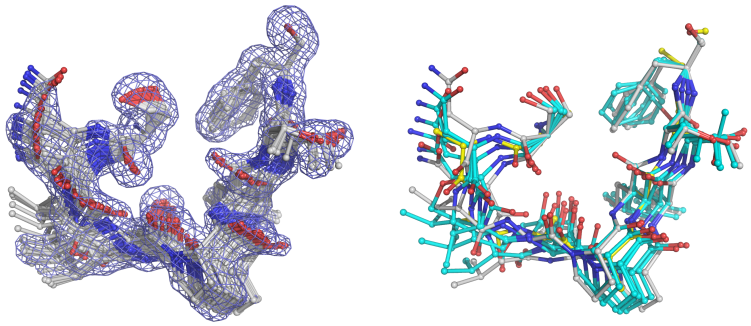
Table 3. Details of calculated dual conformers for loop 104-112 of 2R4I. The true conformers were added in the sampling protocol. Each row lists occupancies for the conformers (Occ), map resolution (Res, in Å), RMSD of calculated conformers to PDB conformers, the cumulative calculated occupancies for the conformers (Calc Occ), and average temperature factor of calculated conformers ($\bar{B}$calc, in Å^2). Average, observed temperature factors are 19.0Å^2.

Occ	Res	RMSD	Calc Occ	$\bar{B}$ calc
0.5/0.5	1.3	0.00	0.47	25.2
		0.00	0.53	
0.5/0.5	1.5	0.00	0.48	22.3
		0.00	0.52	
0.5/0.5	1.7	0.00 0.29	0.49	22.0
		0.00 0.23	0.51	
0.6/0.4	1.3	0.00	0.56	20.1
		0.00	0.44	
0.6/0.4	1.5	0.00 0.54	0.61	24.7
		0.00	0.39	
0.6/0.4	1.7	0.00	0.57	23.4
		0.00	0.43	
0.7/0.3	1.3	0.00	0.65	25.9
		0.00	0.35	
0.7/0.3	1.5	0.00 0.33 0.33	0.66	29.0
		0.00 0.61	0.34	
0.7/0.3	1.7	0.00	0.65	20.8
		0.00	0.35	

This example suggests that in general the returned ensemble of conformations should not be treated as a true physical model of the actual heterogeneity present in the crystal, but as a representation of uncertainty in atomic positions due (in part) by this heterogeneity.

2.2 Experimental data: Modeling a Dual Conformer

A structural model for TM0755 was obtained by the JCSG from data at 1.8Å resolution. The asymmetric unit contains a dimer, with a short main-chain fragment around residue A320, and the same fragment around B320, bimodally disordered. Crystallographers had initially abandoned this fragment due to difficulty interpreting the EDM visually. A dual conformation for the fragment A316-A325, separated by 2.96Å, was obtained from semi-automated methods at 0.5/0.5 occupancy [27]. The average, occupancy-weighted temperature factor was 24.9Å^2. The structure together with the heterogeneous



(a) (b)

Fig. 2. (a) A collection of 20 snapshots of an 8 residue loop while it is transitioning between simulated start and finish conformations. (b) Ensemble of 7 conformers computed by the algorithm (cyan), together with the start and finish conformations (grey) of the simulated collection. A single conformer modeled by RESOLVE is displayed in yellow.

fragment was refined, subjected to the JCSG's quality control protocol (unpublished) and ultimately deposited in the PDB.

An experimental electron density map was calculated from diffraction images with σ_A-weigthed $2mF_o - DF_c$ coefficients [21]. Our algorithm returned a 5-conformer ensemble. Two conformations in the ensemble are 0.47 and 1.24 Å RMSD away from one of the conformations obtained at JCSG with occupancies 0.15 and 0.23. The other three calculated conformers are 0.64, 0.72, and 0.82Å RMSD away from the other conformation obtained at JCSG with occupancies 0.27, 0.23, and 0.12 respectively. The average, occupancy-weighted temperature factor of the ensemble is 30.3Å^2.

This result demonstrates that our method is also highly effective with experimental data which, in contrast to data with simulated measurement errors, may contain substantial phase angle errors. Automatic identification of multi-conformers will greatly enhance the structure determination process.

2.3 Experimental data: Modeling Alternate Side Chain Conformations

Locally, side-chains too regularly adopt alternate conformations, accomodated by subtle changes in the main-chain [6]. At present, modeling alternate side-chains onto a known main-chain in the final stages of the structure determination process is time-consuming and subject to individual preferences. To assess the value of our algorithm for a high-throughput structure determination pipeline such as the JCSG's, it was modified to model alternate side-chain conformations onto a known main-chain. At a fixed position in the protein chain, trial positions for the C_β atom are generated, and the entire residue is repositioned by adjusting flanking dihedral angles. For each trial C_β position, neighborhoods of rotamers are sampled to obtain a large set of candidate con-

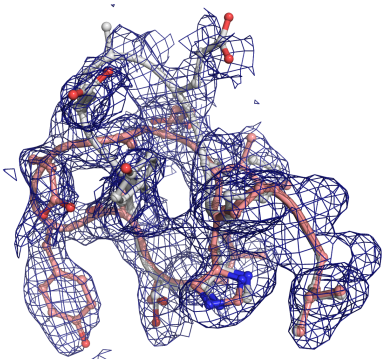


Fig. 3. Two conformations from the 5-conformer ensemble computed for the fragment A316-A325 in the experimental EDM. The conformers deviate by 0.47 Å RMSD and 0.64 Å RMSD from the conformations obtained by the JCSG. Alternate conformers are difficult if not impossible to identify and model visually in ambiguous electron density. For clarity, the main-chain is represented by a cartoon in the figure.

formations. This set is then subjected to a selection step to obtain occupancy values. Finally, the coordinates are refined with a standard crystallographic refinement suite [1].

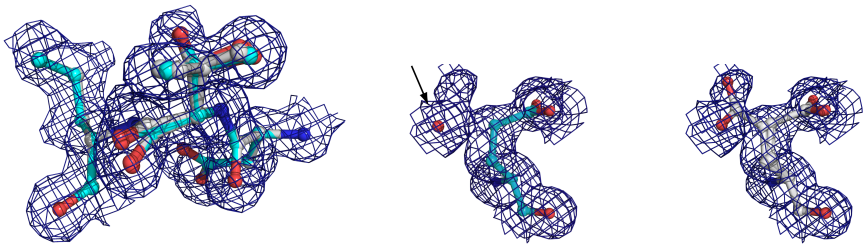
A structural model for Xisl protein-like solved to 1.3 Å resolution (PDB ID 2NLV) was used to test the procedure. The protein is 112 residues in length, and was deposited in the PDB with 20 residues of the A-chain in alternate conformations. The algorithm successfully identified and modeled 85% of residues with alternate conformations, see Figure 4. The side-chain conformations that were not found were outside the sample set. Additionally, 12 multi-conformer alternatives for single-conformer residues were identified for which the data fit improved substantially, see Figure 4.

3 Method

Our goal is to compute an ensemble of conformations, the occupancy of each conformation, and the temperature factor of each atom in every conformation, that together optimally represent the data in an input EDM E of a protein fragment.

One approach – let us call it initialize-optimize – consists of formulating this problem as an optimization problem:

1. Pick an ensemble of k initial conformations, along with their occupancies and temperature factors.
2. Compute the simulated EDM that corresponds to this ensemble.
3. Iteratively modify the k conformations, their occupancies, and the temperature factors to minimize the difference between the experimental and the simulated one EDMs.



(a) (b) (c)

Fig. 4. (a) Residue 36THR from the A chain in 2NLV. The PDB model is shown in cyan (2 conformations at 0.5/0.5 occupancy), and our model is shown in grey (3 conformations at 0.31, 0.35, and 0.36 occupancy). Note that the carbonyl oxygen shifts considerably to accomodate an alternate conformation. The side-chain EDM correlation coefficient improved from 0.77 to 0.81. (b) Residue 81GLU as modeled in the PDB conformation, and (c) as modeled by our algorithm. Observe that the PDB conformation mistakingly modeled a water molecule at full occupancy at the position of a carboxyl oxygen, a common mistake. Our alternate side-chain is modeled at 0.33 occupancy.

We actually tried this approach, but even on dual-conformer examples the number of parameters to optimize is huge and the optimization process (Step 3) gets easily trapped into local minima. Monte-Carlo methods with simulated annealing protocols were unable to handle these issues.

This led us to develop a completely different approach, which we call sample-select. Instead of incrementally modifying conformations, we first sample a very large set of conformations and then select the best ensemble from this set. More precisely, our method alternates two steps, SAMPLE and SELECT:

1. SAMPLE samples a large set Q of conformations (and the temperature factors of the atoms in each conformation) that is highly likely to contain a subset S representing E well.
2. SELECT simultaneously identifies this subset S and computes the occupancy factor of each conformation in S.

The space sampled by SAMPLE has high dimensionality, so each run of SAMPLE uses the conformation subset selected at the previous iteration to sample a new set of candidate conformations, which in turn is submitted to SELECT. The core of our method is an efficient linear-programming algorithm that is able to select pertinent ensembles from very large sets of sampled conformations. We first describe this algorithm.

3.1 Selection step

SELECT is handed a large set $Q = \{q_1, \ldots, q_N\}$ of N conformations, together with a vector t_i specifying the temperature factor of each atom in every conformation q_i. It identifies the subset S of conformations that *collectively* provides the best explanation for the input EDM E, over all possible subsets of Q.

Let G be the grid over which E is defined. Let E_i be the simulated EDM that corresponds to the configuration q_i with the temperature factors in t_i. Let $E(p)$ and $E_i(p)$ denote the values of E and E_i, respectively, at point $p \in G$. The value at p of the EDM that corresponds to $Q = \{q_1, ..., q_N\}$ with occupancies $\alpha_1, \ldots, \alpha_N$ is $\sum_i \alpha_i E_i(p)$. SELECT minimizes the L_1 difference between E and this EDM. Since each $E_i(p)$ is constant, this amounts to solving the following linear problem (LP):

$$\begin{array}{ll} \text{Minimize} & \sum_{p \in G} \mid E(p) - \sum_i \alpha_i E_i(p) \mid \\ \text{such that} & \alpha_i \geq 0, \text{ for } i = 1, \ldots, N \\ & \sum_i \alpha_i = 1. \end{array}$$

The solution is the vector of optimal values for α_i, $i = 1, \ldots, N$. SELECT retains only the conformations q_i whose occupancies are greater than a given threshold (set to 0.1 in our implementation). It returns the set S of retained conformations with occupancies re-normalized to sum up to 1. We use Coin-OR libraries [14] to solve the above LP.

3.2 Conformation sampling

The goal of SAMPLE is to generate a set $Q = \{q_1, \ldots, q_N\}$ of candidate conformations, together with temperature-factor vectors $t_1, \ldots, t_N$, such that a subset S of Q (with suitable occupancies) provides an optimal explanation of the EDM E. Each SAMPLE step uses the outcome of the previous SELECT step and samples a distinct subspace of reasonably small dimensionality.

Let $n > 3$ be the number of residues in the fragment. We fix bond lengths and dihedral angles ω around peptide bonds to their canonical values. This leads us to treat the fragment as a kinematic linkage [8] whose degrees of freedoms are the dihedral ϕ and ψ angles around N-C_α and C_α-C bonds, the bond angles in the main chain, and the χ angles in the side-chains. We divide the fragment into a front and a back half, each with $p = \lceil \frac{n}{2} \rceil$ residues. We first incrementally build conformations of these two halves. Then, we connect them using an inverse kinematics (IK) algorithm.

We describe the various steps in more detail below. However, it should be noted that there are many possible variants, some of which might work equally well. The key idea is to consider fractions of the front and back halves of increasing size, so that the number of conformations sampled by each SAMPLE step can be handled by the next SELECT operation.

(a) Sampling the main chains of the two halves. Let us temporarily ignore side-chains and temperature factors. We incrementally build candidate partial conformations of the front half's main chain by sampling one ϕ or ψ angle at a time, starting from the N terminus. We sample the first ϕ angle at some uniform resolution ϵ (set to 2 degrees in our implementation). We also sample the bond angle centered at the N atom preceding this ϕ angle at the same resolution ϵ in the 12-degree interval around its corresponding Engh-Huber

value. We thus obtain a set of $6 \times 2\pi/\epsilon$ candidate positions for the following C_β and C atoms, on which we run SELECT. Let k_1 be the number of *partial* conformations retained by SELECT. Next, we sample in the same ways the following dihedral angle (a ψ angle) and the two following bond angles centered at C_α and C atoms. We thus get a set of $12 \times 2\pi/\epsilon \times k_1$ candidate positions for the following O, N, and C_α atoms. We run SELECT on this set and obtain an ensemble of size k_2.

At this point, we re-sample the two ϕ and ψ angles at a finer resolution (0.5 degrees) in small neighborhoods (± 1 degree) of their values in the ensemble of size k_2. This re-sampling step yields an expanded set of candidate conformations to which we apply SELECT. We proceed in the same way with the remaining $p - 1$ residues in the front half's main chain.

The same procedure is applied in reverse to the back half, starting from its C terminus.

(b) Inserting side-chains. Immediately after a pair of consecutive ϕ and ψ angles have been re-sampled, the side-chain of the residue containing those two angles is inserted. We use a rotamer library [18] to obtain the values of the χ angles. Adding the side-chain multiplies the number of partial conformations of the front half by the number of rotamers for the side-chain. We apply SELECT to this new set. The same procedure is applied to the back half.

(c) Assigning temperature factors. Temperature factors are assigned whenever a ϕ or ψ angle is sampled or a side-chain is inserted. Their values are taken from a finite set T input by the user. However, assigning a distinct temperature factor to every atom would quickly lead to large sets of candidate conformations. So, we define groups of atoms that are assigned the same temperature factors. The C_β and C atoms following a ϕ angle forms one group, so do the O, N, and C_α atoms following a ψ angle and the atoms in a side-chain.

Consider the case where we sample a ϕ angle in the front half. As described in paragraph (a), this gives a number of candidate conformations for the following C_β and C atoms. We pair each of these conformations with a distinct temperature factor from T. Similarly, when we insert a side-chain, we pair each rotamer with a distinct temperature factor from T.

(d) Connecting the front and back halves. We enumerate all pairs of conformations of the fragment's front and back halves computed as above. For each pair, complete closed conformations of the fragment's main chain are obtained by computing six dihedral angles using an analytical IK algorithm [5]. More precisely, for each pair, we consider every three consecutive residues such that at least one belongs to the front half and another one to the back half, and we re-compute the ϕ and ψ angles in those residues using the IK algorithm, so that the fragment's main chain gets perfectly closed. The side-chain conformations and temperature factors for each of these residues are set as in either the front or back half conformation.

We collect all the closed conformations into a candidate set, on which we run SELECT. The result is the final conformation ensemble built by our method. If desired, a clustering algorithm can be run on this ensemble to merge conformations that are pairwise very close.

The above method sometimes eliminates a pertinent partial conformation. This is due to the fact that partial conformations are retained based on their fit with only a subset of the EDM. So, a SELECT step might retain one conformation and discard another based on this *local* fit, while the inverse result could have been obtained if larger fractions of the fragment had been considered. Unfortunately, when a pertinent partial conformation has been discarded, it cannot be recovered later. So, to reduce the risk of eliminating a pertinent partial conformation, we retain a greater number of partial conformations at each selection step. This is done as follows. Let m be the size of the set of conformations given to a selection step and m' the size of the ensemble retained by SELECT. We run SELECT again on the remaining $m - m'$ conformations, and we repeat this operation until a pre-specified number of conformations have been obtained.

4 Conclusion

This paper presents a new method to model structural heterogeneity in an EDM by computing an ensemble of conformations, with occupancies and temperature factors, that collectively provide a near-optimal explanation of the EDM. Instead of being based on an initialize-optimize approach that is classical in single-conformer programs, our method is based on a sample-select approach that adaptively alternates sampling and selection steps. We successfully tested our method on both simulated and experimental EDMs of protein fragments ranging from 4 to 12 residues in length and side-chains.

Modeling structural heterogeneity from EDMs is of major importance and may have a major impact on the way protein models are stored in the Protein Data Bank. However, our work is only a step in that direction. Several issues must still be investigated.

We need to further analyze errors caused by the finite resolution and the locality of our sampling protocol. Experiments with simulated EDMs show that if we include the correct conformations in the set of sampled conformations submitted to a SELECT step, this step reliably returns the exact ensemble of conformers (see Table 3). This suggests that an adaptive sampling protocol could generate better results than our current protocol.

Although is some cases the ensemble returned by our method is a physical model of the actual heterogeneity present in the crystal, this is not always the case. The example in Section 2.1 is a good counter-example. In general, we can only say that an ensemble returned by our method is a macromolecular representation that near-optimally fits the EDM, and thus also represents uncertainty in atomic positions. Additional physicochemical evidence is usually

required to determine if this outcome is a physical model of the conformational states present in the crystal.

Finally, to be really useful and actually used by crystallographers, our method will have to be integrated into existing suites of modeling software.

Acknowledgements: This work was partially supported by NSF grant DMS-0443939. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF. Test structures used in this work were solved and deposited as part of the JCSG pipeline (www.jcsg.org). The authors thank all members of the JCSG for their assistance in providing data. The JCSG is funded by NIH Protein Structure Initiative grants P50 GM62411, U54 GM074898. This research made use of the BioX cluster that has been partially funded by NSF award CNS-0619926.

References

1. P. V. Afonine, R. W. G. Kunstleve, and P. D. Adams. The phenix refinement framework. *CCP4 newsletter*, 42, 2005.
2. H. M. Berman, K. Henrick, H. Nakamura, and E. Arnold. Reply to: Is one solution good enough? *Nature Structural and Molecular Biology*, 13:185, 2006.
3. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 2000.
4. F. T. Burling, W. I. Weis, K. M. Flaherty, and A. T. Brunger. Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. *Science*, 271(5245):72–77, 1996.
5. E. A. Coutsias, C. Seok, M. P. Jacobson, and K. A. Dill. A kinematic view of loop closure. *Journal of Computational Chemistry*, 25(4):510–528, 2004.
6. I. W. Davis, W. B. Arendall, D. C. Richardson, and J. S. Richardson. The backrub motion: How protein backbone shrugs when a sidechain dances. *Structure*, 14:265–274, 2006.
7. M. A. DePristo, P. I. de Bakker, and T. L. Blundell. Heterogeneity and inaccuracy in protein structures solved by x-ray crystallography. *Structure*, 12:831–838, 2004.
8. A. Dhanik, P. Yao, N. Marz, R. Propper, C. Kou, G. Liu, H. van den Bedem, and J. C. Latombe. Efficient algorithms to explore conformation spaces of flexible protein loops. In *7th Workshop on Algorithms in Bioinformatics*, pages 265–276, 2007.
9. F. DiMaio, D. A. Kondrashov, E. Bitto, A. Soni, C. A. Bingman, G. N. P. Jr., and J. W. Shavlik. Creating protein models from electron-density maps using particle-flitering methods. *Bioinformatics*, 23(21):2851–2858, 2007.
10. F. DiMaio, J. Shavlik, and G. N. Phillips. A probabilistic approach to protein backbone tracing in electron density maps. *Bioinformatics*, 22(14):e81–e89, 2006.
11. J. Drenth. *Principles of protein X-ray crystallography.* Springer Verlag, New York, 1999.
12. H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

13. N. Furnham, T. L. Blundell, M. A. DePristo, and T. C. Terwilliger. Is one solution good enough? *Nature Structure Molecular Biology*, 13(3):184–185, 2006.
14. R. L. Heimer. The common optimization interface for operations research: promoting open-source software in the operations research community. *IBM Journal of Research and Development*, 47(1):57–66, 2003.
15. T. Ioerger and J. Sacchettini. The textal system: Artificial intelligence techniques for automated protein model building. In C. W. Carter and R. M. Sweet, editors, *Methods in Enzymology*, pages 244–270. Springer, 2003.
16. J. Kuriyan, G. A. Petsko, R. M. Levy, and M. Karplus. Effect of anisotropy and anharmonicity on protein crystallographic refinement: An evaluation by molecular dynamics. *Proteins: Structure, Function, and Bionformatics*, 190:227–254, 1986.
17. E. J. Levin, D. A. Kondrashov, G. E. Wesenberg, and G. N. P. Jr. Ensemble refinement of protein crystal structures: validation and application. *Structure*, 15:1040–1052, 2007.
18. S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins: Structure, Function, and Bionformatics*, 40(3):389–408, 2000.
19. K.-I. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *PNAS*, 103:11844–11849, 2006.
20. A. Perrakis, T. K. Sixma, K. S. Wilson, and V. S. Lamzin. wARP: Improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Crystallographica D*, 53:448–455, 1997.
21. R. J. Read. Improved fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallographica A*, 42:140–149, 1986.
22. V. Schomaker and K. N. Trueblood. On the rigid-body motion of molecules in crystals. *Acta Crystallographica B*, 24:63, 1968.
23. T. C. Terwilliger. Automated main-chain model building by template matching and iterative fragment extension. *Acta Crystallographica D*, 59:38–44, 2003.
24. T. C. Terwilliger. Improving macromolecular atomic models at moderate resolution by automated iterative model building, statistical density modification and refinement. *Acta Crystallographica D*, 59:1174–1182, 2003.
25. T. C. Terwilliger, R. W. Grosse-Kunstleve, P. V. Afonine, P. D. Adams, N. W. Moriarty, P. Zwart, R. J. Read, D. Turk, and L.-W. Hung. Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models. *Acta Crystallographica D*, 63:597–610, 2007.
26. S. C. E. Tosatto, E. Bindewald, J. Hesser, and R. Manner. A divide and conquer approach to fast loop modeling. 15(4):279–286, 2002.
27. H. van den Bedem, I. Lotan, J. C. Latombe, and A. M. Deacon. Real-space protein-model completion: an inverse-kinematics approach. *Acta Crystallographica D*, 61:2–13, 2005.
28. D. Vitkup, D. Ringe, M. Karplus, and G. A. Petsko. Why protein R-factors are so large: A self-consistent analysis. *Proteins: Structure, Function, and Genetics*, 46:345–354, 2002.
29. M. A. Wilson and A. T. Brunger. The 1.0 Å crystal structure of Ca2+-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. *Journal of Molecular Biology*, 301(5):1237–1256, 2000.