CHEMISTRY 456 / INTERNATIONAL WORKSHOP

Structure Determination

Andy Howard
Associate Professor of Biology & Physics,
Illinois Tech

Monday 28 June 2021

WHERE DID ALL THESE LOVELY STRUCTURES COME FROM?

- Structures as studied by computational chemists are based on experimentally-determined structures
- Determining the structure of a big molecule is a challenging endeavor no matter how you do it.

WHAT WE'LL DISCUSS

- Structure-function relationships
- High-resolution methods
  - X-ray crystallography
  - High-field NMR
  - Cryoelectron microscopy
- Additional methods
  - Solution scattering
  - Circular dichroism
  - Fiber diffraction
- The PDB's role
- Applications in drug design
- Applications in molecular dynamics

STRUCTURE-FUNCTION RELATIONSHIPS

- If you wish to sound wise in a room full of structural biologists, train a parrot to say *structure-function* over and over.
- But the point is that if we know something about the structure of a macromolecule, we can use that knowledge to deduce useful insights into how it works.
- From there we can go on to design ligands or do simulations

EXPERIMENTS AND MODELS

- Computational chemists and biologists are accustomed to regarding the entities they analyze as models rather than primary realities
- But even the experimentally derived structures that we're talking about today are models.
- The task of the experimentalist is to obtain data and then build a model that is as consistent as possible with those data without over-fitting!

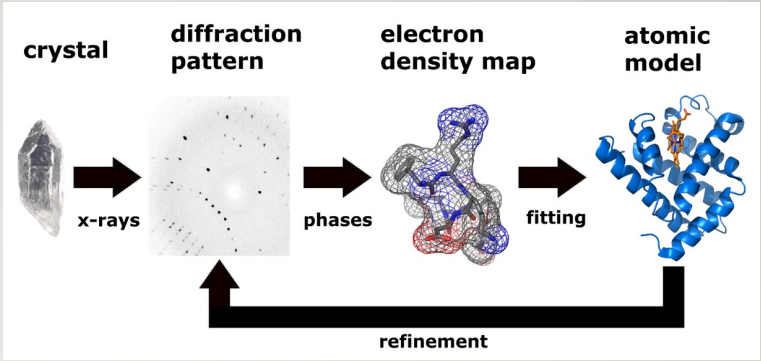
RESOLUTION

- Resolution refers to the size of the smallest detail that is discernable within a structural model
- Thus high resolution is associated with a small value of the resolution, and low-resolution goes with a big value; and high resolution is almost always better than low.
- We can characterize any structure according to the resolution with which we can distinguish neighboring model entities (atoms, residues, subunits, …)
- With some methods it's easy to characterize the resolution of the structural models; not with some others

HOW HIGH A RESOLUTION IS REQUIRED?

- That depends on what you want to do with the structure.
- For drug design we generally need near-atomic (~1.8Å resolution so we can tell how the ligand is going to fit into the structure
- We might want even better resolution if we want to analyze alternate conformations or solvent dynamics
- For molecular dynamics, we need high resolution if we're trying to do QM in parts of the structure or if we want to begin the simulation with the atoms precisely located; but if we only want to model big changes, we can make do with 3Å or even 5Å resolution

X-RAY CRYSTALLOGRAPHY

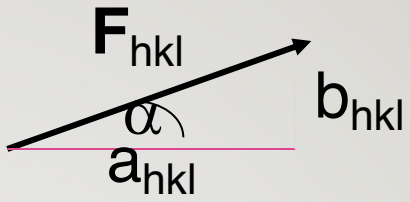


- Crystals are translationally ordered arrays of molecules, and most solids are crystalline
- Entraining an X-ray beam at a crystal causes coherent interference in specified directions (diffracted rays) associated with contents of repeating unit of crystal
- If you intercept the diffracted rays with a detector, you can measure intensities of spots ("Bragg reflections")
- Each intensity is proportional to the square of a complex Fourier coefficient associated with the structure

CRYSTALLOGRAPHY ON MACROMOLECULES

- The physics of diffraction isn't different for macromolecules
- What *is* different?
  - More data to a given resolution
  - Fewer unit cells in the crystal's volume
  - Static and dynamic disorder limit resolution to considerably higher than $\lambda/2$
  - Phasing methods that work for small molecules don't work for macromolecules

SOLVING THE PHASE PROBLEM



- Early structures determined by soaking heavy atom compounds (Hg, Pt, …) into the crystal and carefully measuring $F_{PH} - F_P$
- That can tell you where the heavy atom is, and then you can bootstrap to find the light atoms (C, N, O, S)
- Alternative: vary the X-ray wavelength and measure $F(\lambda_1) - F(\lambda_2)$
- OR survey the PDB for similar structures and use one or more of those as a rotated & translated start model

HOW DO YOU KNOW THE CRYSTAL STRUCTURE IS INFORMATIVE?

- This was a big concern in the early days of protein crystallography: is the solution structure really all that similar to the crystal structure?
- Answer: generally yes, especially on the interior
  - Crystalline proteins still have biological activity!
  - Multiple crystal forms yield the same structure
  - Structures determined by other methods look similar
- *Interior* relevant for enzymes: active sites are inside

WHAT CAN YOU LEARN?

- (X,Y,Z) coordinates of atoms in the macromolecule
- (X,Y,Z) coordinates for well-ordered water molecules
- Mobility and fuzziness of individual atoms
- Multiple conformations visible at high-resolution

```
ATOM   64 CB AVAL A 10 13.809 27.281 23.136 0.50 13.07   C
ANISOU 64 CB AVAL A 10 1479 727 2759 71 -70 297          C
ATOM   65 CB BVAL A 10 13.611 27.375 22.837 0.50 11.05   C
ANISOU 65 CB BVAL A 10 1470 952 1778 -14 -121 4          C
```

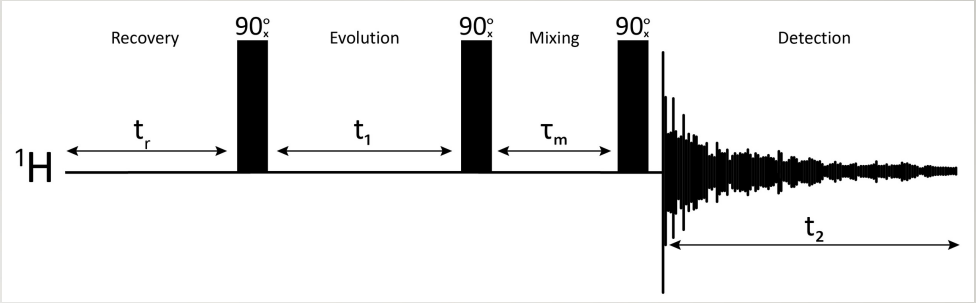
STATIC & DYNAMIC STRUCTURES

- Initially, crystal structures were static or time-averaged images of the macromolecule, and that's often still true
- Some approaches are now available that enable us to look at each distinct step within an protein's mechanism: a movie, or at least a slide-show, of the mechanism
- Part of what makes that possible is the availability of free electron laser facilities like the LCLS at Stanford, where structures can be derived from 10^{-15} sec X-ray bursts

HIGH-FIELD NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

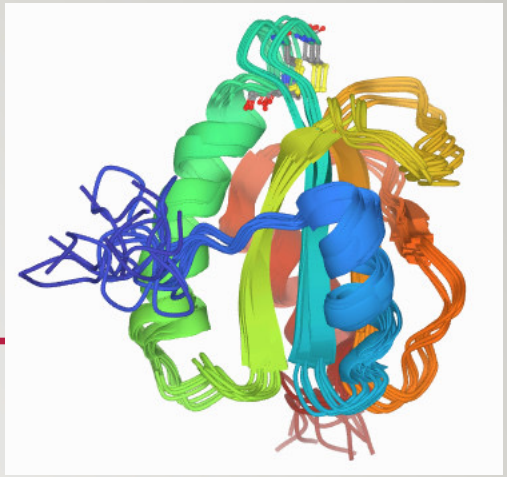
- NMR relies on resonant interactions between an applied, microwave-frequency electromagnetic field and unpaired nuclear spins in a sample held in a strong magnetic field
- For small molecules each individual resonance can be identified as arising from a specific atom or cluster
- Simple spectra from big molecules are uninterpretable: too many peaks, and they overlap disastrously
- So we have to resort to multiple pulse trains or examine two or more resonating atoms, typically ^{1}H and ^{15}N

INTERATOMIC DISTANCES



- Most productive approach with big molecules uses Nuclear Overhauser Effect measurements (NOEs): measure the change in one resonance when another resonance is saturated
- Enables measurement of interatomic distances between the two affected spins
- Many measurements required and a considerable amount of hypothesis-testing to assign specific NOESY peaks to specific atom-atom pairs
- Ultimately these can be brought together to define coordinates for H (and N) atoms; bootstrap the rest.

FAMILIES OF STRUCTURES



Fungal protein disulfide
PDB 2RUE

- Ultimate outcome for many macromolecular experiments is a *family of structures*, typically with certain atoms having well-defined locations and multiple paths between those atoms
- Why this lack of specificity?
  - It partly reflects real heterogeneity in the conformations in solution
  - It also results from the structure-definition problem being under-determination

COMPARING X-RAY TO NMR

- Big overlap in the achievable domains of these two methods (5 kDa – 60 kDa for NMR; 1 kDa – 500 kDa for Xray)
- So what do you do when they differ?
  - Trust the surface in the NMR structure because it isn't influenced by intermolecular contacts
  - Rely on H's in NMR because the data are based on them
  - Trust the X-ray structure elsewhere because it's probably obtained at higher resolution

POLITICS AND ECONOMICS

- NMR equipment capable of measurements on > 20 kDa proteins have > 17 T magnets, i.e 700 MHz; that's a $1M instrument, and it costs $100K/yr to maintain it
- State-of-the-art Lab X-ray equipment costs about the same
- BUT most X-ray structures are determined at government-funded storage rings or FELs, which (for non-commercial users) are essentially free!
- So one can equip an "X-ray" lab without X-ray equipment

ELECTRON MICROSCOPY

- EM: long history of imaging biological samples at organelle size-scale (> 10 nm)
- ~0.5 nm-resolution EM on macromolecules mostly failed before 1985 because of sample damage
- Two experimental advances enabled ~0.4 nm by 1990:
  - Cooling the sample (< 100K): slows sample damage
  - Plating samples on a 2D grid and using image software to merge the images in 3D at improved resolution

MATURATION OF CRYO-EM

- 3 additional advances bring cryogenic EM into the near-atomic resolution range (2Å in ideal circumstances, 3Å in many cases, 4Å in most):
  - Highly sensitive and linear electronic detectors
  - Better sample-prep techniques
  - Software innovations that improve sample averaging and error correction
- Definitely the method of choice for large, multi-component protein or protein-RNA complexes
- Recent advances enable good outcomes even < 50 kDa

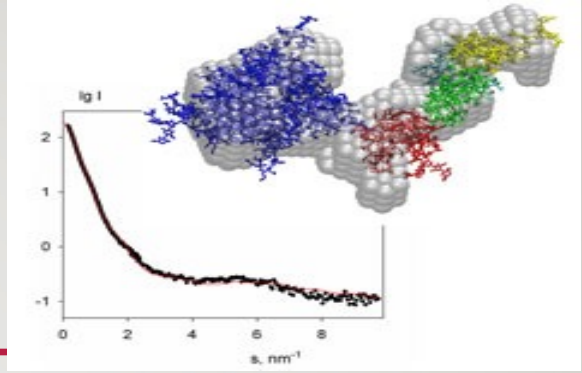
MOLECULAR LEGO

- Many biological systems consist of a collection of medium-sized polypeptides that together constitute a protein complex.
- Frequently it's possible to obtain a 3.5-4Å cryoEM structure of the intact complex
- Meanwhile, it's possible to obtain < 2Å X-ray crystal structures of the individual components
- Then the best way to use the data from both is to combine them: molecular lego

IS CRYO-EM MORE REALISTIC THAN CRYSTALLOGRAPHY?

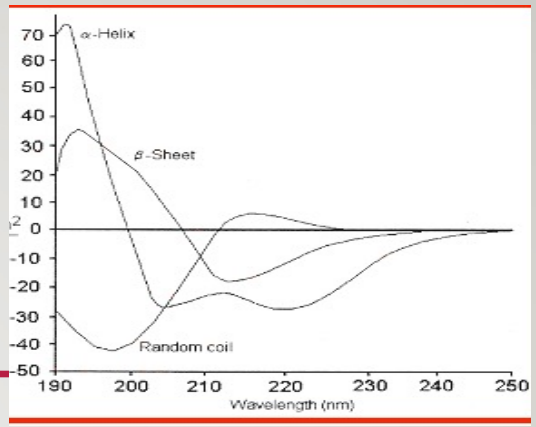
- Perhaps: CryoEM doesn't require growth of crystals
- But it does require exposure of biological samples to cryo conditions and dehydration on an EM grid
- CryoEM is probably less sensitive to impurities carried through to the structure determination step

SOLUTION SCATTERING



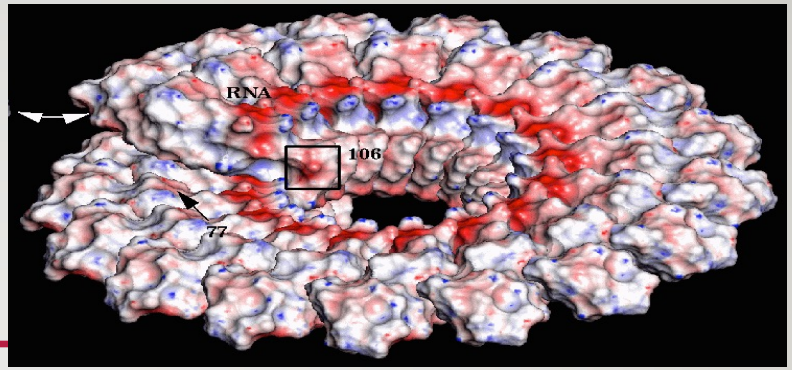
- Proteins and polynucleotides scatter X-rays or neutrons even when the macromolecules are in solution
- One-dimensional experiment because angular dependence is averaged out over many molecules
- Intensity as a function of angle yields low-resolution information about the size and shape of the scatterer
- Hypothesis-testing: "I think this molecule is shaped like a dumb-bell. Does the scattering pattern match that of a dumb-bell"?

CIRCULAR DICHROISM



- Biopolymers have asymmetric carbons in them: they're *chiral*
- Chiral molecules rotate circularly polarized light: direction and magnitude of the rotation varies with wavelength
- Secondary structure, particularly alpha helices, produce characteristic; spectra in the 190-250 nm range
- In practice this is usually used only to monitor helical content
- Sample holder can be heated, so you can monitor loss of helicity as f(temp)

FIBER DIFFRACTION



- Many biopolymers naturally form aggregates that have 2-D (fiber-like) order:
  - Actomyosin in muscles
  - Some viral coat assemblies
  - DNA (!)
  - Collagen (most plentiful human protein)
- Fiber diffraction patterns can be interpreted using math that is similar to 3-D crystals
- IIT and the APS have long been centers for fiber studies

ROLE OF THE PROTEIN DATA BANK

- Roughly 180,000 macromolecular structures are contained in the PDB: 90% proteins, 10% polynucleotides
- 89% determined by X-ray crystallography; rest by NMR, CryoEM, neutron diffraction, combined techniques
- CryoEM fraction is increasing rapidly
- Perhaps 40,000 independent structures; rest are mutants, ligand-bound, or redeterminations
- Drug designers and dynamicists use these as input data

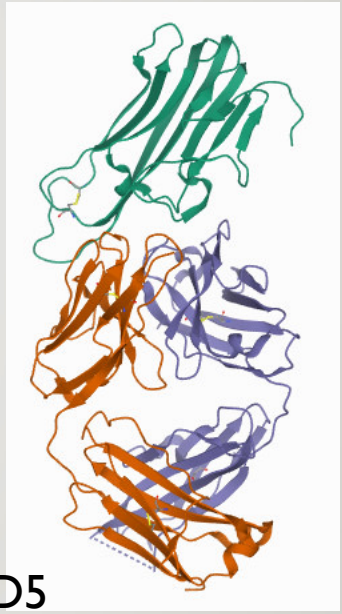
DRUG DESIGN

- Until about 2005 virtually all drug targets were enzymes
  - Essential proteins in pathogens
  - Human proteins that were overexpressed or expressed under the wrong circumstances
  - Proteins produced in cancer cells at higher levels than in normal tissue
- In each case, designing an effective inhibitor (competitive, mixed, or non-competitive) enabled a desirable health outcome

HOW HAS THAT CHANGED ?

6MU9

- For pathogens:
  constant concern over resistance
- Many drugs (the majority of the big sellers) now biologics, i.e. macromolecules, not small-molecule inhibitors
- ~ half of the biggest revenue-producers are monoclonal antibodies



3WD5

CONNECTION WITH THE REST OF THIS WORKSHOP

- Clearly, design of effective pharmaceuticals is one of the goals of this workshop
- Predicting effectiveness before (and after) doing wet-lab experiments is part of the point of molecular modeling and docking studies

MOLECULAR DYNAMICS

- Every MD simulation has to begin with a plausible structure
- Typical starting point is an experimentally-determined structure
  - Remove all or most of the water
  - Fill in the blanks and the hydrogens
  - Energy-minimize the structure before beginning molecular dynamics