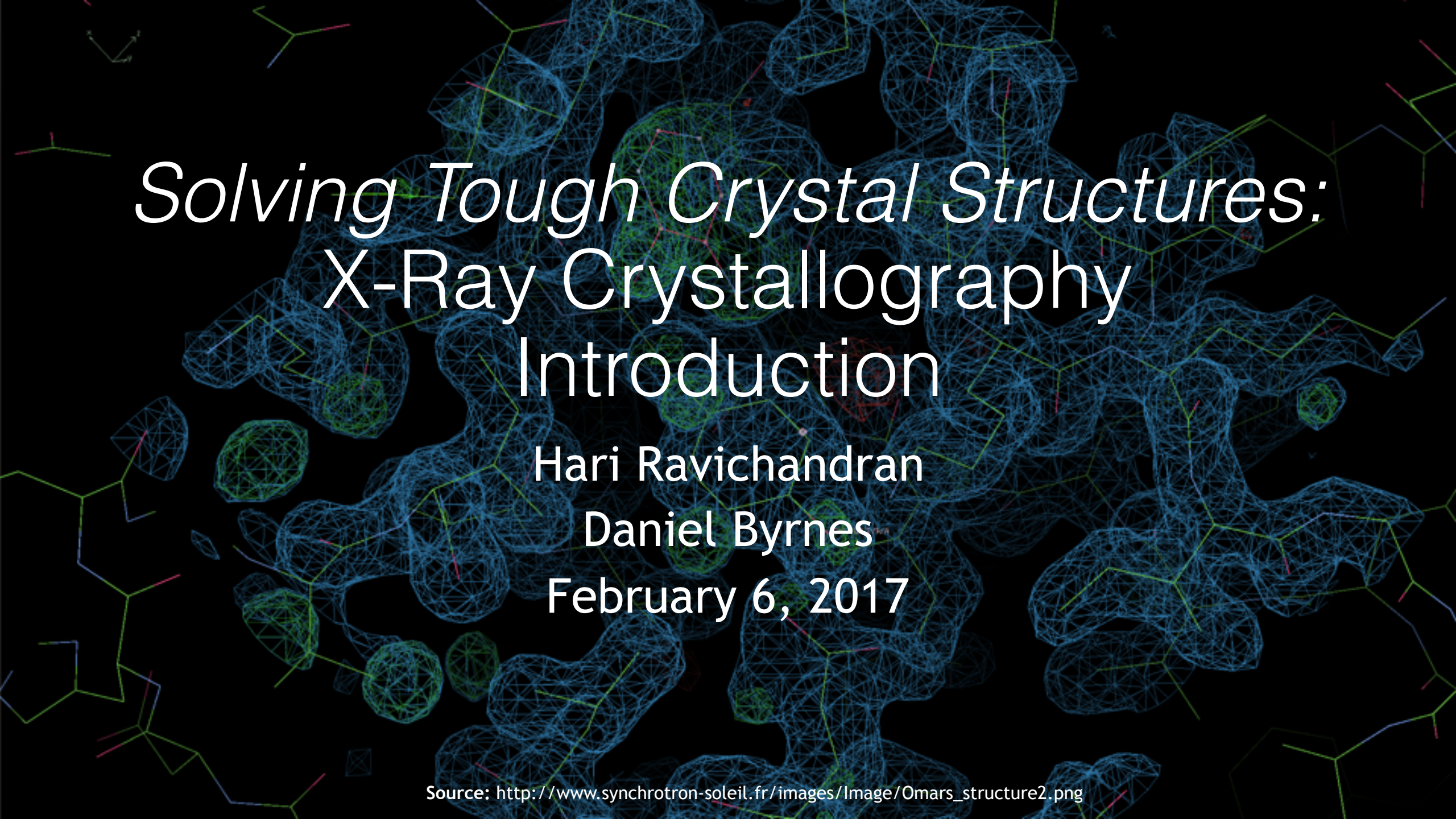


Solving Tough Crystal Structures: X-Ray Crystallography Introduction

Hari Ravichandran
Daniel Byrnes

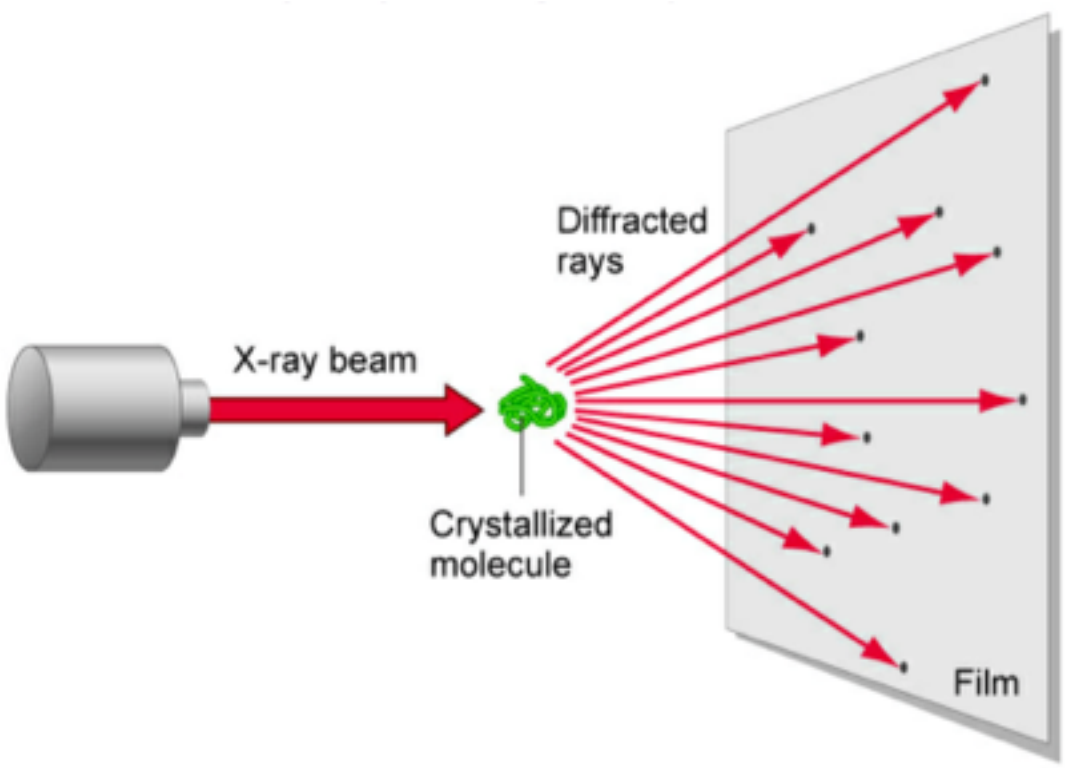
February 6, 2017

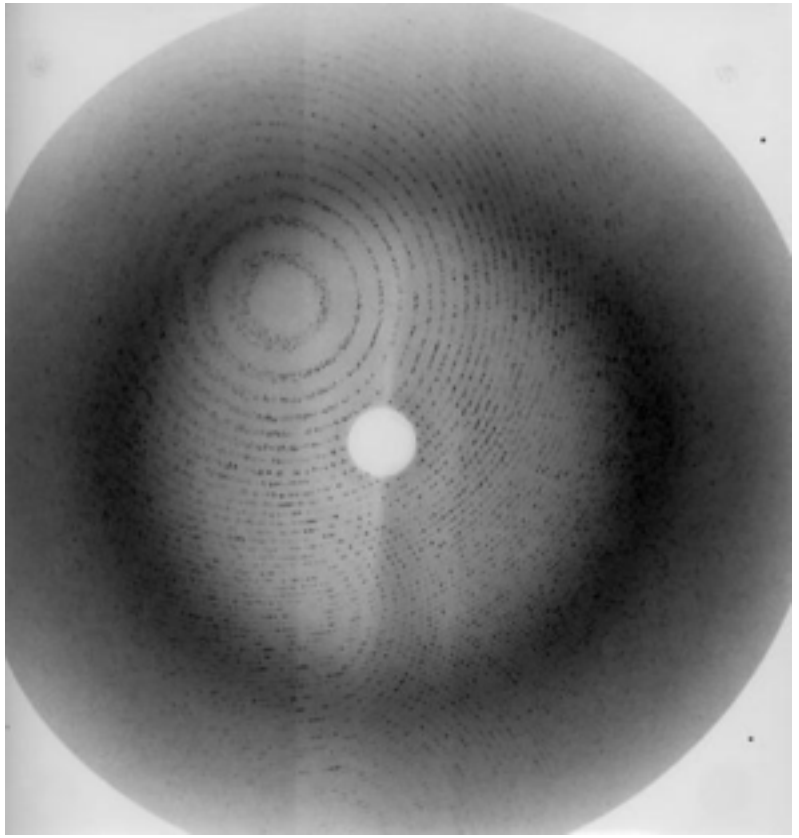
Source: http://www.synchrotron-soleil.fr/images/Image/Omars_structure2.png

X-Ray Crystallography Overview

- **Procedure Overview**
  - Pure high-concentration sample crystallized (e.g. protein)
  - Shine X-rays on crystals (diffraction)
- **Goal:** Obtain 3D Molecular Structure
- **Relevant Application:** Experimentally determining the structures of proteins and other biological structures

X-Ray Crystallography Setup

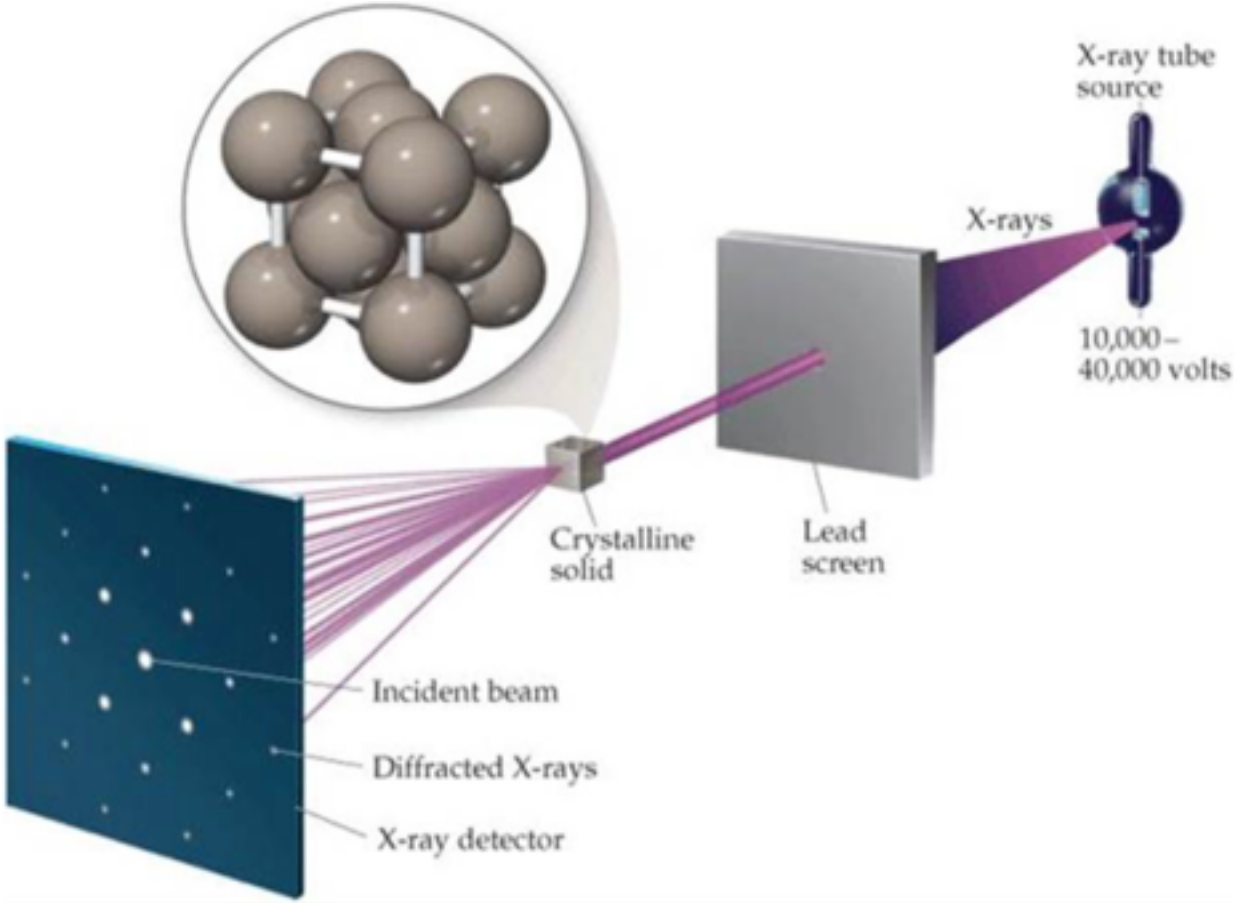




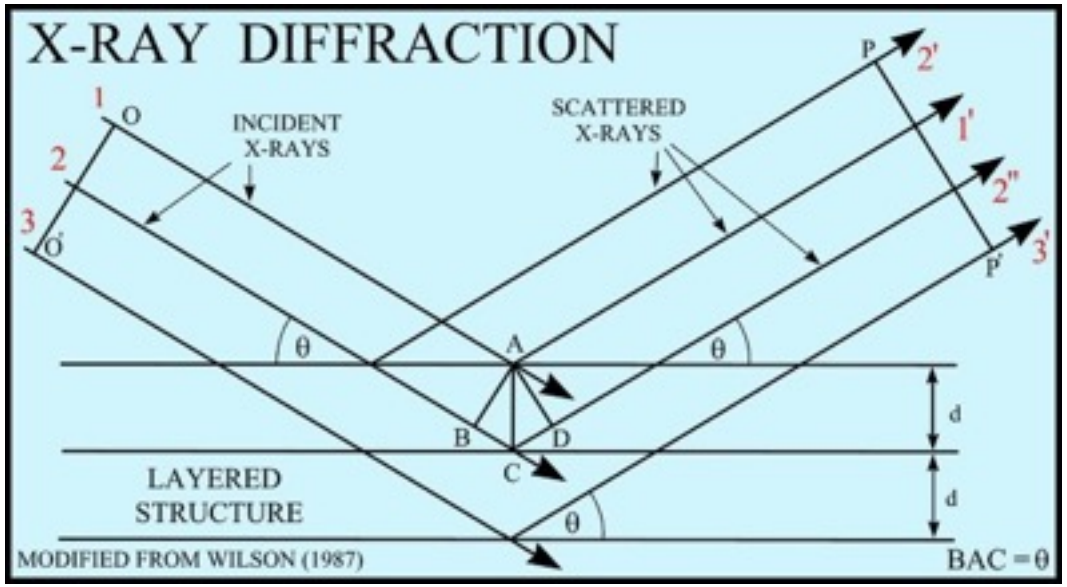
Source: http://web.chem.ucla.edu/~harding/ec_tutorials/tutorial73.pdf x-ray_cryst_setup

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1186895/\#__sec5titl

X-Ray Crystallography Setup



Source: http://schoolbag.info/chemistry/central/109.html



X-Ray Diffraction.

Source: https://pubs.usgs.gov/of/2001/of01-041/htmldocs/xrpd.htm

X-ray Crystallography Overview

- **How do we determine structure from the scatter pattern**?
  - Inverse Fourier transform takes the amplitudes and phases and returns the electron-density map.

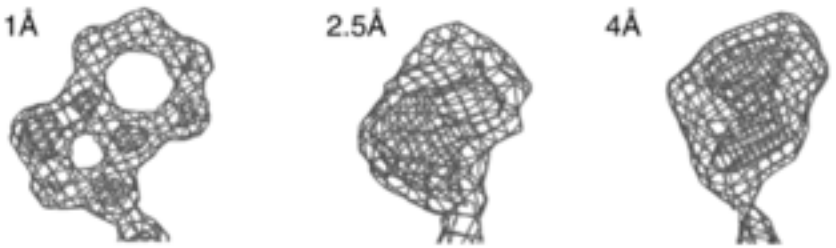


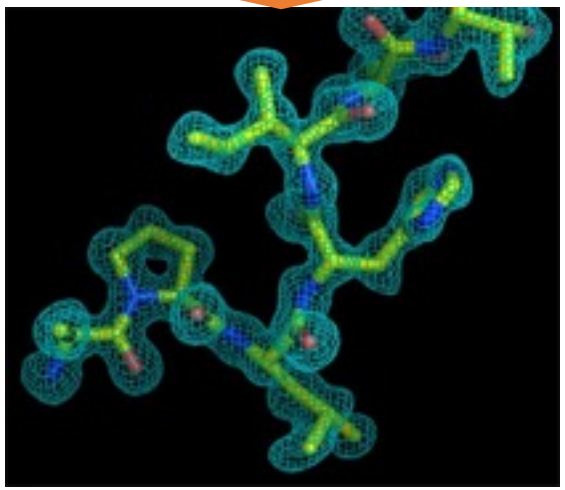
Figure: Varying resolution on electron density map of tryptophan sidechain. Bioinformatics. 2007;23(21):2851-2858. doi:10.1093/bioinformatics/btm480

- **Crystallographic resolution**: minimum spacing of crystal lattice planes that still produces discernible diffraction of x-rays.

‘Typical’ Approach

- Grow “well-ordered” crystals
  - Required for high-resolution diffraction patterns
- Get Diffraction Patterns using Crystallography
- Compute Electron Density Map
  - 3D Grid within unit cells, computing electron density at each point
- Manually (or Computationally) build structural model using the Electron Density Map
  - Homology modeling is common computational approach

Interpretation



Electron Density Map

Source: http://www.xtal.iqfr.csic.es/Cristalografia/archivos_07/densidad-mapa2.jpg

Source: Smyth, M. S., & Martin, J. H. J. (2000). *x* Ray crystallography. *Molecular Pathology*, *53*(1), 8–14.

nature Vol 464|22 April 2010|doi:10.1038/nature08892

LETTERS

Super-resolution biomolecular crystallography with low-resolution data

Gunnar F. Schröder[1,2], Michael Levitt[2] & Axel T. Brunger[2,3,4,5,6]

Presenter: Hari Ravichandran

CS 371 – Professor Ron Dror – February 6, 2017

Overview

- Published in *Nature* in 2010
- When crystallizing larger biological systems such as ribosomes, diffraction generally yields low resolutions (> 4 Å)
- Require new methods, as current (2010) methods need high-resolution initial structure
- This paper uses homologous structures with some allowances for modifications – e.g., “wiggle room”
  - **Key Assumption:** As a protein evolves and its amino acid sequence changes, its structure (at least locally) will more often than not remain very similar or even identical
- **Results:** Refining low-resolution structures yields significant improvements
  - Applications in studying crystals that diffract weakly

Paper Approach

- Low-resolution (~5 Å) X-ray diffraction (XRD) data is theoretically good enough to ascertain real sample structures
  - Have to find the torsional angles between each atom
  - Although a conformational search (in which we vary the torsional angles until we find a fit) would theoretically work, it is too computationally demanding
- Approach uses other information to narrow possibilities
  - **General**: Ideal bond lengths, bond angles, atom sizes
  - **Specific**: Homolog Information from a “reference model”

Deformable Elastic Network (DEN)

- **Problem**: Real structure is different from homolog structure
- Need to describe these differences mathematically
- Enter the Deformable Elastic Network (DEN) Approach
  - Implemented in Crystallography and NMR System (CNS) Software
  - Selects atom pairs and defines springs between them
  - Equilibrium spring length set to distance between atoms
  - MD Simulation runs, changing torsional conformations, recalculating energies as stipulated in Equation (1), and adjusting based on reference model

$$E_{\text{total}} = E_{\text{geometric}} + w_{\text{a}} E_{\text{ML}} + w_{\text{DEN}} E_{\text{DEN}}(\gamma) \qquad (1)$$

- **Requirements**
  - At least 30% sequence similarity for homologs
  - High resolution homolog (< 3.5 Å)

Results – DEN Refinement of Synthetic Structures

Table 1 | DEN refinement improves structures refined against four synthetic data sets

Target function	Resolution (Å)	R_{free}		
		DEN	noDEN	Improvement
MLHL	3.50	0.331	0.357	0.0256
MLHL	4.00	0.322	**0.328**	*0.0058*
MLHL	4.50	**0.293**	0.358	0.0651
MLHL	5.00	0.300	0.400	**0.0991**
MLF	3.50	*0.378*	0.390	0.0123
MLF	4.00	0.347	0.391	0.0445
MLF	4.50	0.348	0.413	0.0655
MLF	5.00	0.341	*0.425*	0.0841
Average	4.25	0.332	0.383	0.0503
Minimum	3.50	**0.293**	**0.328**	*0.0058*
Maximum	5.00	*0.378*	*0.425*	**0.0991**

Source for MLF: Pannu, S. N. & Read, R. J. (1996). Improved structure refinement through maximum likelihood. *Acta Crystallogr.* A 52, 659–668.
Source for MLHL: Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). Incorporation of prior phase information strengthens maximum-likelihood structure refinement. *Acta Crystallogr.* D 54, 1285–1294.

- Synthetic Data Sets
  - Compared DEN vs. NoDEN
- Target Function
  - Least-squares – "Traditional Function"
  - MLF – Max Likelihood Function
  - MLHL – Max Likelihood Function that takes phase information into account
- R_{free} – measure of 'goodness of fit' of predicted vs. actual structure
  - Low R_{free} more favorable
- Values in Bold are the best values for that column
- In every case shown, DEN-based model better than no DEN

Results – DEN Refinement of PDB Structures

Table 2 | DEN refinement improves low-resolution structures in the PDB

PDB ID	Resolution (Å)	No. of residues	R_{free}			Comments on differences
			DEN	noDEN	Improvement	
1AV1	4.00	804	0.335	0.336	0.0012	
1ISR	4.00	448	**0.233**	**0.237**	0.0043	
1JL4	4.30	557	0.353	0.354	0.0009	
1PGF	4.50	1,102	0.284	0.295	0.0108	Small throughout the chains
1R5U	4.50	3,517	0.334	0.335	*0.0003*	
1XDV	4.10	1,517	0.358	0.367	0.0089	
1XXI	4.10	3,532	0.407	0.465	**0.0582**	Large (~4 Å domain motions)
1YE1	4.50	574	0.312	0.350	0.0381	Small throughout
1YI5	4.20	1,356	0.323	0.336	0.0139	Local in several chains
1Z9J	4.50	821	0.317	0.331	0.0135	Large in chain A (domain motion)
2A62	4.50	319	0.340	0.353	0.0131	
2BF1	4.00	304	*0.479*	*0.492*	0.0131	
2I36	4.10	962	0.387	0.401	0.0137	Local in chain B
2QAG	4.00	702	0.392	0.401	0.0091	
2VKZ	4.00	10,941	0.327	0.337	0.0095	Large in subdomain placements
3BBW	4.00	543	0.304	0.334	0.0304	Significant local
3CRW	4.00	485	0.324	0.338	0.0136	Large in one domain (hinge motion)
3DMK	4.19	2,127	0.407	0.428	0.0211	Throughout, ref. model only 50%
3DU7	4.10	1,839	0.332	0.336	0.0039	
Average	4.19	1,708	0.345	0.359	0.0146	
Minimum	4.00	304	**0.233**	**0.237**	*0.0003*	
Maximum	4.50	10,941	*0.479*	*0.492*	**0.0582**	

- R_{free} – measure of 'goodness of fit' of predicted vs. actual structure
  - Low R_{free} more favorable
- Known Structures from Protein Data Bank (PDB)
- PDB ID - represents a Protein in the Data Bank
- Values in Bold are the best values for that column
- In every case shown, DEN-based model better than no DEN

Results – Summary & Electron Density Map

- The R_{free} values using DEN were all better than the ones using no DEN
  - Better Model Fit using DEN
- The use of DEN in addition to homology modeling allows for a better prediction of protein structures than previous approaches
  - 4% improvement in R_{free}

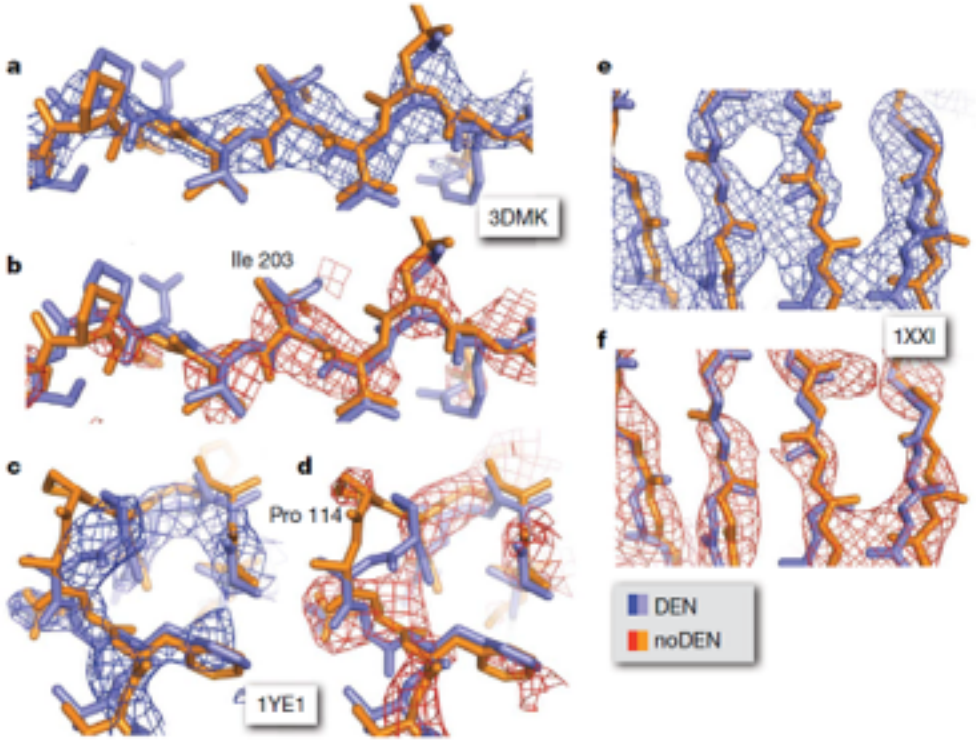


Figure 3 from Paper.

Strengths of Paper

- DEN method a significant improvement over previous methods
  - Uses homolog comparison to increase computational viability
- Verified method by testing against known structures
- Clear, practical research applications
  - Experimental determination of larger biological structures such as ribosomes
  - X-ray crystallography, cryo-electron microscopy, and potentially optical imaging once its resolution is high enough

Weaknesses of Paper

- Communication Style
  - Paper written for experts in the field
- Technical
  - DEN Refinement is very useful for larger deformations, but not so much for smaller changes in structure (Figure 4)
  - Variations in homologous structure families
  - Further refinement needed

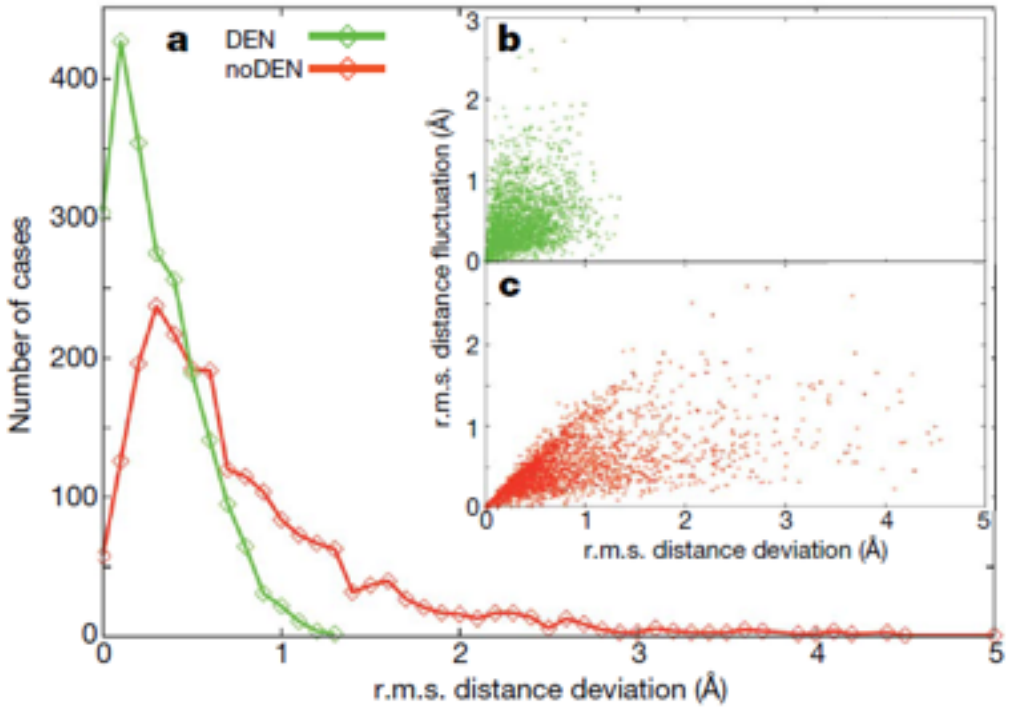


Figure 4 From Paper.


Question

New Methods for Solving Tough Crystal Structures

Daniel Byrnes

February 6, 2017

X-Ray Crystallography: From Electron-Density Map to Protein-Structure

- **Task**: Trace the protein sequence of amino acids through the 3D electron density map.
- Difficult protein structures and *low-quality* density maps can require a great deal crystallographer effort.
- It would be great if we could automate this process!

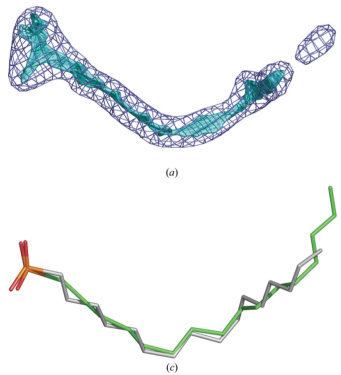


Figure: Figure 5 from "Automated crystallographic ligand building using the medial axis transform of an electron-density isosurface"

Using Low-Resolution Density Map to Determine Protein-Structure

Problem (Quality of Map)

Low resolution of electron-density map makes tracing the protein difficult and time consuming.

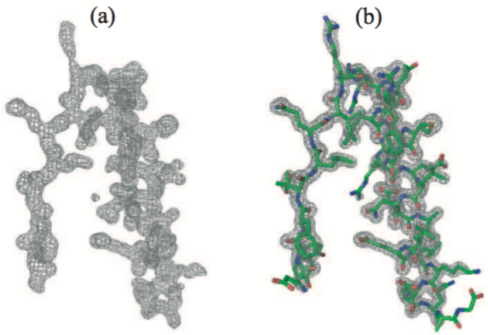


Figure: Left: Electron density map of protein. Right: Non-hydrogen atoms of protein-structure that fits map. (Figure 2 from paper)

Using Low-Resolution Density Map to Determine Protein-Structure

Guiding Belief Propagation using Domain Knowledge for Protein-Structure Determination

Ameet Soni
Dept. of Computer Sciences
Dept. of Biostatistics and Medical Informatics
University of Wisconsin
Madison, WI 53706
soni@cs.wisc.edu

Craig Bingman
Dept. of Biochemistry
Center for Eukaryotic Structural Genomics
University of Wisconsin
Madison, WI 53706
cbingman@wisc.edu

Jude Shavlik
Dept. of Computer Sciences
Dept. of Biostatistics and Medical Informatics
University of Wisconsin
Madison, WI 53706
shavlik@cs.wisc.edu

- Overview: This paper expands upon a previously developed probabilistic model to trace a protein backbone in poor quality maps ($\sim$ 3 to 4 Å resolution).
- Automated Crystallographic Map Interpretation (ACMI)
- Improved Belief Propagation Protocol

Automated Crystallographic Map Interpretation (ACMI)

- Probabilistic framework to sample all-atom protein-structure models.
- Three-phase process of ACMI pipeline:

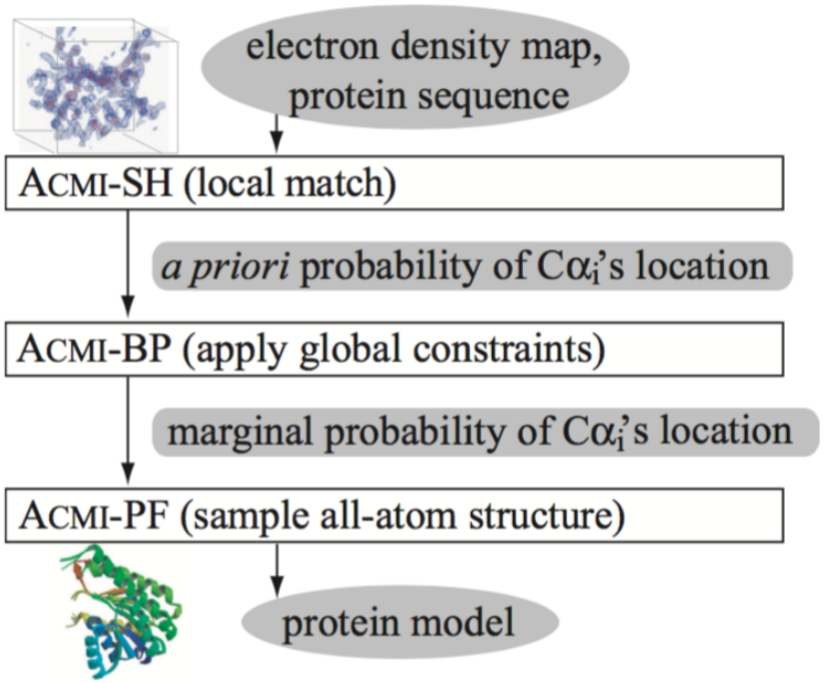


Figure: Figure 3 from paper

Novelties of ACMI Methodology

- Ties local density information and global constraints to infer possible locations of residues.
- Each residue's location is represented as a distribution over the entire electron-density map.
- To do this, ACMI uses a probabilistic graphical model:
  - **Pairwise Markov-field model (MRF)**.
- MRF allows us to probabilistically represent all possible structures in a compact manner and perform inference on subsets of the graph.

Graphs and Pairwise Markov-Field Models

- Markov Random Field is an undirected graphical model that defines a probability distribution on a graph.
- Vertices are associated with random variables.
  - Vertex i corresponds to the i^{th} amino acid in the sequence.
  - Random variables describe the location, $\vec{u_i}$, of each C_α.
- Undirected edges form pairwise constraints on connected random variables.
  - An edge exists between each vertex to signify 3D folding constraints.

Markov Random Field: Example

"Based on my current belief, I would expect you to be located (with probability) here."

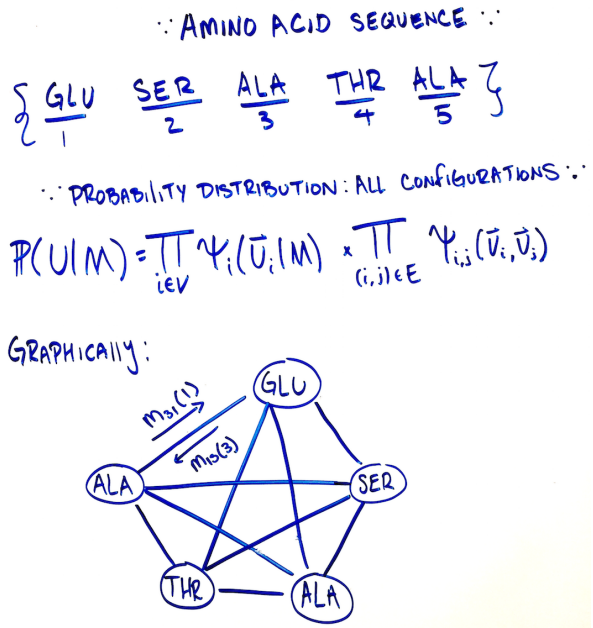


Figure: Markov Random Field for example amino acid sequence. Graph represents the full-joint probability distribution over all possible configurations for all residues in the target protein.

ACMI phase I: Local Match

$$\mathbb{P}(U|M) = \prod_{i \in V} \psi_i(\vec{u_i}|M) \times \prod_{(i,j) \in E} \psi_{i,j}(\vec{u_i}, \vec{u_j})$$

- $\psi_i(\vec{u_i}|M)$ (observation potential function): prior probability on the location of an amino acid given map M. Ignores all other amino acids in protein.

- $\psi_{i,j}(\vec{u_i}, \vec{u_j})$ (Edge potential function): global constraints on protein structure.
  - Adjacency potential: adjacent residues must maintain $\sim 3.8\,\text{Å}$ spacing and proper angles.
  - Occupancy potential: no two residues can occupy the same space.

ACMI phase 2: Belief Propagation

- We cannot calculate this probability in large graphs with cycles!
- ACMI uses **loopy belief propagation** to approximate marginal probability distribution.
- This paper focuses on improvements in ACMI-Belief Propagation (BP) phase.

ACMI phase 3: Particle Filtering

- Recall: the resulting marginal probability distribution describes the location of the C_α in each residue.
- This does not account for residue side chains, only the backbone.
- Phase 3 uses a sequential sampling algorithm (**particle filtering**) to produce a physically feasible, all-atom, protein structure.

ACMI and Belief Propagation

- Iterative process: for each vertex (amino acid) compute the marginal distribution over locations in the unit cell using local probability and incoming messages.
- Then compute the outgoing messages to the connected neighbors of each vertex.

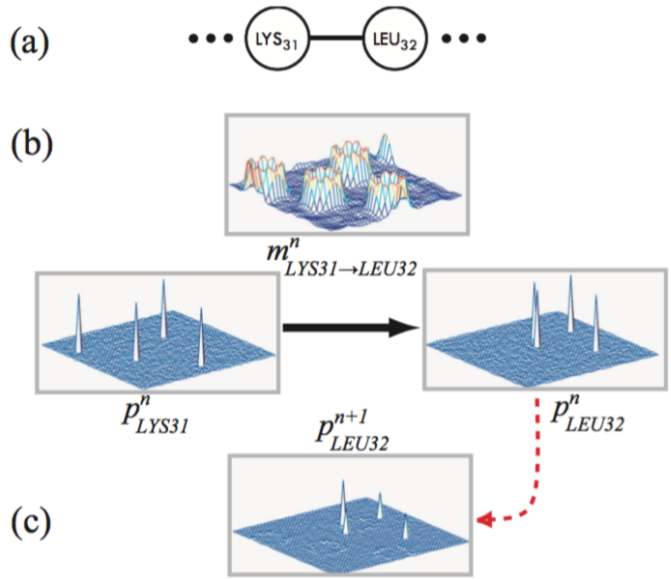


Figure: Figure 5 from paper. Lysine sending a message to Leucine. Notice change in confidence of each peak after message is sent.

ACMI-BP Message Scheduling

- Belief propagation algorithm requires a message passing protocol.
- Message scheduling protocols:
  - Round-robin: each vertex treated equally; no priority based on evidence of information gain.
  - Residual Belief Propagation: prioritize messages with the most new information.
  - **Domain Knowledge**: well-structured regions of protein sequence are more likely to contain accurate information regarding local conformation.
- Domain knowledge approach prioritizes random variables deemed a priori more accurate.

Belief Propagation and Domain Knowledge

- Domain knowledge approach prioritizes residues that are likely to be in well-structured regions of the final 3D solution.
- Decay factor in probabilities allows less reliable amino acids to work up the queue.

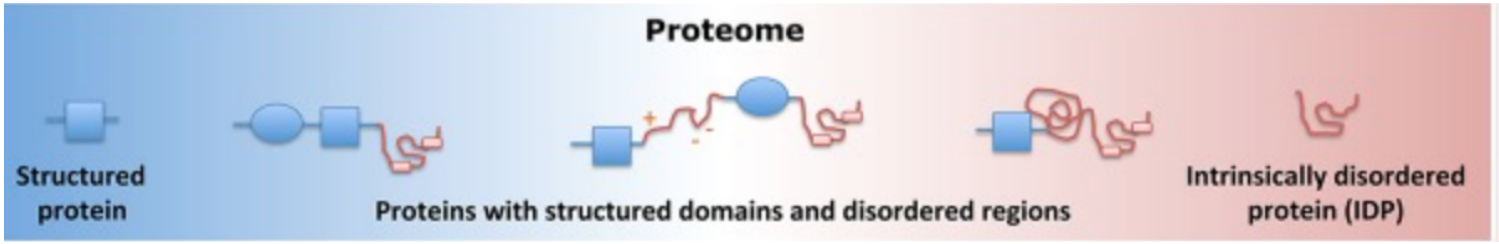


Figure: Figure 1 from Chem Rev. 2014 Jul 9; 114(13): 6589?6631.

Experimental Setup

- Dataset consists of 10 "difficult" experimentally-phased electron density maps.
- Each test only differs in which message scheduling protocol is used during ACMI-BP.
- Each of the three ACMI-BP algorithms was used to produce a marginal probability distribution.
- ACMI-PF then samples all-atom structures from the marginal probability distributions produced by phase 2.

Experimental Results

- Each point represents one of the 10 protein structures in the dataset.
- The **Rank** of a (correct) residue is defined as the fraction of points in the marginal probability distribution that have greater probability than the true solution. The rank for all residues were averaged for each protein.
- Rank metric allows us to compare prediction results across differing probability space sizes for each protein.

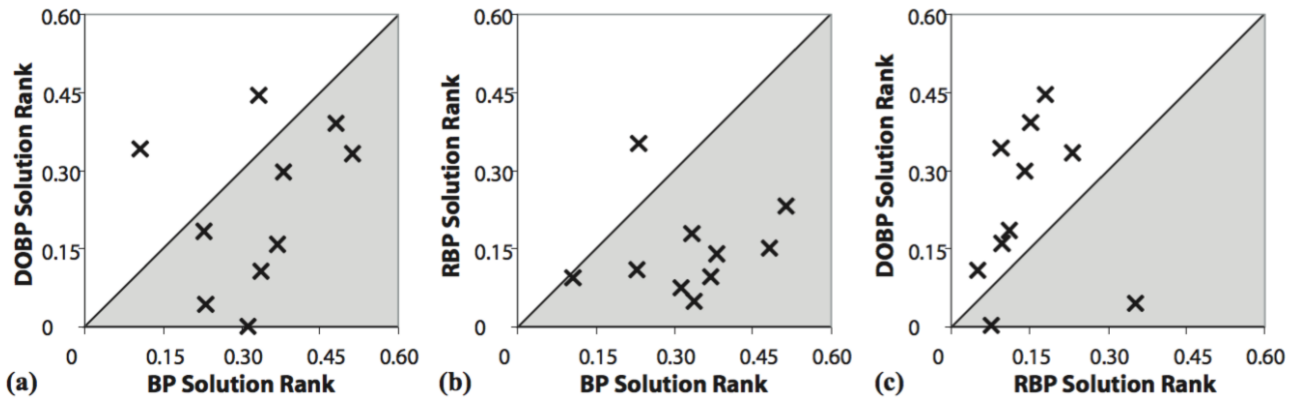


Figure: Figure 7 from paper.

Experimental Results

- ACMI-PF was used to sample physically-feasible protein structures from the set of marginal probability distributions returned from the belief propagation phase.
- ACMI-PF fails to sample results produced by RBP protocol.

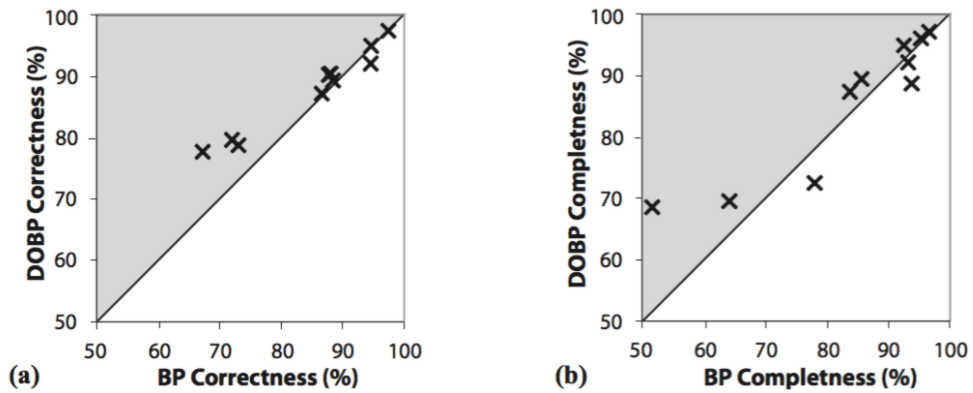


Figure: Figure 8 from paper. Correctness and completeness of predicted protein structures using marginal probability distribution produced by BP and DOBP message scheduling protocol.

Paper Critique

- Disordered proteins are abundant in eukaryotic cells; if a protein does not have enough well-structured regions then the domain-knowledge based priority function might be insufficient to push belief propagation towards convergence.

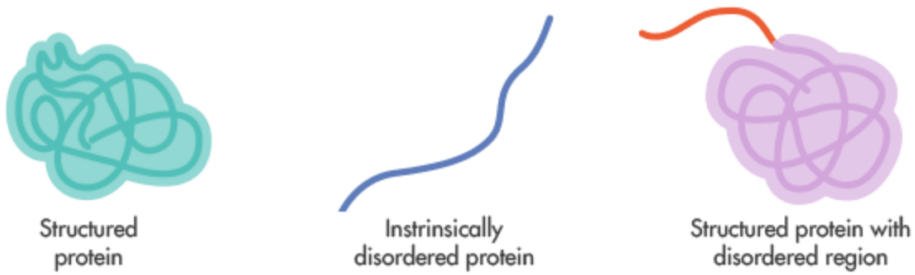


Figure: Lucy Reading-Ikkanda/Quanta Magazine, "The Shape Shifting Army Inside Your Cells".

- Develop a method to filter the locations with non-negligible probabilities returned by the residual belief propagation (RBP). This message scheduling protocol seems promising, but ACMI-PF requires a smaller search space within density map.

References I

Soni, Ameet and Bingman, Craig and Shavlik, Jude, "Guiding Belief Propagation using Domain Knowledge for Protein-Structure Determination", (August 02, 2010). In Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology (BCB '10). ACM, New York, NY, USA, 285-294.; available from http://dx.doi.org/10.1145/1854776.1854816.

Soni, Ameet and Shavlik, Jude, "probabilistic ensembles for improved inference in protein-structure determination", Journal of Bioinformatics and Computational Biology. 2012;10(1):1240009. doi:10.1142/S0219720012400094.

Bin Xue and A. Keith Dunker and Vladimir N. Uversky, "Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life", Journal Of Biomolecular Structure And Dynamics Vol. 30 , Iss. 2,2012.; available from http://dx.doi.org/10.1080/07391102.2012.675145.

Frank P. DiMaio, Ameet B. Soni, George N. Phillips, and Jude W. Shavlik. "Spherical-harmonic decomposition for molecular recognition in electron-density maps", Int. J. Data Min. Bioinformatics 3, 2 (May 2009), 205-227. DOI=http://dx.doi.org/10.1504/IJDMB.2009.024852