Nucleic Acids

See the full presentation at:
https://dialpuri.github.io/NucleicAcidsPresentation

Jordan Dialpuri

York Structural Biology Laboratory

27/11/25

Nucleic Acids

What are they?

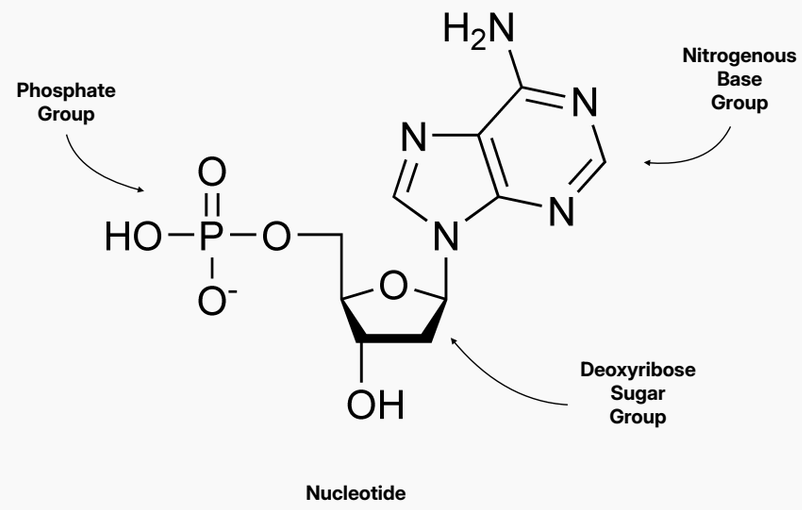
B-DNA
PDB Code: 1BNA

T-RNA
PDB Code: 1EHZ

Nucleic acids like DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), are polymers of nucleotides that carry genetic information and play essential roles in regulation, catalysis, and structural organisation within the cell.

Nucleic Acids

What are they?



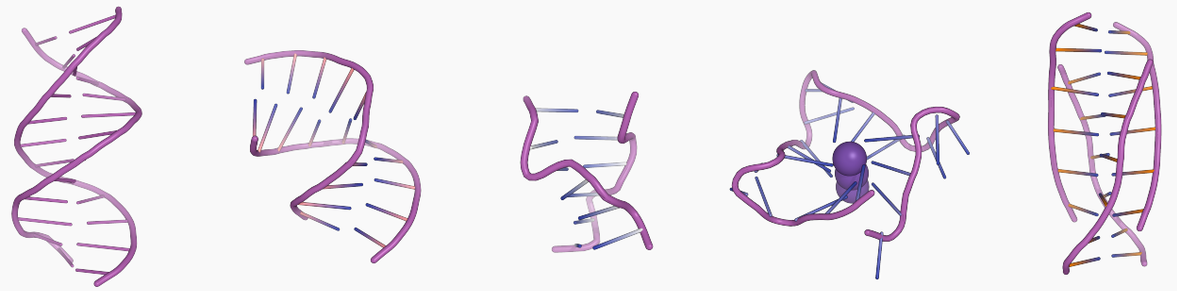
Nucleotide

Nucleic acids are polymers of monomeric units called nucleotides which consist of the phosphate (P atom coordinated by Oxygens), sugar (five membered ring ribose sugar) sugar and base groups (planar, aromatic bases), the base encodes the genetic information, and also allows for tertiary structure formation. The nucleotide, under normal conditions is negatively charged (anionic), leading to nucleic acids commonly having a polyanionic backbone.

Nucleic Acids

DNA

Many different possible DNA conformations



B-DNA
PDB Code: 1BNA

A-DNA
PDB Code: 1ANA

Z-DNA
PDB Code: 180D

G Quadruplex
PDB Code: 1KF1

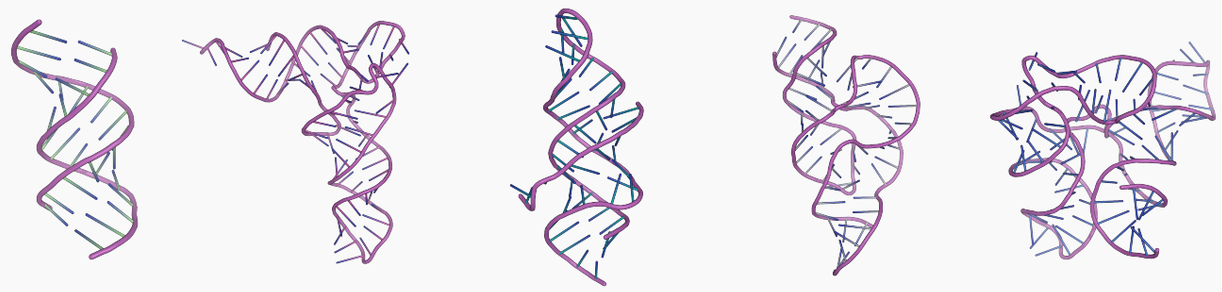
I Motif
PDB Code: 1I9K

When deoxyribose nucleotides form polymers, they form DNA but DNA exists in many more interesting conformations. Traditionally DNA is often thought of as exclusively existing in the B-DNA double helix, but the world of DNA is more complex than that.
Depending on the environment (the amount of water) around the DNA, this B-DNA form can transform into the A-DNA form which can be important in specific enzymatic processes.
But it is not just the environment that can influence the conformation of DNA, of course the sequence is incredibly important too.
If you have lots of CG (alternating purine-pyrimidines) repeats in your sequence, it may form the Z-DNA helix which is a left handed helix which has a zig-zag backbone formation and has been shown to be important in gene expression.
If you have lots of G (Guanine-rich) in your sequence, you may form a quadruplex (four strands) conformation with an internal cation, or if you have lots of C (Cytosine-rich) in your sequence you could form the I-motif which are both important in gene regulation.

Nucleic Acids

RNA

Many different possible RNA conformations



A-RNA
PDB Code: 1RNA

T-RNA
PDB Code: 1EHZ

Triple Helix
PDB Code: 3P22

Ribozyme
PDB Code: 3ZP8

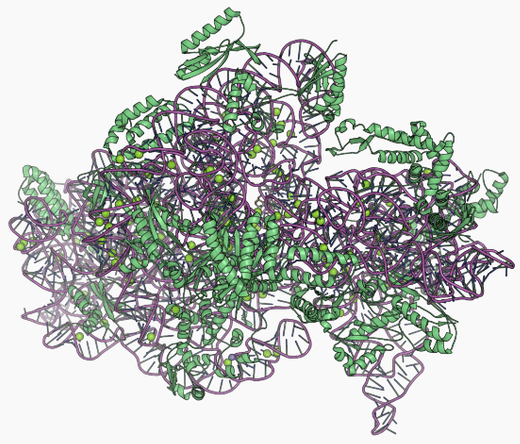
Riboswitch
PDB Code: 2GIS

DNA is very conformationally diverse, but the other common nucleic acid RNA, can exist in many more conformations.
When RNA forms a double stranded structure (like in some viral RNA), the conformation is known to be similar to that of A-DNA, but since RNA commonly exists in a single stranded form, the single strand can fold onto itself to create a diverse range of structures.
For example, T-RNA forms a cloverleaf structure from a single strand and is of course incredibly important in protein synthesis.
RNA (and DNA too) can form triple helices where a third RNA strand binds into the major groove of an RNA duplex, often to shield itself from degradation.
RNA molecules can also have catalytic activity (ribozymes) which are used for peptide synthesis or phosphodiester bond cleavage.
Or RNA can form switches which, upon binding of a metabolite, can change structure to regulate the expression of specific genes.

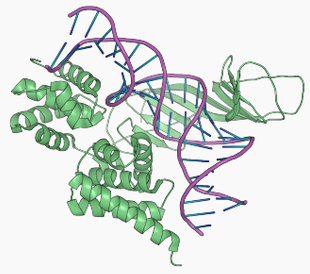
Nucleic Acids

Protein Complexes

Cas9 RNA/DNA Complex
PDB Code: 4UNS



30S Ribosomal Subunit
PDB Code: 1IBK



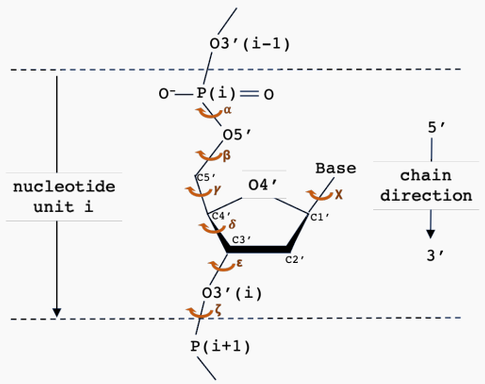
Transcription Factor DNA Complex
PDB Code: 1D3U

Although DNA or RNA can exist by itself, more often it is functionally relevant for DNA or RNA to interact with proteins for some specific function. One important example is of course the ribosome, here the 30S ribosomal subunit is shown, where ribosomal RNA is very important in the production of proteins. Proteins and nucleic acids are often in complex through electrostatic, hydrogen bonding, hydrophobic interactions and van Der Waals forces.

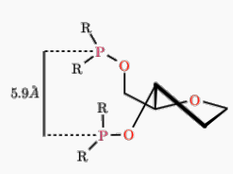
Nucleic Acids

Heterogenity and Flexibility

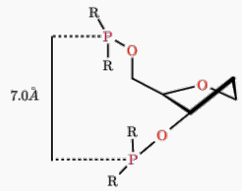
Multiple Torsion Angles



Sugar Pucker



C3′ Endo

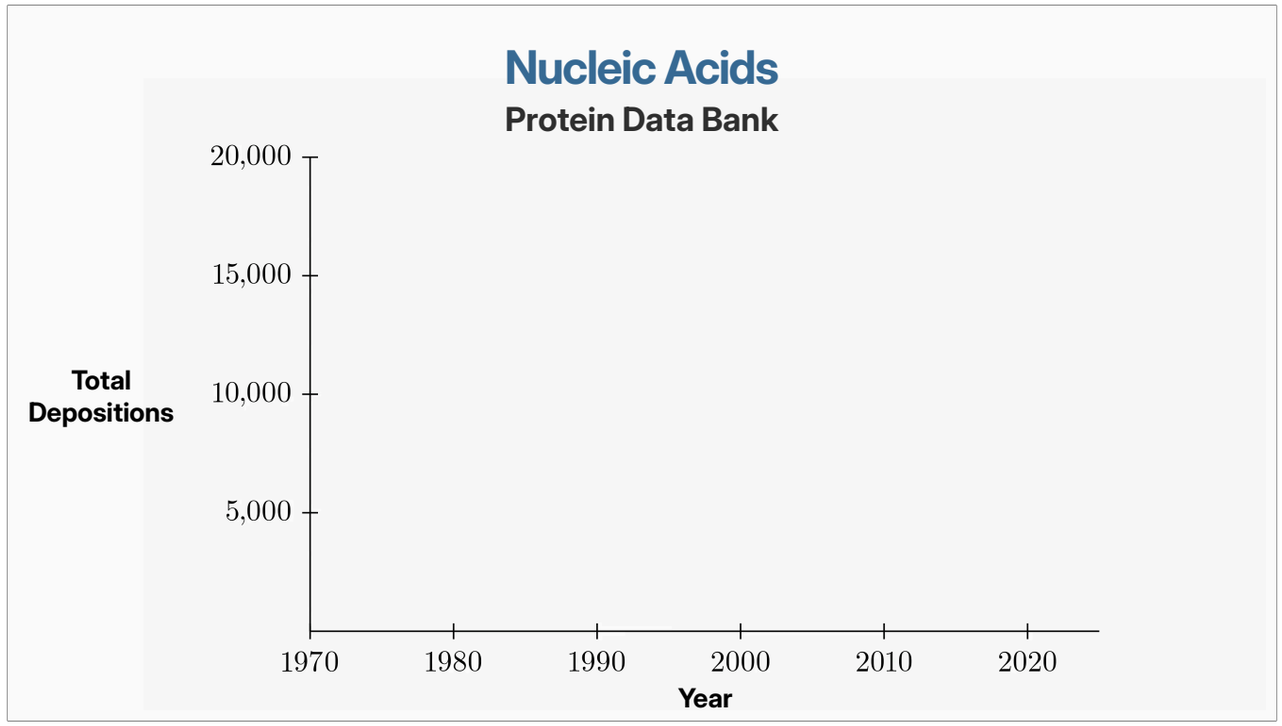


C2′ Endo

Lawson CL, et al. Nucleic Acids Research 52, D245-D254. doi:10.1093/nar/gkad957

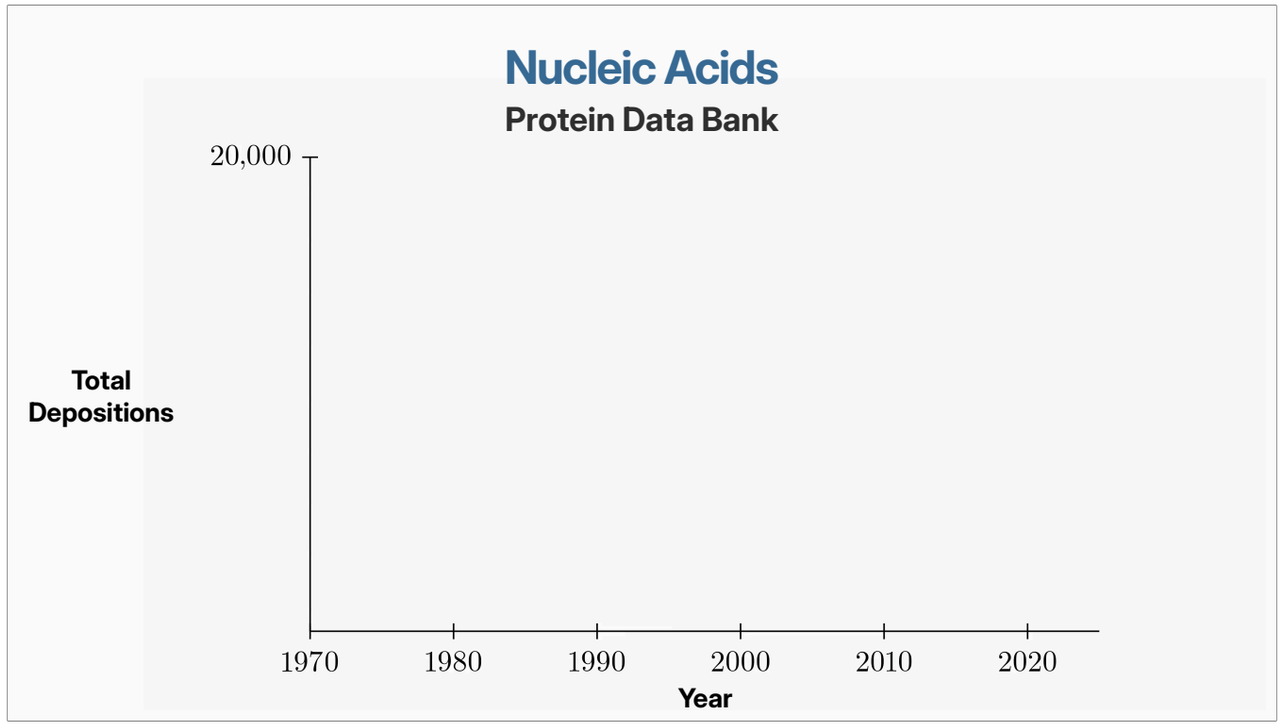
Nucleic acids are incredibly diverse in conformation and often they are also flexible structures, and there are two main geometric properties which allow for this heterogeneity and flexibility. The sugar-phosphate backbone of the nucleic acid, contains 6 important torsion angles (and therefore rotatable bonds) which allows for more conformational freedom in the backbone. Additionally, the sugar pucker, which defines which atom is out of the ring plane of the ribose, affects how tall the nucleotide is which can affect the global helical conformation.

So these two geometric properties, in addition to sequential and environmental factors allow nucleic acids to adopt a wide array of different conformations which are functionally relevant.



Nucleic acids are incredibly important, but we can see their importance by looking at how many nucleic acids have been deposited in the PDB.

As of 2025, around 5,000 NAs have been deposited into the PDB, and around 15,000 protein-nucleic acid complexes have been deposited.



If we compare that to proteins, we have to change the scale of our y axis, with an order of magnitude more more proteins than nucleic acids in the PDB. We know nucleic acids are interesting, they are very important, so why aren't there many structures? Likely this is due to difficulties that often occur during structure solution.

Since nucleic acids are so flexible and they have a polyanionic backbone so obtaining a well ordered and packed crystal can be difficult. And this flexibility and heterogeneity also hinders cryo-EM.

Nucleic Acid Crystallography

Structure Solution

How can we estimate the phase of our reflections?

The phase

h	k	l	F	phi
1	1	1	20	

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|e^{2\pi i\phi_{hkl}}e^{-2\pi i(hx+ky+lz)}$$

So imagining we have been able to obtain a crystal of our nucleic acid or protein - nucleic acid complex. We have to solve the phase problem which is inherent to crystallographic structure solution and has been spoken about.

The phase is a number which is associated with every reflection, we have a list of reflections after data processing but we are missing a column. To be able to get an electron density map (we need it so we know where our structure should go), we have to do this sum which contains the phase and the amplitude. If we don't have that number we can't do this sum. It's like attempting to do an addition sum where we are adding 1 + blank, it doesn't make sense, we need a number in that second position, like we need the phase which is just a number (angle).

Nucleic Acid Crystallography

Structure Solution

How can we estimate the phase of our reflections?

The phase

h	k	l	F	phi
1	1	1	20	90

$$\rho(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l |F_{hkl}| e^{2\pi i \phi_{hkl}} e^{-2\pi i(hx+ky+lz)}$$

Once we have a phase in this column we can calculate an electron density map, we could put any random number in here but it should be at least somewhat related to our structure, otherwise the map we get from the calculation is likely to be nonsense.

Nucleic Acid Crystallography

Structure Solution

How can we estimate the phase of our reflections?

Use prior knowledge

The first way we can phase our reflections (associate an angle with every reflection) is to use prior knowledge.

Nucleic Acid Crystallography

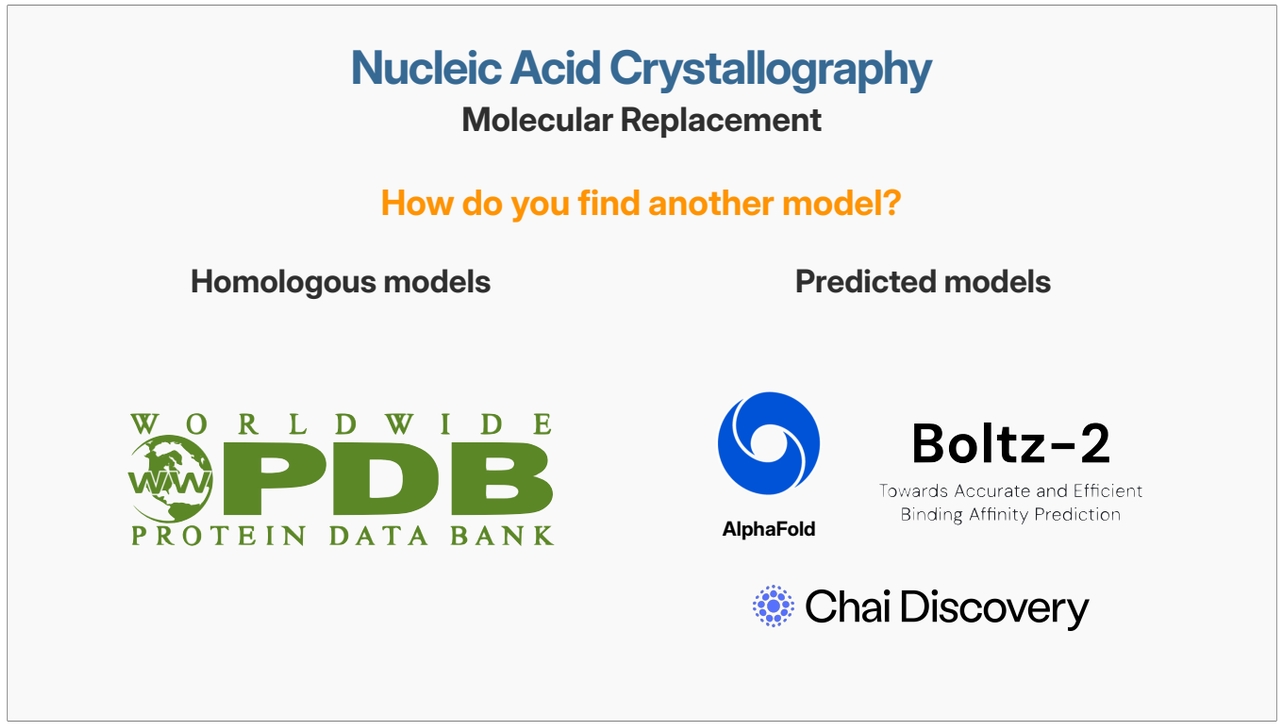
Structure Solution

How can we estimate the phase of our reflections?

Use prior knowledge

Estimate the phase of each reflection by using the information from another model

If we have information from a model which already exists, maybe we could get some information from that model and use it to get that column of phases. This is the process of molecular replacement. But we can’t just take that column and copy and paste it over (unless you’re sure), because the orientation and position of the macromolecule in the crystal is unlikely to be exactly the same. So we have to attempt to locate this position for use in molecular replacement.



So how can we get this information? Well we most commonly get it from another model which has already been solved (a homologous model) or one that we can predict in silico, commonly for nucleic acids we can use AlphaFold 3, Boltz-1/2, Chai-1, RosettaFoldNA, or VFold.

Maybe for protein molecular replacement, these can be seen as two distinct sources since predictions can work when homology cannot. But for nucleic acids these are more intimately linked, when you have a good homologous model then likely you will get a good prediction, and when you don’t then things get more tricky.

Nucleic Acid Crystallography

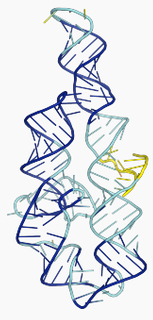
Molecular Replacement

How do you find another model?

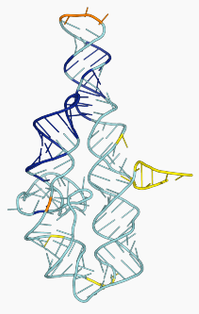
Homologous models

Homologue
PDB Code: 2R8S

Predicted models



AlphaFold 3



Boltz 2

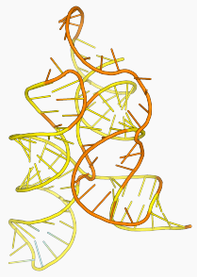
Chai 1

If we look at this example of a group 1 intron, we have a few homologous structures in the PDB, so prediction models (which are trained on the PDB) have a good idea of what is going on and can give us a sensible solution. If we attempted to run a molecular replacement software like *Phaser* using this as a search model, then it may work.

Nucleic Acid Crystallography

Molecular Replacement

How do you find another model?

Homologous models	Predicted models		
No homologues			
	AlphaFold 3	Boltz 2	Chai 1

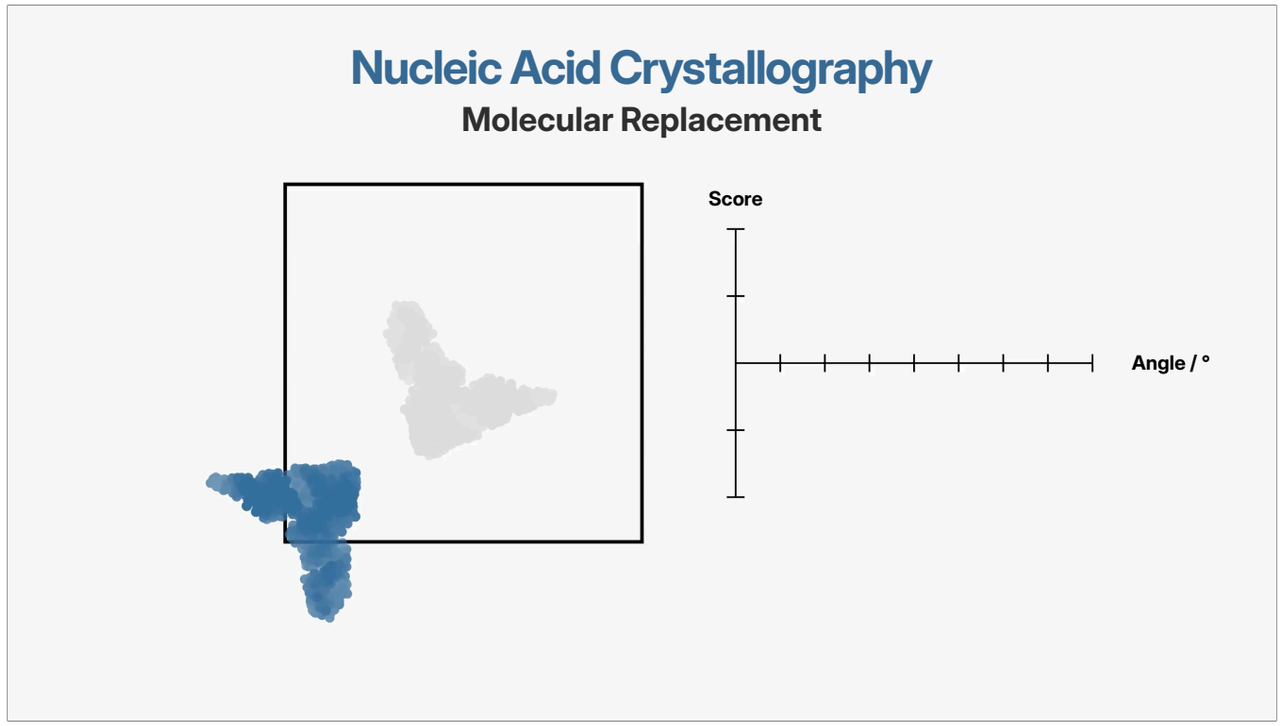
If we can’t find a homologue, chances are we may not find a good prediction, this is an example of a Rous sarcoma virus deposited in the PDB recently with a novel fold (nothing with >30% sequence identity).

But why are nucleic acid structure predictions so closely tied to available templates?

RNA structure prediction relies on the data available in the PDB, intricate networks of nucleic acid base pairing and other intramolecular interactions influence the fold, and when they are not well sampled in the training data (the prediction software packages haven’t seen them before) then they are unlikely to know what to do.

If this is a situation that you are in, it’s not the worst idea to at least attempt molecular replacement with one of the predictions, in case the predicted model got it correct somehow.

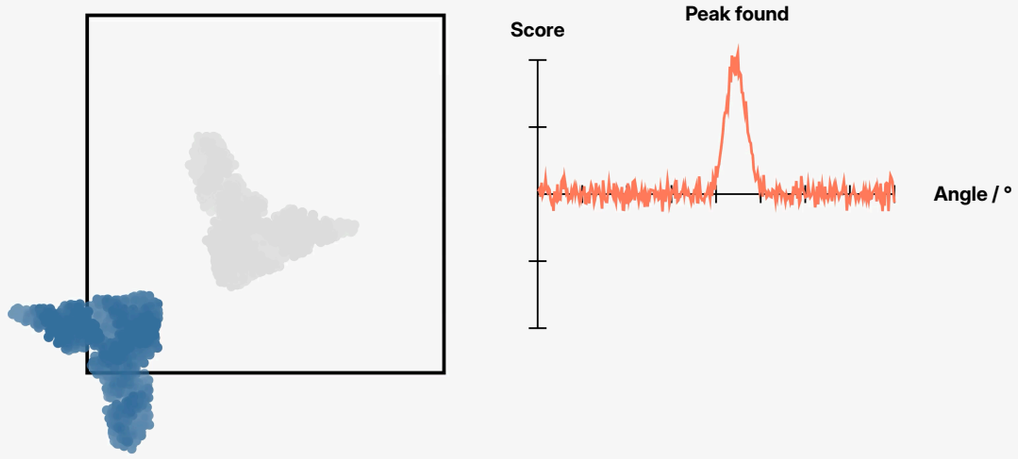
But remember nucleic acids are conformationally flexible so even though we have something which is homologous, it may not be isomorphous to our crystalline state and so molecular replacement may or may not be able to handle it.

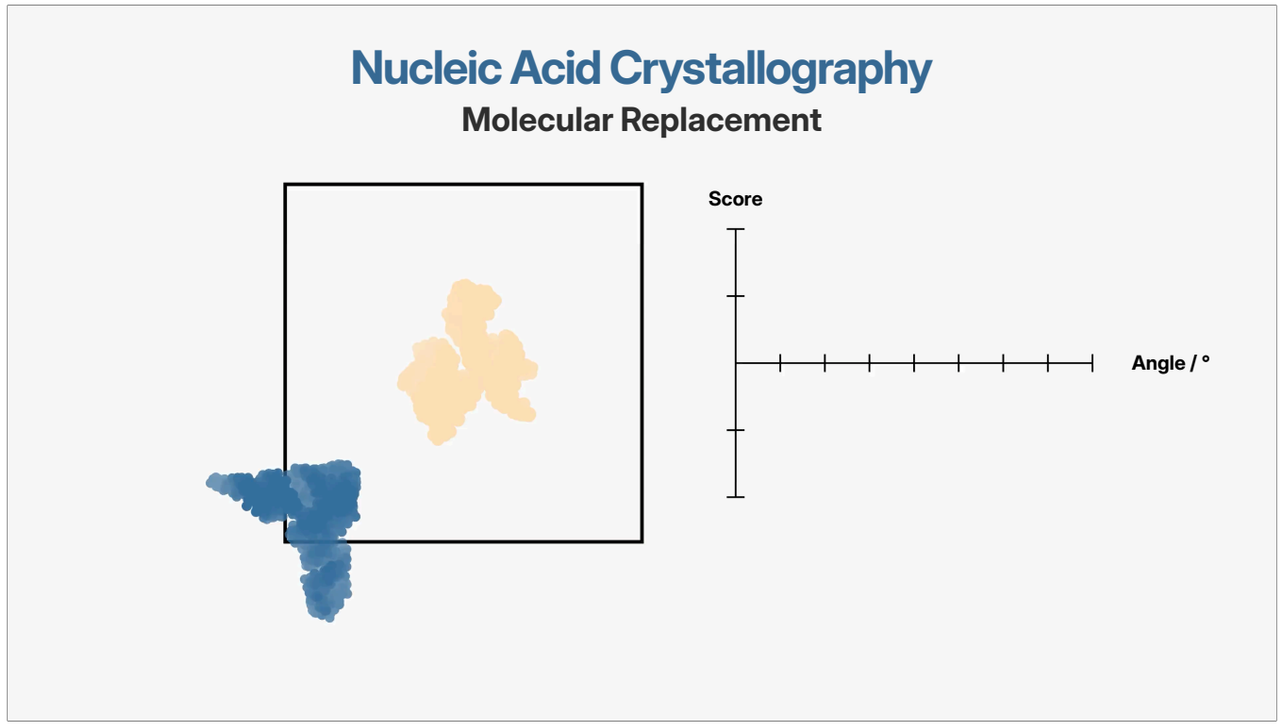


We can attempt to see this if we look at what is going on inside molecular replacement, the first step is a rotation function which attempt to see what rotation the molecule is in. The score here effectively tells us how well oriented our molecules are (between the search model and the actual structure in our crystal), and is in relatively either a Patterson rotation function or a likelihood rotation function.
As we rotate, when our molecules line up we see a peak in our score function and we know we have likely found a good rotation.

Nucleic Acid Crystallography

Molecular Replacement

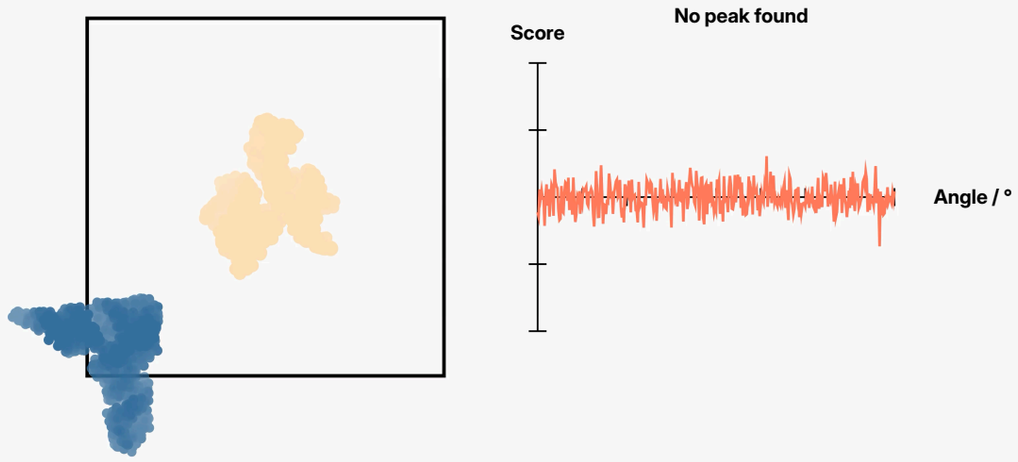




However, if our crystal and search model aren’t so similar (because they aren’t sequentially similar, or crystal contacts make it form a different conformation, or it is in a different conformation for some other reason, or the environment is different) as we rotate, the molecules never really align, and we don’t see a good solution to the rotation function.

Nucleic Acid Crystallography

Molecular Replacement



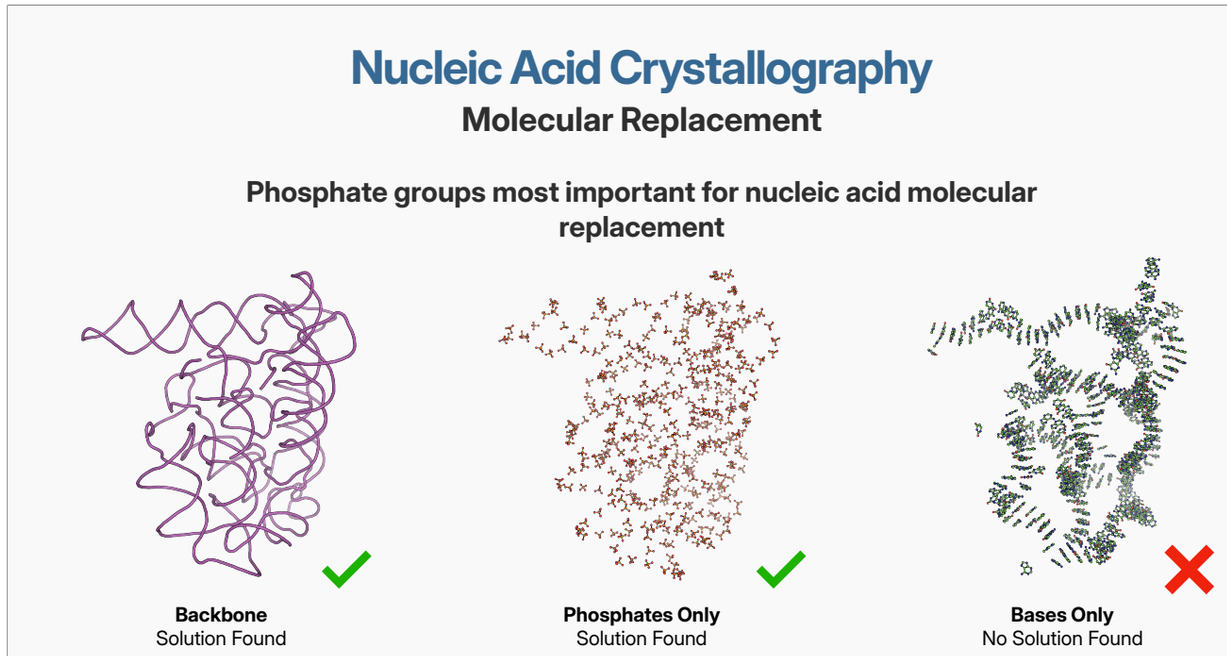
Nucleic Acid Crystallography

Molecular Replacement

Since nucleic acids are so conformationally diverse, structural similarity is more important than sequence similarity

Marcia M, et al. Acta Crystallogr D Biol Crystallogr. 2013;69(Pt 11):2174-2185. doi:10.1107/S0907444913013218

This underscores a point about molecular replacement in general, but particularly for nucleic acids, the structure of the model is more important than how similar the sequence is.



In fact for nucleic acids, the backbone and particular the phosphate positions are the most important. The bases are not enough to be able to get a molecular replacement solution.
Great paper on nucleic acid molecular replacement.

Nucleic Acid Crystallography

Molecular Replacement

How do you find another model?

Homologous models

Homologue
PDB Code: 2R8S

Software methods

Infernal

R3D-BLAST2

LocARNA

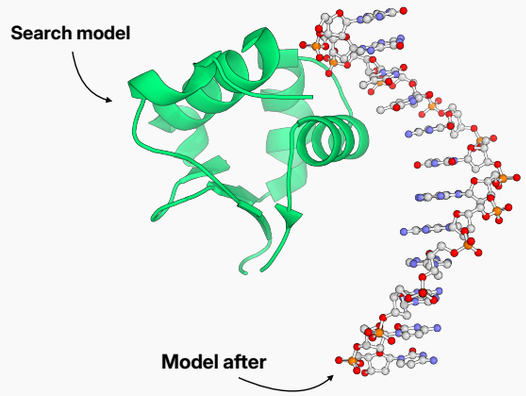
CMfinder

So since structure is important, there are some tools which attempt to help locate homologues which are structurally similar, but their success is limited by what is known in the PDB of course.
If prediction software and these software packages, are struggling to a search model, then maybe molecular replacement of your nucleic acid isn't going to work. There have been reports of more manual molecular replacement with smaller A-form RNA helices but these solutions may not fit every case.

Nucleic Acid Crystallography

Molecular Replacement

May be easier to use a protein search model if available



Heat-shock transcription factor 1 with DNA
PDB Code: 5D5W

If you have a protein in complex with your nucleic acid, it is almost always better and easier to use that protein as a search model, likely you know what the binding protein is, or if you don’t the predictions for proteins are much better than those of nucleic acids.
If you are having trouble crystallising a nucleic acid, introducing a chaperone or other protein binding target can also be a strategy that helps.

Nucleic Acid Crystallography

Structure Solution

How can we estimate the phase of our reflections?

Use prior knowledge

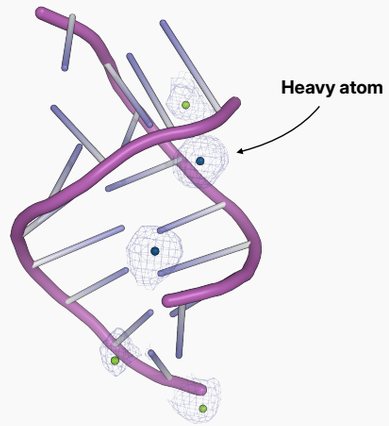
Use physical properties

Estimate the phase of each reflection by using information from heavy atoms

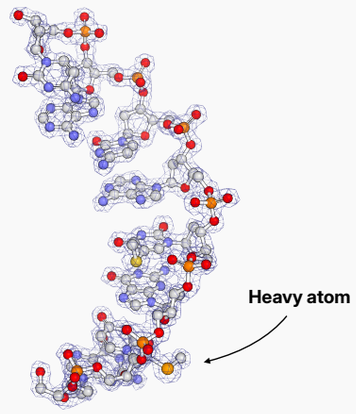
The other way we can get a phase angle associated with every reflection is to use physical properties of the diffraction experiment (which is helped by the presence of heavy atom scatterers).

Nucleic Acid Crystallography

Experimental Phasing



Ir soaked RNA
PDB Code: 8XZE
Partial Model



5-SMe Derivatised DNA
PDB Code: 3HG8

To solve a nucleic acid by experimental phasing, it is possible to soak your structure in a solution of heavy atom cations. This approach can work well with specific nucleic acids since the backbone is polyanionic. Alternatively, it may be possible to engineer a specific binding site if required.
Alternatively, you could incorporate heavy atom deriviatised nucleotides into your structure (purchasable) if you are working with small nucleic acids.

You may not even need to do soaking or derivatisation, since nucleic acids themselves contain a relatively heavy atom (P). If you have an appropriate number of nucleotides (<30) you may even be able to do native-SAD (where you don’t add any heavy atoms) at a beamline like i23 at Diamond, but it is tricky to do.

Nucleic Acid Crystallography

Structure Solution

How can we estimate the phase of our reflections?

Use prior knowledge	Use physical properties
Molecular Replacement	**Experimental Phasing**

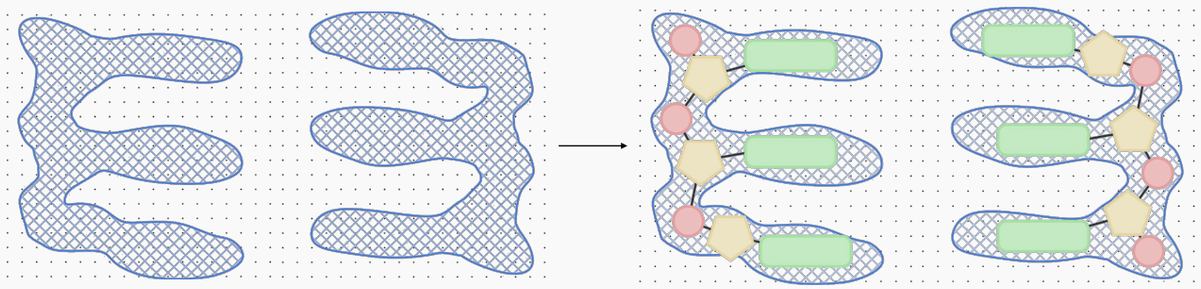
$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|e^{2\pi i\phi_{hkl}}e^{-2\pi i(hx+ky+lz)}$$

So we can either use molecular replacement (using prior knowledge) or use experimental phasing (using physical properties) to get an estimate for every phase angle in our list of reflections.

Now we have a phase for each reflection, we can complete this summation and generate an electron density map (that we can open in up a software package like Coot and look at it).

Nucleic Acid Structure Solution

Model Building

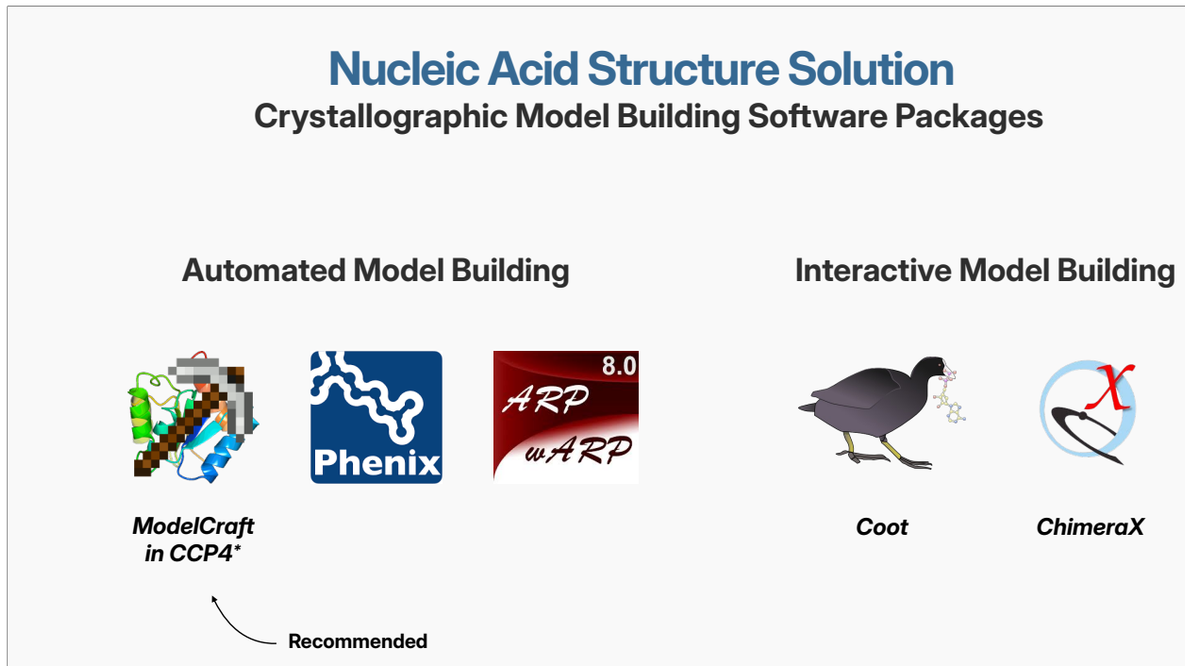


Density
2mFo-DFc or charge density

Atomic Model

After phasing, we have an electron density map in the case of crystallography and a charge density map in the case of cryo-EM, we need to create a model which explains our experimental observations (the electron density).

Depending on how well our phasing attempts have gone, the electron density may look nice an interpretable, or it may look poor and difficult to interpret. This is commonly the case in the regions of nucleic acids after phasing with protein molecular replacement for example.



In crystallography, there are a number of software methods that can help us with this, there is ModelCraft in CCP4, Phenix, Arp/wARP (also in CCP4) which attempt to do this process automatically. They are often decent for proteins but for nucleic acids they can sometimes be challenging. The best software package currently is ModelCraft, since it contains the new automated nucleic acid model building package NucleoFind (my PhD work) which builds models of nucleic acids with more completeness and accuracy than existing crystallographic methods.

If there is some ambiguity, you may have to look at interactive model building for example in a software package like Coot or ChimeraX.

Nucleic Acid Structure Solution

cryo-EM Model Building Software Packages

Automated Model Building

ModelCraft in CCP4*

Recommended for nucleic acids

Phenix

ModelAngelo

Recommended for protein-nucleic acid complexes

cryoREAD

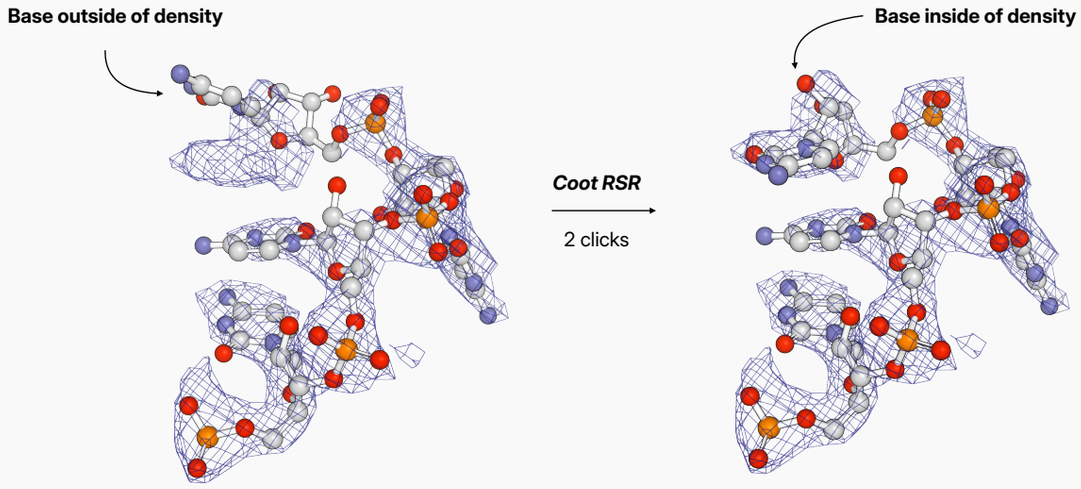
Recommended for low resolution cases

For cryo-EM the software packages are relatively more recently developed. There is ModelCraft in CCP4, Phenix, ModelAngelo, and cryoREAD which can all model nucleic acids.
ModelCraft, Phenix and ModelAngelo will build protein-nucleic acid complexes and cryoREAD is just for nucleic acids.

Best for protein-nucleic acid modelling is ModelAngelo
Best for lower resolution cryoREAD
Best for RNA only and good for P-NA ModelCraft

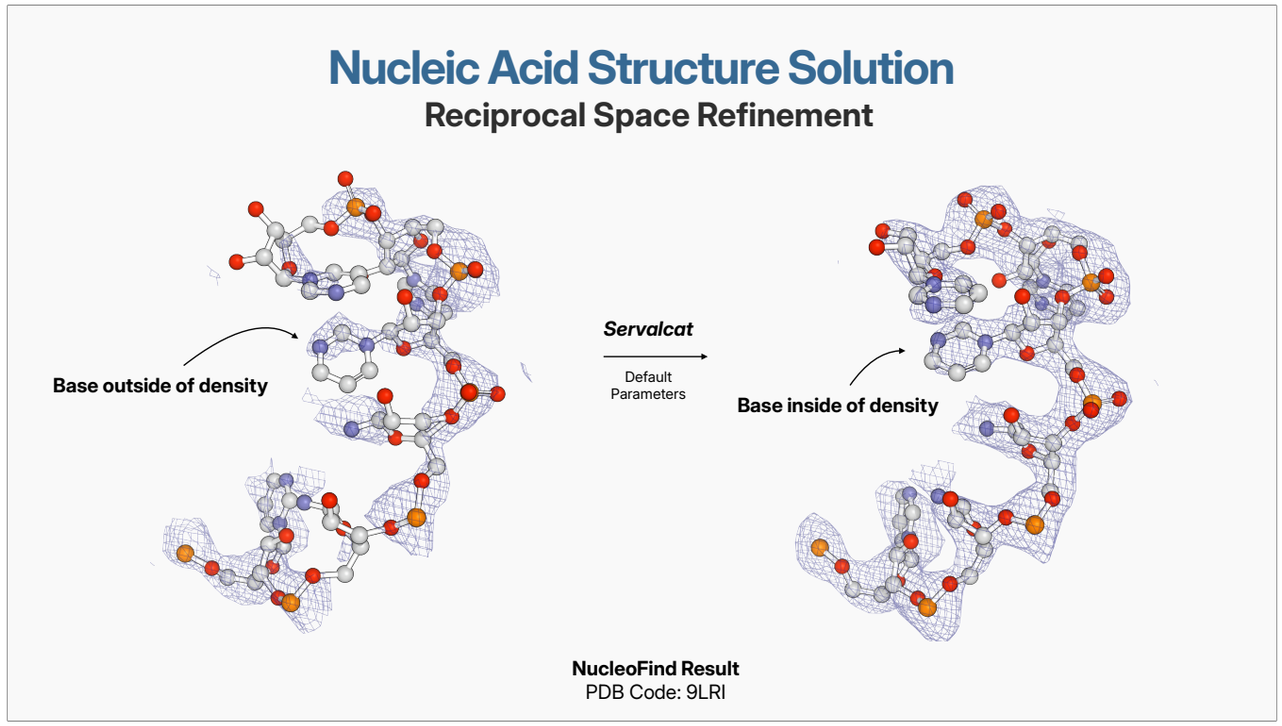
Nucleic Acid Structure Solution

Real Space Refinement



Synthetic Example
PDB Code: 9LRI

After model building, we may have certain residues which are outside of the density, we can use Coot real space refinement to deal with these, particularly good for terminal residues where model building programs may get confused.

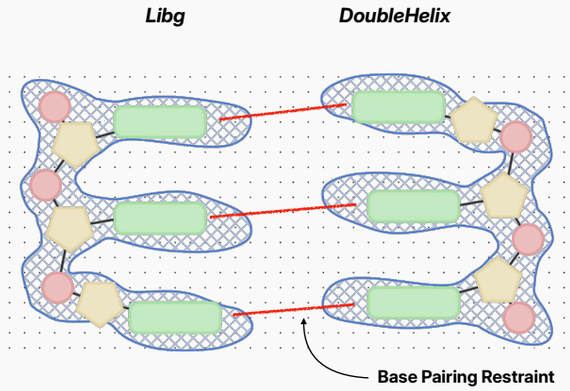


And then you will want to refine in reciprocal space in order to update your phase estimations and update the geometry of your model. Servalcat is a good program which can do this for both crystallography and cryo-EM.

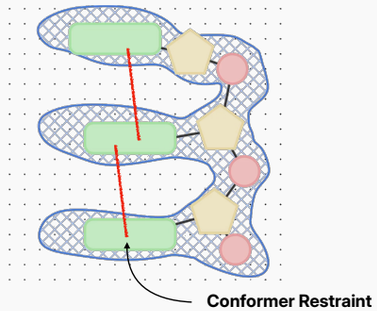
Nucleic Acid Structure Solution

Refinement Restraints

Base pairing and puckering restraints



Conformer restraints



With lower resolution data you may want to incorporate specific nucleic acid refinement restraints like those you could generate from libg or from another program called doubleHelix. These will ensure that our geometry resembles pre-existing expectations and can help us keep consistent bond angles/torsions etc.

Another refinement restraint you may want to use is conformer restraints. There is web server DNATCO (which is coming to CCP4 soon) which compares dinucleotides to high resolution structurally validated dinucleotides, and you are able to select a specific conformation (called an nucleotide confomers) and enforce those restraints during refinement.

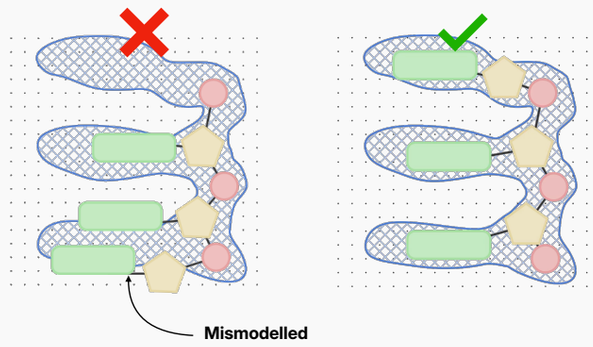
Nucleic Acid Structure Solution

Validation

MOL PROBITY

DNATCO

DoubleHelix



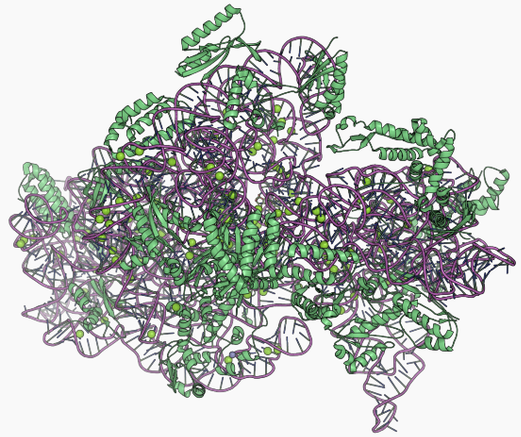
During the process of iterative model building and certainly after model building we want to validate our structures to ensure that the model we have made is sensible. We want to ensure it fits to the experimental observations and also we want to make sure that the geometry of our model is biochemically sensible.

There are many programs which do this like MolProbity, DNATCO and DoubleHelix. For nucleic acids, DNATCO is a great resource which can allow you to look at the known well modelled high resolution nucleotide conformers and see if your model has dinucleotides which fit a known type, if they don’t maybe there could be another explanation that fits the data better.

Nucleic Acid Structure Solution

Conclusions

Nucleic acid structure solution is more difficult than proteins but achievable



Email: jordan.dialpuri@york.ac.uk

30S Ribosomal Subunit
PDB Code: 1IBK

Once we have validated and finished our structure we can deposit it to the PDB.

In general nucleic acid structure solution is more difficult than that of proteins, but with better tooling and advances in methods it is becoming easier. The next update of CCP4 coming in a few months will bring big updates for nucleic acid crystallography and cryo-EM and we look forward to it.

You are welcome to email me if you have any questions about nucleic acid structure solution