Center for Eukaryotic Structural Genomics

Protein Structure Initiative

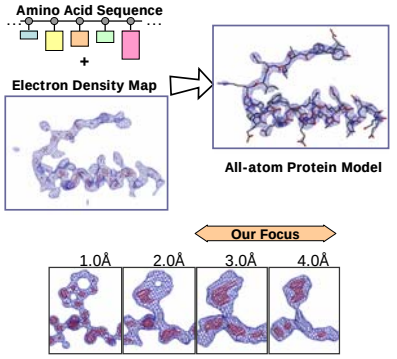
NEW APPROACH TO AUTOMATIC FITTING OF ELECTRON-DENISTY MAPS

Craig A. Bingman[1], Frank DiMaio[2,3,4], Ameet Soni [1,2], Jude W. Shavlik [1,2], Dmitry Kondrashov [2], Eduard Bitto[1,5], George N. Phillips, Jr.[1,2]

University of Wisconsin-Madison, Department of Biochemistry, 433 Babcock Drive, Madison, WI USA 53706-1549 http://www.uwstructuralgenomics.org

Introduction

One bottleneck in high-throughput crystallography is the interpretation of a macromolecular model from the electron density map. When the map is well phased and at sufficiently high resolution, this interpretation can usually be completed more or less automatically by existing methods, such as Arp/warp (1), Textal (2), or Resolve (3). However, when the resolution is lower than about 2.5 Å or the phasing is not particularly accurate, it can sometimes take weeks or months of a crystallographer's time to complete a chain trace, registered with the sequence. This process requires substantial trial and error.



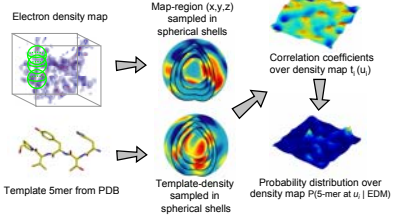
We have developed the automatic interpretation tool, ACMI (for *Automatic Crystallographic Map Interpreter*) (4-6). ACMI employs probabilistic inference to compute the probability distribution of each amino acid's 3D location given this density map.

Phase 1: Pentapeptide Matching

• Estimate probability of amino acid locations using pattern matching.

• Spherical Harmonic decomposition(6) allows for rapid search over all rotations for each amino acid template for each location in the map.

ACMI's observational potential is computed by searching for a pentapeptide centered at each amino-acid location in the protein. That is, ACMI walks along the protein, one amino-acid at a time, and considers the 5-amino-acid sequence centered at each position. It searches the PDB for all observed conformations of that particular sequence. Spherical harmonics are used to quickly compute the correlation coefficient between some region in the (unsolved) map and a pentapeptide fragment over all rotations for each location on the map. .

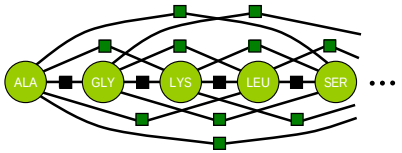


Phase 2: Probabilistic Protein Model

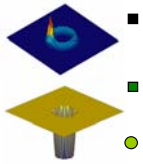
• Refine local search using protein structural constraints.

• Marginal probabilities approximated using Loopy Belief Propagation.

ACMI models a protein using a pairwise Markov random field, a probabilistic model where a probability distribution over some set of random variables is defined on an undirected graph. Specifically, the probability is given as the product of potential functions associated with vertices and edges in the graph.

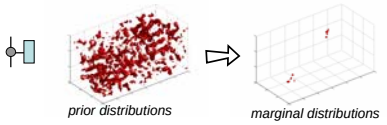


To model a protein, ACMI constructs a graph where each vertex corresponds to an amino acid. A random variable associated with each node describes the position and orientation of that amino acid's Cα. The joint probability of some configuration of Cα's is the product of …



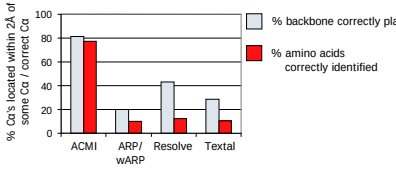
- An adjacency potential function between neighboring amino acids. This ensures that adjacent amino acids are the proper distance (3.8Å) apart and in the proper relative orientation.
- An occupancy potential function over all other amino acid pairs. This ensures that no two amino-acids occupy the same 3D space.
- An observational potential function associated with each vertex (that is, each amino acid). This is from Phase 1 (prior probabilities).

The marginal probability distributions are intractable to compute; Belief Propagation (Pearl 1988) is a message passing algorithm that provides an approximation (in loopy graphs) that is not guaranteed to converge, but works well in practice. Final result is an approximate marginal distribution of each amino acid's Cα position/orientation.



Phase 2 Results

We compare ACMI's backbone trace to three popular tools on a set of 10 experimentally-phased density maps over a wide range of phase error and resolutions. We compare the resultant models (Cα locations only) both in terms of average RMS error and in terms of correctly-identified amino acids. ACMI is producing a more accurate and more complete Cα trace.



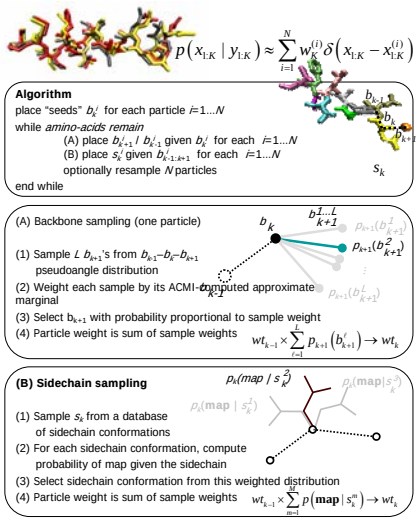
Phase 3: Producing an All-Atom Model

• Sampling using particle filtering produces a set of physically feasible all-atom models.

• Phase 2's belief and domain knowledge guide the search.

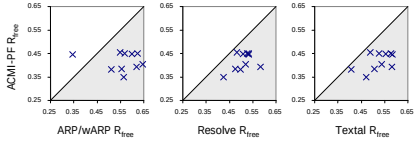
Given the approximate marginal distributions provided by ACMI, we want to compute an ensemble of physically feasible all-atom protein models. Our approach uses a probabilistic method know as particle filtering (PF).

A particle here refers to the refers to one specific 3D layout of all the non-hydrogen atoms in a contiguous subsequence of the protein. PF represents the distribution of some subsequence's layout using a set of particles and their weights



Phase 3 Results

ACMI's all-atom model using particle filtering produces physically feasible along with fewer chains and better RMSD than Phase 2 alone. Compared to current methods, ACMI-PF produces lower R_{free} values on all maps with only one exception.

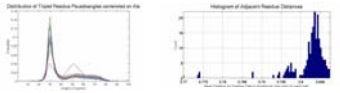


Introducing Domain Knowledge

Importance sampling's performance is dependent on the accuracy of our *importance distribution* (the sampling function - angle/bond constraints for backbones and PDB instances for sidechains). The likelihood of keeping a sample is dependent on the accuracy of *importance evaluation function* (quality of match to Phase 2 beliefs for the backbone and density correlation for sidechains).

Sampling Function

Currently, we assume Cα-Cα distances are the same for all pairs of residues (~3.8 Å with a tight distribution). In actuality, specific pairs exhibit different distances and variation (proline in particular). In addition, sampling psuedoangles for Cα-Cα-Cα triplets from a bimodal distribution hides the fact that not all residues exhibit the same distribution of psuedoangles (vary by secondary structure, surface accessibility, etc.). By accumulating data from the PDB, we can capture higher order information such as secondary structure into the sampling function.



Recent Activities

We used ACMI to construct atomic models on 20 crystallographic data sets with the initial phases available before any model building. The structures had been previously solved and deposited in the PDB, providing correct models to assess the performance of the automated model-building. The input files provided to the software are the protein sequence, the crystal symmetry and cell parameters, and the electron density map computed with the initial, experimental phases. The density maps provided different levels of challenge for model-building, depending on the accuracy of initial phasing and the available resolution. Two measures of quality and the number of residues in the asymmetric unit have been examined, with the phase error computed by comparing the initial phases with the calculated phases from the final model. In addition to the resolution, phase error is an important factor for interpretability of an electron density map: a high-resolution data set with poor initial phases map produce an indistinct map. Based on our experience and expert judgement of the quality of the initial density maps, we divided the data sets into two categories: easy to interpret and difficult. We now have the results of automated model building from ACMI and three other programs for the two categories of data sets. The model building performance is assessed for the backbone completeness and sidechain identification, in terms of prediction sensitivity (fraction of the correct model predicted) and specificity (fraction of the predicted model which is correct). Below is one sample table (manuscript in preparation). ACMI has also been used to solve a new 2.7 Angstrom resolution structure that is currently in refinement.

Sidechain Identification Comparison for Difficult Data Sets

PDB ID	ACMI-PF	Arp/Warp	Resolve	Textal
2nxf[c]	73.5/73.5	96.1/99.3	6.8/10.8	3.9/4.1
2q7a[b]	80.7/84.8	2.0/7.0	35.1/47.1	28.9/31.9
3bus	89.5/91.9	1.0/3.7	11.3/17.2	4.7/6.8
1xri	87.7/84.8	4.3/6.1	17.2/26.0	37.7/40.0
1ztp	95.0/96.7	0.58/2.8	63.6/81.3	52.0/57.6
1v0z	81.2/85.0	21.9/38.5	15.6/17.1	
2a3q	87.8/71.4	10.4/17.8	5.9/ 8.3	
2ifu	63.0/79.9	0.2/0.4	5.9/12.9	3.8/7.6
2bdu	78.9/85.3	0.5/1.8	12.5/20.1	5.0/6.6
2ab1	56.0/70.1	0.9/4.9	3.4/5.9	3.9/7.5

[b]The two numbers reported are the percent of the correct structure predicted/percent of the predicted structure correct.

Other Improvements

Phase 1: Filtering, Fragment Selection

Phase 1 is the largest consumer of CPU time, although high parallelization mitigates the problem. The proportion of Cα locations to points in the map is miniscule, thus most of the computations over the map are useless. A SVM trained to predict Pr (Cα | EDM) allows a large proportion of the map to be eliminated *apriori* without losing many positives (80% precision at 95% recall). This leads to ~35% reduction in computation over using density values alone, and up to 5x improvement overall.

Fragment selection has been expanded to select larger n-mers if they exist in the DB. In addition, work is being done to search for larger homologous domains and more accurately select "good" 5mers (particularly in maps where only secondary structure can be determined).

Phase 2-3: Ensembles and Diversity

One key extension is to overcome error accumulation in BP by sampling more diverse distributions. Ensemble generation in both Phase 2 and 3 can allow greater sampling of the conformational space. These can be created via dynamic scheduling in BP, early stopping of BP, selecting numerous starting points in PF (already done), and testing particle filtering variations.

References

(1) R. Morris, A. Perrakis, and V. Lamzin. (2003). *Meth Enz* 374, 229-244.
(2) T. Ioerger and J. Sacchettini. (2002). *Acta Cryst* D5, 2042-2054.
(3) T.C. Terwilliger. (2003). *Acta Cryst* D59, 38-44.
(4) F. DiMaio, J. Shavlik, and G. Phillips, Jr. (2006). *Bioinformatics* 22, e81-89.
(5) F. DiMaio, D. Kondrashov, E. Bitto, A. Soni, C. Bingman, G. Phillips & J. Shavlik (2007). *Bioinformatics*. doi: 0.1093/bioinformatics/btm480
(6) F. DiMaio, A. Soni, G. Phillips & J. Shavlik (2007). *BIBM'07*, Fremont, CA.

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative NIH/NIGMS grant number U54 GM074901.