Full text provided by www.sciencedirect.com

# Conformer generation under restraints

Paul IW de Bakker[1], Nicholas Furnham[2], Tom L Blundell[2] and Mark A DePristo[3]

Conformational sampling by direct optimization of an all-atom energy function is ineffective and inefficient because of the ruggedness of the energy landscape. Discrete sampling schemes represent an attractive alternative for generating ensembles of conformers consistent with spatial restraints derived from empirical data. Conformational sampling is becoming increasingly important for structure prediction as the bottleneck in accurate prediction shifts from energy functions to the methods used to find low-energy conformers. Experimental structure determination remains a perennial challenge as investigators tackle larger macromolecular systems, and begin to incorporate more complete descriptions of uncertainty, heterogeneity and dynamics into their models. Computational approaches that combine dense, discrete sampling with all-atom energy evaluation and refinement may help to overcome the remaining barriers to solving these problems.

**Addresses**
[1] Department of Molecular Biology and Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, CPZN-6818, Boston, MA 02114-2790, USA, and Broad Institute of Harvard and MIT, Cambridge, MA 02141, USA and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[2] Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK
[3] Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Street, Cambridge, MA 02138, USA

Corresponding author: DePristo,
Mark A (mark_depristo@harvard.edu)

**Current Opinion in Structural Biology** 2006, **16**:160–165

This review comes from a themed issue on
Theory and simulation
Edited by Joel Janin and Michael Levitt

Available online 17th February 2006

0959-440X/$ – see front matter

**DOI 10.1016/j.sbi.2006.02.001**

## Introduction

Three-dimensional structures provide important mechanistic insights into the various molecular and cellular processes mediated by proteins. To date, experimental structure determination by X-ray crystallography and NMR spectroscopy has been most successful, allowing researchers to propose three-dimensional models at atomic resolution and make functional inferences from these structures. In the absence of experimental data, homology modeling enables structure to be inferred from an evolutionary or structural relationship established on the basis of amino acid sequence similarity [1]. *Ab initio* prediction has recently begun to produce remarkably accurate models of small, single-domain proteins [2,3••].

A unifying view of these seemingly different approaches is that available information — in the form of experimental data points, homologous structures and sequences, interaction potentials, and stereochemistry of amino acids and the polypeptide chain — is converted into a three-dimensional model of the protein. Model accuracy depends on the amount and specificity of available information; for example, protein crystals diffracting to atomic resolution provide more information than crystals that diffract to low resolution. Not surprisingly, homology modeling provides even less specific information. Obtaining an accurate all-atom model without any experimental (or homology) information remains the ultimate challenge: model accuracy is then solely dependent on the energy function and the algorithmic ability to locate its global minimum.

Here, we review recent advances in restraint-based modeling of proteins for structure prediction and determination structure (Figure 1).

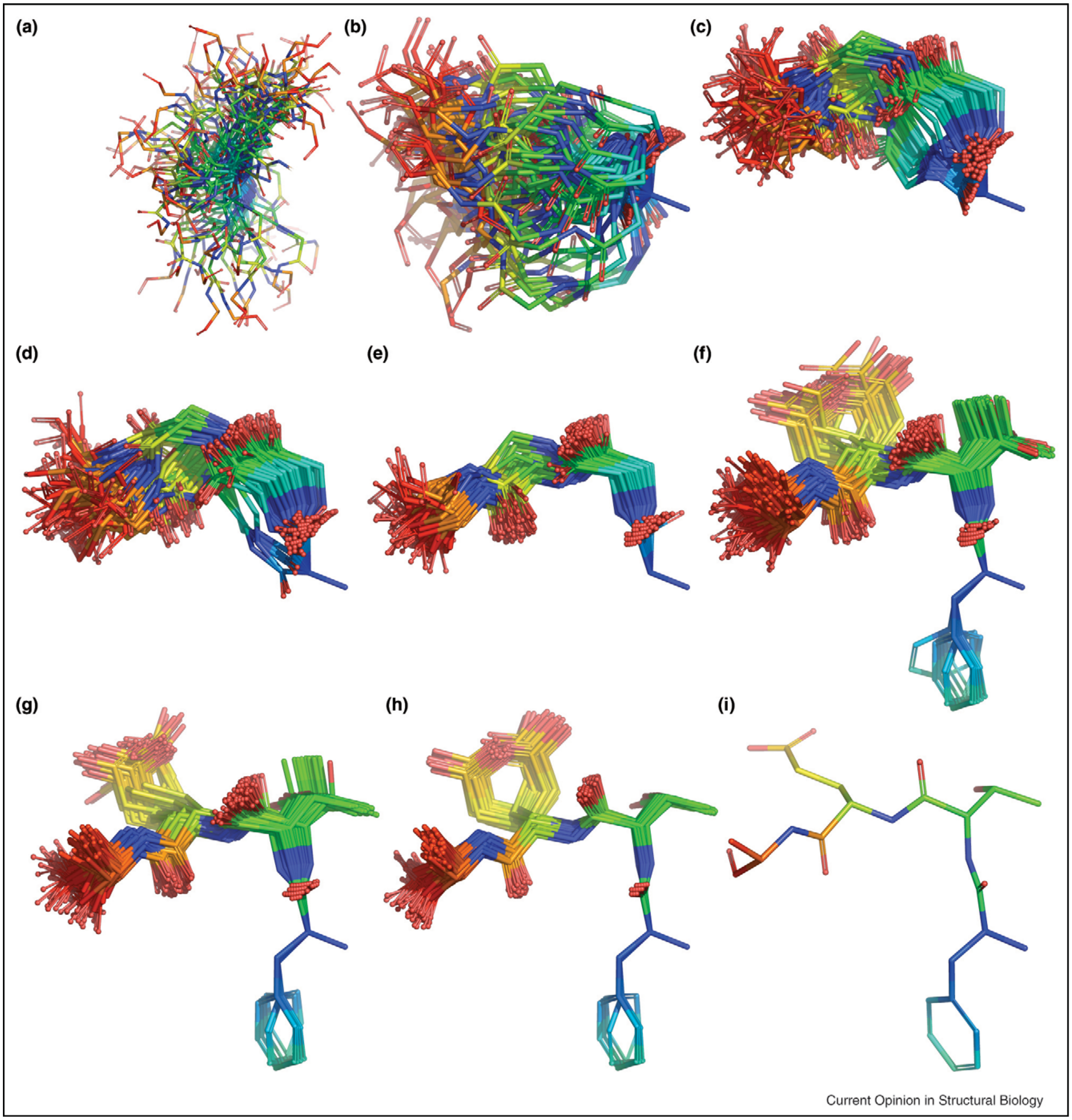
## Model representation of protein structure

The advantage of simplified, coarse-grained models of protein structure [4] is the efficiency of conformational sampling over their fewer degrees of freedom. Although such models can provide insight into folding pathways, they generally lack the atomic detail required for accurate prediction. By contrast, models that represent explicitly all atoms provide the detail required to address many biologically interesting questions and allow accurate discrimination by all-atom energy functions, but at the cost of a vast conformational space. Promising approaches address these tradeoffs by adopting a hybrid approach, in which conformational space is explored initially with simplified models that are subsequently used as seeds for dense sampling using an all-atom representation [3••,5,6].

## Conformational biases and spatial restraints

Interatomic forces constrain protein conformations and vastly reduce their effective conformational space. Atoms cannot overlap due to steric repulsion, but instead make favorable van der Waals contacts at close range [7]. Ultrahigh resolution crystallographic studies of small molecules [8] and quantum mechanical calculations have revealed tight equilibrium values for bond lengths and

**Figure 1**



Ensembles of 100 RAPPER [20,32,34••,39] conformations for residues 82–85 (FTEA) of amicyanin (PDB code 1AAC) [40] under a variety of restraints. Each subsequent panel was generated with the noted restraints in addition to those in the previous panel, except h and i. **(a)** Random coils under internal excluded volume only, **(b)** plus gap closure, **(c)** plus excluded volume against the framework protein, **(d)** plus 3 Å Cα restraints, **(e)** plus 1 Å Cα restraints, **(f)** plus 3 Å sidechain centroid restraints, **(g)** plus 1 Å sidechain centroid restraints, **(h)** excluded volume, gap closure and fit into the electron density map and **(i)** crystal structure. All panels, except (a), are from an identical perspective. Note that two conformations of Glu84 satisfy all restraints in (h). Oxygen atoms are colored red, nitrogens in blue and carbons from blue (F), green (T), yellow (E) and red (A) along the polypeptide chain. Figures generated with PyMOL [41].

angles between constituent atoms of amino acids. The principal degrees of freedom of the polypeptide chain are the φ/ψ dihedral angles along the polypeptide chain. These dihedral angles exhibit marked preferences, famously illustrated by the Ramachandran plot [9•]. Sidechains also exhibit conformational preferences (rotamers), which depend, in turn, on the mainchain conformation [10,11]. These conformational biases can be sufficiently strong so as to predispose short sequences towards particular structural motifs [12–14].

## Sampling and optimization

Once a representation of protein structure has been adopted and restraints formulated, an algorithm must be selected to find conformers consistent with these restraints. We briefly review general strategies for solving restraint satisfaction problems.

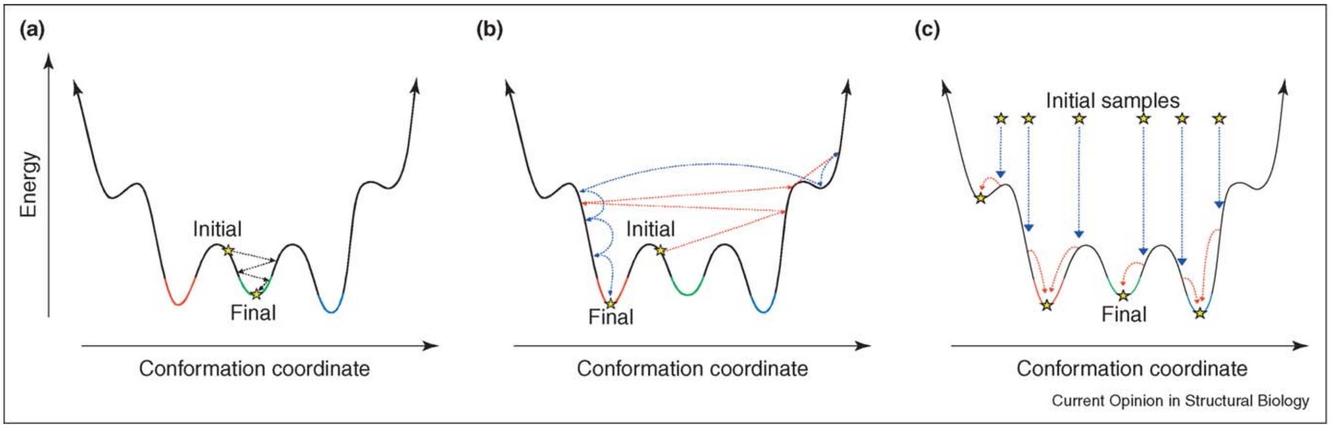
Continuous optimization aims to assign values to a set of variables ($X$) that minimize the continuous scoring function $f(X)$ that encodes the restraints and potential energy terms operating on the system (see [15] for an introduction). The continuous optimization framework is well developed, general and powerful, ranging from efficient first-derivative minimization algorithms limited to local optimization (Figure 2a) to simulated annealing algorithms for non-local optimization (Figure 2b). Not surprisingly, continuous optimization has been employed with great success in many modeling programs, such as CNS [16] and MODELLER [1,17].

The application of continuous optimization algorithms has been fraught with difficulties. The potential energy landscapes encountered in structural biology are rugged and pocked with local minima [18,19] that must be avoided as poor solutions to the restraint satisfaction problem. The penchant of local optimization algorithms to become trapped in local minima on such landscapes drove the adoption of non-local optimization strategies, such as simulated annealing [16] (Figure 2b). Though superior to direct minimization, simulated annealing still suffers from an inability to escape local minima. Indeed, there are no, nor are there likely to ever be, efficient methods to find the global energy minimum on arbitrary potential energy landscapes. What, then, is an effective heuristic search strategy for generating low-energy conformers?

Discrete conformational sampling has proven a viable, if not superior, alternative to continuous optimization for restraint-based modeling of protein structure on the rugged landscapes common in structural biology (see Table 1) [4,20–23]. The single biggest advantage of discrete sampling is that it freely crosses energy barriers that separate conformations (Figure 2c). This barrier-crossing problem is a general challenge for all approaches that explore conformational space by generating a series of related structures in which each successive structure is a perturbation of the previous one. As the probability of visiting a state is proportional to its energy, such methods have difficulties traversing rugged landscapes with deep valleys separated by high peaks.

Ideally, the sampling of conformational states should be in proportion to the energies of the low-energy states and not the height of the barriers separating them. Discrete sampling schemes in part realize this by sampling directly from empirical propensities. In this way, only low-energy conformations are examined and kinetic traps, such as those due to *cis-trans* peptide bond or sidechain rotamer flips, are bypassed. Consequently, discrete sampling is, for many applications, an efficient and effective means to explore conformational space and rugged energy landscapes [21].

**Figure 2**



Methods for exploring complex energy landscapes. A rugged energy landscape with multiple nearly iso-energetic minima (red, green, blue) separated by energy barriers. **(a)** During direct minimization, the initial conformation (upper star) undergoes a series of minimization steps that reduce its energy. The final conformation (lower star) is at the bottom of the local energy well. **(b)** Simulated annealing begins by heating (red line) the initial conformation (star), so that local energy barriers can be overcome and a range of conformations sampled. Slow cooling (blue line) then returns the system to a low-energy well (lower star). **(c)** Discrete conformational sampling first generates a range of initial conformations (upper stars), independent of the underlying energy landscape. These conformations typically have poor initial energies with respect to the underlying energy landscape (blue projection lines). Nevertheless, direct energy minimization (red lines) produces final conformations (lower stars) at the bottom of their local energy wells.

**Table 1**

**Applications of restraint-based conformer generation.**

| Application/field | Problem | Specific restraints | References |
|---|---|---|---|
| Structure prediction | | | |
| Loop modeling | Construct loop region given knowledge of surrounding protein structure | Reattachment to framework structure. Scoring with potential energy functions | [5,32,33,39] |
| Sidechain assignment and design | Find optimal configuration of sidechain rotamers given fixed mainchain conformation | Pairwise interaction potential | [42–44] |
| Comparative modeling | Construct model of a sequence of unknown structure given homology to a sequence of known structure | Atomic positions, dihedral angles, secondary structure inferred from homologs | [1,17] (a) |
| Structure determination | | | |
| Crystallographic model building and refinement | Rebuild model to improve consistency with X-ray data | Electron density map and reciprocal-space reflections | [21,29,31,34••] |
| NMR | Find and/or improve models consistent with NMR data | Interatomic distances, dihedral angles, interatomic angles relative to global coordinate system | [25,26,45•,46] |
| Electron microscopy and tomography | Dock atomic-level structures into low-resolution micrograph or tomograph | Individual protein structures, low-resolution electron density map, known interactions between proteins | [27,28•] |
| Structural dynamics and heterogeneity | Determine the accuracy, heterogeneity and dynamics of proteins | X-ray or NMR data | [34••,38,47] |

(a) N Furnham, PIW de Bakker, MA DePristo, DF Burke, TL Blundell, unpublished.

## Applications and specific restraints

Restraint-based conformer generation has been successfully employed in a variety of contexts in structural biology (Table 1). These applications segregate into two main areas: structure determination from a set of experimental data; and structure prediction by homology or *ab initio* modeling. Restraint-based conformer generation has proven especially fruitful for structure determination because of the difficulties inherent to deriving atomic positions from the reciprocal-space reflections produced by X-ray crystallography [24], interatomic distance and angular relationships generated by NMR [25,26], or the low-resolution electron maps from microscopy [27] and tomography [28•]. The crystallographic community early appreciated the usefulness of restraint-based modeling to produce models that fit into real-space electron density maps [29,30]. These early techniques have evolved into fully automated methods that identify and rebuild poorly fit regions, improving on the refinement process itself [21,31].

Conformer generation has also enjoyed widespread success in structure prediction (Table 1), from short loops [5,32,33] to whole proteins [3••,17]. Baker and colleagues [2,3••] have recently achieved remarkable accuracy in *ab initio* prediction. Their hybrid approach first generates an ensemble of conformers using a coarse-grained representation of protein structure; these are subsequently converted to an all-atom representation and refined against an all-atom energy function. Although most predictions are to within a few angstroms of the experimental structure, the poor predictions are due to an inability to produce initial models close to the native structure and are not limitations of the potential energy function, which scores the native and neighboring structures better than any of the sampled conformations [3••]. They conclude that “the primary bottleneck to consistent high-resolution structure prediction appears to be conformational sampling”.

## Structural heterogeneity and dynamics

A range of equivalent solutions can often be found for a given set of restraints. For example, in resolving X-ray crystallographic data, several different conformers of equivalent quality can be independently determined from the structure factors [34••]. Such an ensemble of solutions captures the uncertainty in the determined structure, and its associated heterogeneity and dynamics. Traditionally, uncertainty, dynamics and heterogeneity have been represented by atomic B factors, a measure of the mean square atomic displacement. However, B factors cannot adequately describe correlated motions among atoms or discrete conformational substates [35–37]. One solution to this shortcoming is to explicitly represent heterogeneity and dynamics with an ensemble of conformers [38], as is standard in the NMR and molecular dynamics communities. We hope that the crystallography community will adopt this practice in the near future, too.

## Conclusions

Discrete conformational sampling is an effective strategy for the efficient generation of ensembles of structures consistent with spatial restraints. A fruitful direction for future research is to better understand the relationship between the information content of restraints and the accuracy of three-dimensional atomic positions inferred from those restraints (e.g. Figure 1). Such a theory would be helpful in the hitherto unsuccessful effort to marry disparate sources of structural information, such as simultaneous refinement against both NMR and X-ray data, or the incorporation of secondary structure predictions in comparative modeling. On the practical side, we expect that discrete conformational sampling should improve structure determination and prediction in so far as these are limited by conformational sampling. For example, simply generating conformers around a putative model followed by minimization results in a significant improvement in structure prediction [32] and determination [21]. Restraint-based conformational sampling is helping to overcome some of the remaining obstacles to fully automated structure determination and atomic-resolution structure prediction.

## Acknowledgements

MAD is a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRG-1861-05). NF is supported by the UK Biotechnology and Biological Sciences Research Council.

## References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Sali A, Overington JP, Johnson MS, Blundell TL: **From comparisons of protein sequences and structures to protein modelling and design**. *Trends Biochem Sci* 1990, **15**:235-240.

2. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D: **Progress in modeling of protein structures and interactions**. *Science* 2005, **310**:638-642.

3. •• Bradley P, Misura KM, Baker D: **Toward high-resolution *de novo* structure prediction for small proteins**. *Science* 2005, **309**:1868-1871.

This article describes *ab initio* structure prediction of small, single-domain proteins to near-atomic resolution using a combination of coarse- and fine-grained models of protein structure, sophisticated methods of conformational sampling and accurate empirical potential energy functions.

4. Tozzini V: **Coarse-grained models for proteins**. *Curr Opin Struct Biol* 2005, **15**:144-150.

5. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA: **A hierarchical approach to all-atom protein loop prediction**. *Proteins* 2004, **55**:351-367.

6. Li X, Jacobson MP, Friesner RA: **High-resolution prediction of protein helix positions and orientations**. *Proteins* 2004, **55**:368-382.

7. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC: **Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms**. *J Mol Biol* 1999, **285**:1711-1733.

8. Engh RA, Huber R: **Accurate bond and angle parameters for X-ray protein structure refinement**. *Acta Crystallogr A* 1991, **47**:392-400.

9. • Lovell SC, Davis IW, Arendall B III, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC: **Structure validation by C$\alpha$ geometry: $\varphi$, $\psi$, and c$\beta$ deviation**. *Proteins* 2003, **50**:437-450.

The authors derived fine-grained $\varphi/\psi$ propensity maps for all amino acids from a non-redundant collection of high-quality crystallographic structures. These maps can be used to assess the local quality of structures or for propensity-weighted conformational sampling.

10. Lovell SC, Word JM, Richardson JS, Richardson DC: **The penultimate rotamer library**. *Proteins* 2000, **40**:389-408.

11. Dunbrack RL: **Rotamer libraries in the 21st century**. *Curr Opin Struct Biol* 2002, **12**:431-440.

12. Han KF, Bystroff C, Baker D: **Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns**. *Protein Sci* 1997, **6**:1587-1590.

13. Srinivasan R, Rose GD: **A physical basis for protein secondary structure**. *Proc Natl Acad Sci USA* 1999, **96**:14258-14263.

14. Shortle D: **Composites of local structure propensities: evidence for local encoding of long-range structure**. *Protein Sci* 2002, **11**:18-26.

15. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C*. Cambridge: Cambridge University Press; 1997.

16. Brunger AT, Kuriyan J, Karplus M: **Crystallographic R factor refinement by molecular dynamics**. *Science* 1987, **235**:458-460.

17. Fiser A, Sali A: **Modeller: generation and refinement of homology-based protein structure models**. *Methods Enzymol* 2003, **374**:461-491.

18. Frauenfelder H, Sligar SG, Wolynes PG: **The energy landscapes and motions of proteins**. *Science* 1991, **254**:1598-1603.

19. Frauenfelder H, Lesson DT: **The energy landscape in non-biological and biological molecules**. *Nat Struct Biol* 1998, **5**:757-759.

20. DePristo MA, de Bakker PIW, Shetty RP, Blundell TL: **Discrete restraint-based protein modeling and the C$\alpha$-trace problem**. *Protein Sci* 2003, **12**:2032-2046.

21. DePristo MA, de Bakker PIW, Johnson RJ, Blundell TL: **Crystallographic refinement by knowledge-based exploration of complex energy landscapes**. *Structure* 2005, **13**:1311-1319.

22. Srinivasan R, Fleming PJ, Rose GD: ***Ab initio* protein folding using LINUS**. *Methods Enzymol* 2004, **383**:48-66.

23. Gibbs N, Clarke AR, Sessions RB: ***Ab initio* protein structure prediction using physicochemical potentials and a simplified off-lattice model**. *Proteins* 2001, **43**:186-202.

24. Hauptman HA: **The phase problem of X-ray crystallography**. *Rep Prog Phys* 1991, **54**:1427-1454.

25. Meiler J, Baker D: **Rapid protein fold determination using unassigned NMR data**. *Proc Natl Acad Sci USA* 2003, **100**:15404-15409.

26. Rohl CA, Baker D: ***De novo* determination of protein backbone structure from residual dipolar couplings using Rosetta**. *J Am Chem Soc* 2002, **124**:2723-2729.

27. Topf M, Sali A: **Combining electron microscopy and comparative protein structure modeling**. *Curr Opin Struct Biol* 2005, **15**:578-585.

28. • Sali A, Glaeser R, Earnest T, Baumeister W: **From words to literature in structural proteomics**. *Nature* 2003, **422**:216-225.

This paper describes the looming challenge to integrate structural information at a hierarchy of resolutions — from atomic (X-ray and NMR structures) to macromolecular complexes and organelles (electron microscopy and tomography) — into a coherent three-dimensional model of the cell.

29. Kleywegt GJ, Jones AT: **Efficient rebuilding of protein structures**. *Acta Crystallogr D Biol Crystallogr* 1996, **52**:829-832.

30. Jones TA, Zou JY, Cowan SW, Kjeldgaard M: **Improved methods for building protein models in electron density maps and the location of errors in these models**. *Acta Crystallogr A* 1991, **47**:110-119.

31. Terwilliger TC: **SOLVE and RESOLVE: automated structure solution and density modification**. *Methods Enzymol* 2003, **374**:22-37.

32. de Bakker PIW, DePristo MA, Burke DF, Blundell TL: ***Ab initio* construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the generalized Born solvation model**. *Proteins* 2003, **51**:21-40.

33. Singh R, Berger B: **ChainTweak: sampling from the neighbourhood of a protein conformation**. *Pac Symp Biocomput* 2005:52.

34. DePristo MA, de Bakker PIW, Blundell TL: **Heterogeneity and •• inaccuracy in protein structures solved by X-ray crystallography**. *Structure* 2004, **12**:831-838.

This paper shows that discrete conformational sampling from an X-ray crystallographic structure can produce ensembles of alternative models of equivalent quality to the PDB structure. It estimates the accuracy of crystallographic structures as a function of diffraction resolution by comparing the heterogeneity of the structures in these ensembles.

35. Burling FT, Weis WI, Flaherty KM, Brunger AT: **Direct observation of protein solvation and discrete disorder with experimental crystallographic phases**. *Science* 1996, **271**:72-77.

36. Vitkup D, Ringe D, Karplus M, Petsko GA: **Why protein R-factors are so large: a self-consistent analysis**. *Proteins* 2002, **46**:345-354.

37. Kuriyan J, Petsko GA, Levy RM, Karplus M: **Effect of anisotropy and anharmonicity on protein crystallographic refinement. An evaluation by molecular dynamics**. *J Mol Biol* 1986, **190**:227-254.

38. Furnham N, DePristo MA, Blundell TL, Terwilliger TC: **Is one solution good enough?** *Nat Struct Mol Biol* 2006, in press.

39. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL: ***Ab initio* construction of polypeptide fragments: efficient generation of accurate, representative ensembles**. *Proteins* 2003, **51**:41-55.

40. Cunane LM, Chen ZW, Durley RCE, Mathews FS: **X-ray structure of the cupredoxin amicyanin, from *Paracoccus denitrificans*, refined at 1.31 Å resolution**. *Acta Crystallogr D Biol Crystallogr* 1996, **52**:676-686.

41. DeLano WL: *The PyMOL Molecular Graphics System*. San Carlos, CA: DeLano Scientific; 2002:. (http://www.pymol.org/).

42. Bower MJ, Cohen FE, Dunbrack RL Jr: **Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool**. *J Mol Biol* 1997, **267**:1268-1282.

43. Dunbrack RL, Cohen FE: **Bayesian statistical analysis of protein side-chain rotamer preferences**. *Protein Sci* 1997, **6**:1661-1681.

44. Desmet J, Maeyer M, Hazes B, Lasters I: **The dead-end elimination theorem and its use in protein side-chain positioning**. *Nature* 1992, **356**:539-542.

45. Bax A: **Weak alignment offers new NMR opportunities to study • protein structure and dynamics**. *Protein Sci* 2003, **12**:1-16.

An excellent introduction to NMR structure determination using residual dipolar coupling. Bax highlights the challenges and opportunities this poses for conformational sampling and refinement methods.

46. Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CA, Nabuurs SB, Guntert P, Livny M, Markley JL, Nilges M *et al.*: **RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank**. *Proteins* 2005, **59**:662-672.

47. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M: **Simultaneous determination of protein structure and dynamics**. *Nature* 2005, **433**:128-132.