Center for Eukaryotic Structural Genomics

Protein Structure Initiative

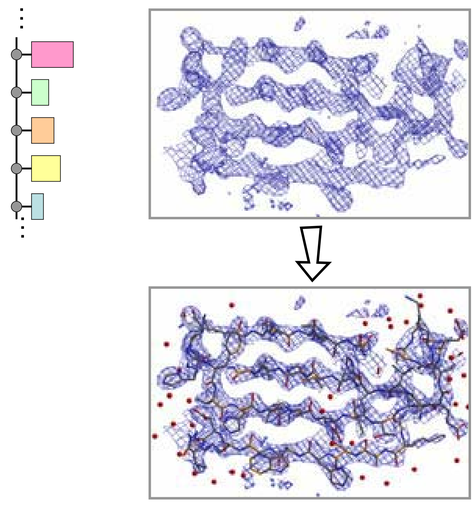
New Approaches to Automatic Fitting of Electron Density Maps

Frank DiMaio[a], Ameet Soni[a], Jude W. Shavlik[a,b], Dmitry Kondrashov[c], Craig A. Bingman[c], Eduard Bitto[c], George N. Phillips, Jr.[a,c]

[a] Computer Sciences Department, [b] Biostatistics and Medical Informatics Department, [c] Center for Eukaryotic Structural Genomics (CESG), Biochemistry Department, University of Wisconsin-Madison,Madison, WI 53706 http://www.uwstructuralgenomics.org

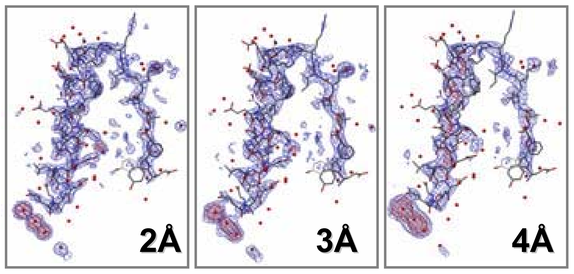
Introduction

One bottleneck in high-throughput crystallography is the **completion 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).

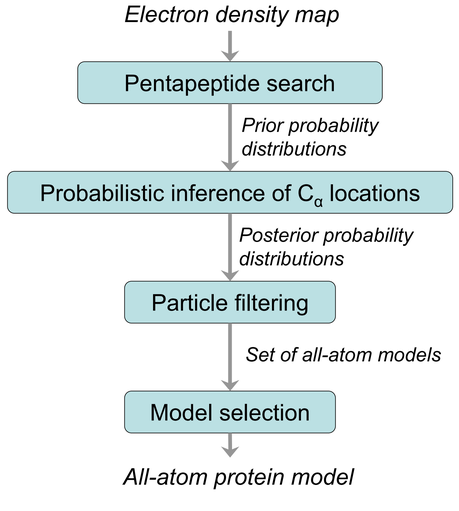


However, when the resolution is lower than about 2.5 Å or the phasing is not particularly accurate, It can sometimes take weeks or months of a crystallographer's time to complete a chain trace, registered with the sequence. This process requires substantial trial and error; often with small and potentially incorrect extensions of the model in any cycle of iterative model/phase extention..

Illustrated here is the same protein fragment at three different resolutions.

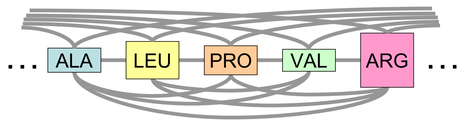


We have developed the automatic interpretation tool, ACMI (for *Automatic Crystallographic Map Interpreter*) (4). ACMI employs probabilistic inference to compute the probability distribution of each amino acid's 3D location given this density map. **A flowchart describing ACMI is illustrated below.** It models the location of a single atom in each amino acid, the alpha carbon (Cα), on a grid with ≈1Å grid spacing. **Probabilistic inference** computes the distribution of each Cα over this grid. Finally, a **particle filtering** method is used to construct an all-atom model from these intermediate distributions.

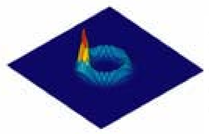


Probabilistic Protein Model

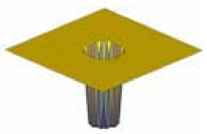
ACMI models a protein using a **pairwise Markov random field**, a probabilistic model where a probability distribution over some set of random variables is defined on an undirected graph. Specifically, the probability is given as the product of **potential functions** associated with vertices and edges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



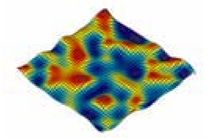
To model a protein, ACMI constructs a graph where each vertex corresponds to an amino acid. A random variable associated with each node describes the position of that amino acid's Cα. **The joint probability of some configuration of Cα's is the product of …**



An **adjacency potential function** associated with the **edges between neighboring amino acids**. This ensures that adjacent amino acids are the proper distance (3.8Å) apart and in the proper relative orientation.



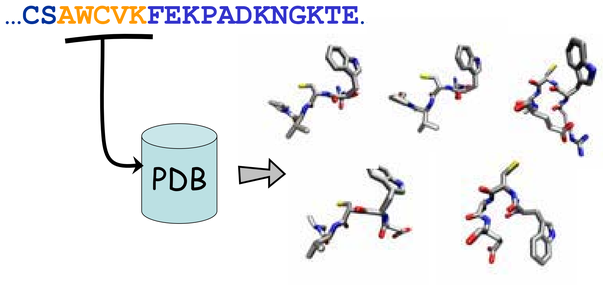
An **occupancy potential function** associated with the **edges between all other amino acid pairs**. This ensures that no two amino-acids occupy the same 3D space. The potential is given as a step function which is only non-zero if two amino acids are sufficiently far apart.



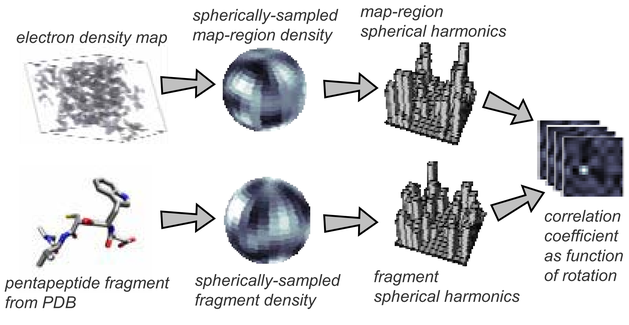
An **observational potential function** associated with **each vertex** (that is, **each amino acid**). This function describes how much a region in the density map looks like a particular amino-acid, and is computed by searching the map for **sequence-specific pentapeptide fragments**.

Pentapeptide Matching

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.



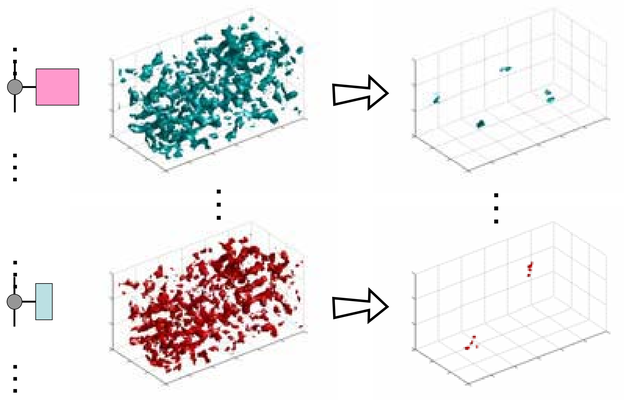
Then, **for each of these fragments, ACMI considers centering the fragment at each location on the 1Å grid**. Spherical harmonics are used to quickly compute the correlation coefficient between **some region in the (unsolved) map and a pentapeptide fragment over all rotations**.



These correlation coefficients are converted to probabilities, which serve as ACMI's observation potentials.

Inferring a Cα Trace

ACMI's pentapeptide matching provides a **prior probability distribution** over the density map (on a 1Å grid). ACMI uses an approximate inference algorithm known as belief propagation to compute each amino acid's **marginal (posterior) probability**. That is, it finds the probability distribution of each Cα's location, taking **constraints on the protein structure** into account. **For example,**

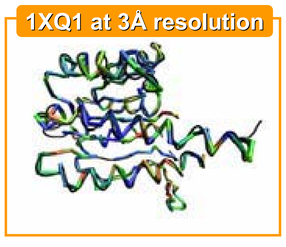


prior distributions *marginal distributions*

Sample Backbone Traces

ACMI computes an (intermediate) backbone trace by taking – for each Cα – the location that **maximizes the marginal distribution**. Sample backbone traces are illustrated below. **Thin black lines** indicate the true (crystallographer-determined) structure while the **thicker segmented lines** indicate ACMI's predicted structure. Line color indicates confidence, from least confident to most confident

0.918Å RMSd
99.6% complete

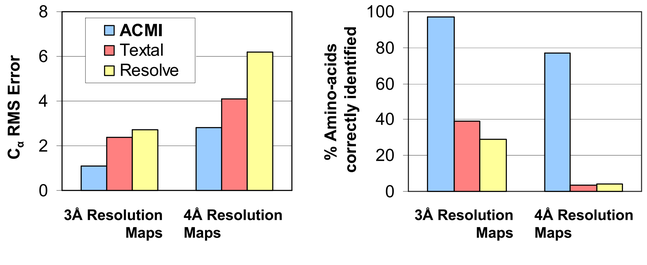


0.844Å RMSd
99.8% complete

3.202Å RMSd
87.6% complete

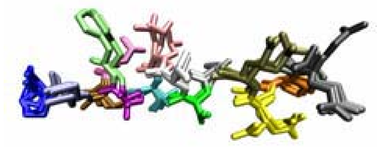
Results

We compare ACMI's backbone trace to both TEXTAL and RESOLVE on a set of 10 model-phased density maps smoothly downsampled to both 3Å and 4Å resolution. We compare the resultant models (Cα locations only) both in terms of average RMS error and in terms of correctly-identified amino acids. At both resolutions, ACMI is producing a more accurate and more complete Cα trace.



Producing an All-Atom Model

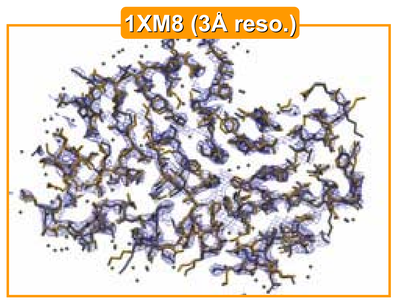
Given the approximate marginal distributions provided by ACMI, we want to compute **an ensemble of physically feasible all-atom protein models** that produces this distribution. Our approach uses a probabilistic method know as **particle filtering (PF)**.



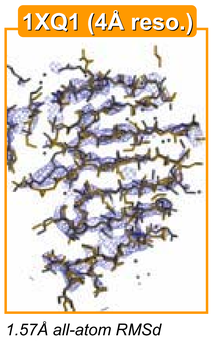
A particle here refers to the refers to **one specific 3D layout of all the non-hydrogen atoms in a contiguous subsequence of the protein** (e.g., from amino acid 21 to 25). PF represents the distribution of some subsequence's layout using a set of distinct layouts for that subsequence, as illustrated above. At each iteration of particle filtering, we grow our particle by one amino acid.

Sample All-Atom Models

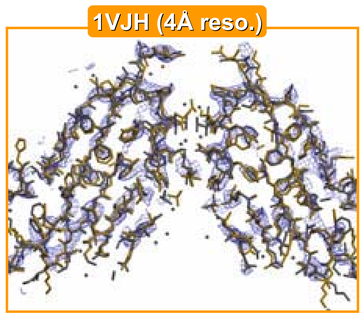
Shown below is a comparison of the best all-atom model (in terms of R-factor) against the true solution. ACMI's model is in **yellow**, while the true (crystallographer-determined) solution is in **grey**.



1.44Å all-atom RMSd *99% complete*



1.57Å all-atom RMSd
95% complete

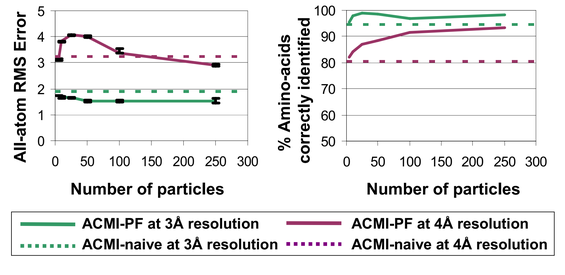


1.68Å all-atom RMSd *98% complete*

Results

We compare ACMI's all-atom model using particle filtering to a naïve approach. ACMI-NAÏVE takes the maximum-marginal backbone trace, and places the best-matching sidechain (using the pentapeptide fragments) at each position.

These results show RMS error and completeness as a function of the number of particles. PF always produces a more-complete model than does the backbone trace alone. In addition, using 250 particles, our approach shows improved accuracy over ACMI-NAÏVE.



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, A. Soni, G.N. Phillips, Jr., and J.W. Shavlik. (2007, submitted).

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