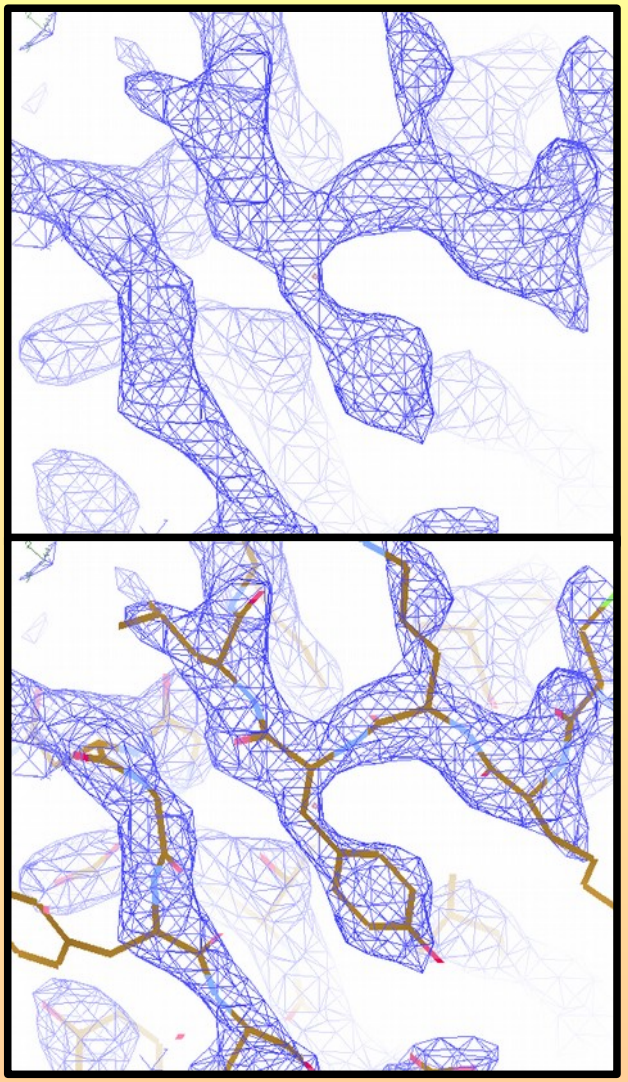
Model Building

Model building software:

- Proteins:
  - Buccaneer
  - ARP/wARP
  - Phenix autobuild
- Nucleic acids:
  - Nautilus/Coot
  - ARP/wARP
  - Phenix autobuild



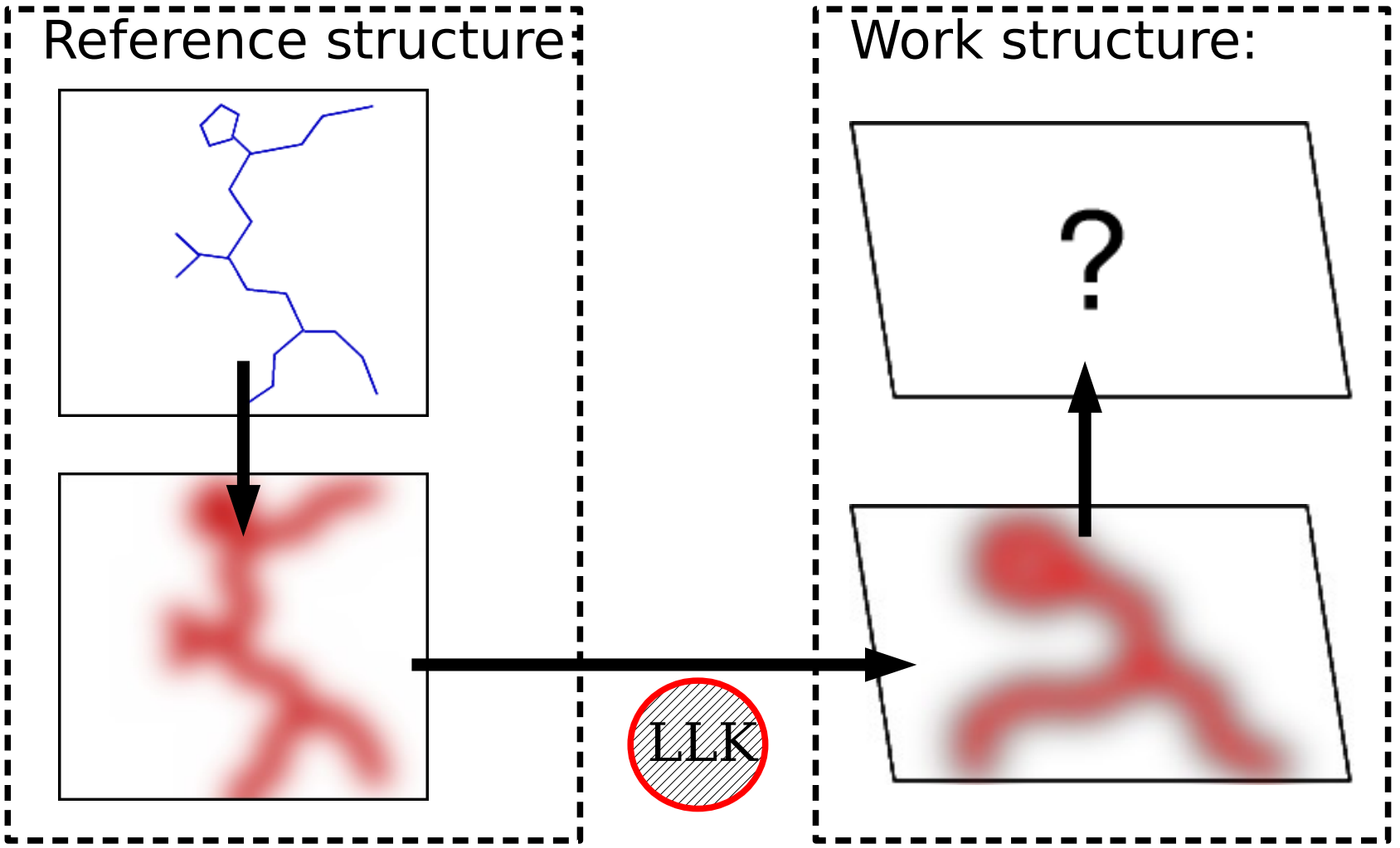
Buccaneer

The buccaneer software for automated model building of protein structures across a broad range of resolutions.

Kevin Cowtan
YSBL, University of York
cowtan@ysbl.york.ac.uk

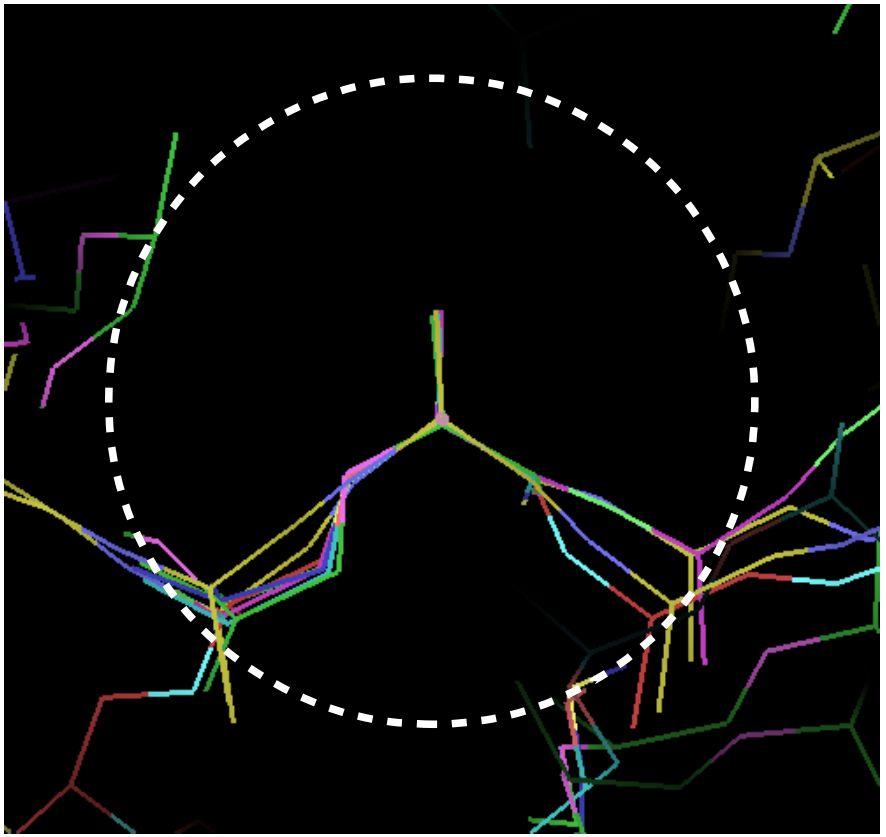
Buccaneer: Method

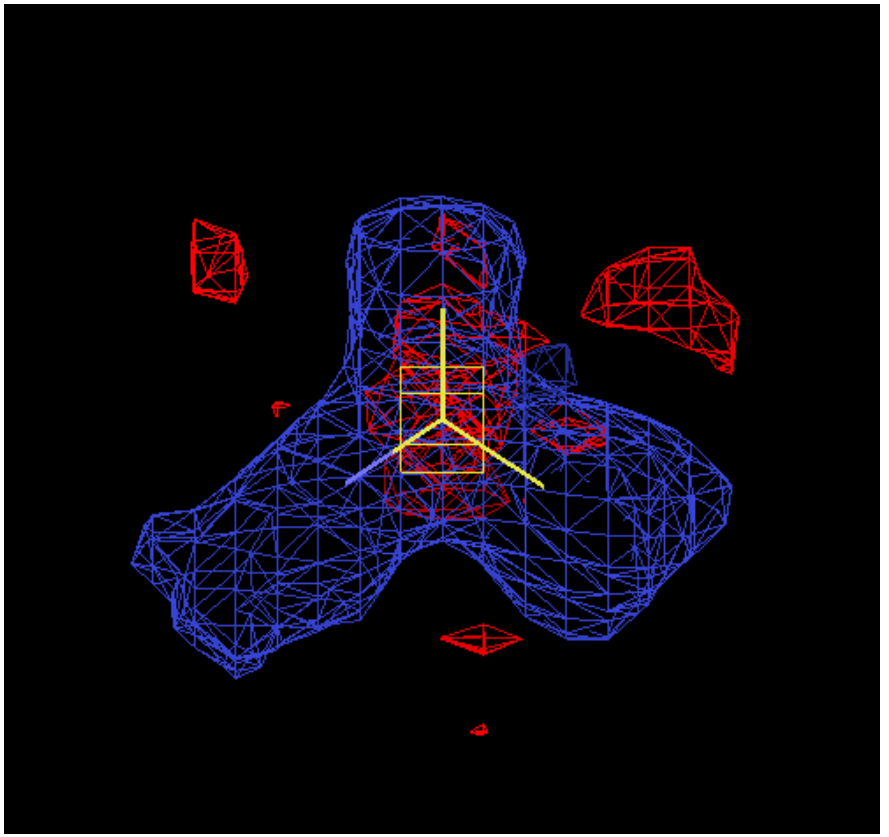
- Compare simulated map and known model to obtain likelihood target, then search for this target in the unknown map.



Buccaneer: Method

- Compile statistics for reference map in 4A sphere about Cα => LLK target.





- Use mean/variance.

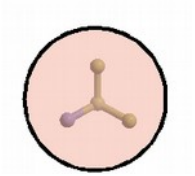
4A sphere about Ca also used by 'CAPRA' loeger et al. (but different target function).

Buccaneer

Use a likelihood function based on conserved density features.

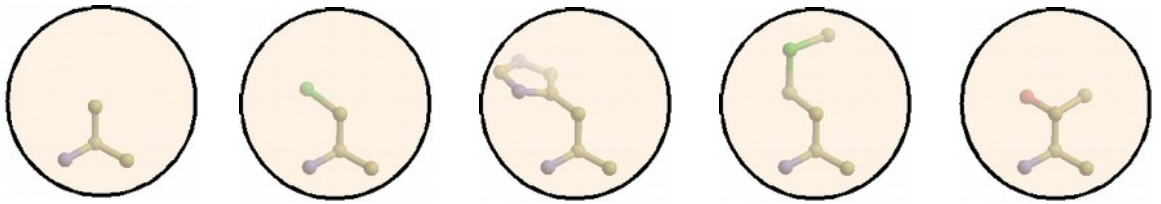
The same likelihood function is used several times. This makes the program very simple (<3000 lines), and the whole calculation works over a range of resolutions.

Finding, growing: Look for C-alpha environment



(4.0A sphere about Cα)

Sequencing: Look for C-beta environment



(5.5A sphere about Cβ)

ALA CYS HIS MET THR ... x20

Buccaneer

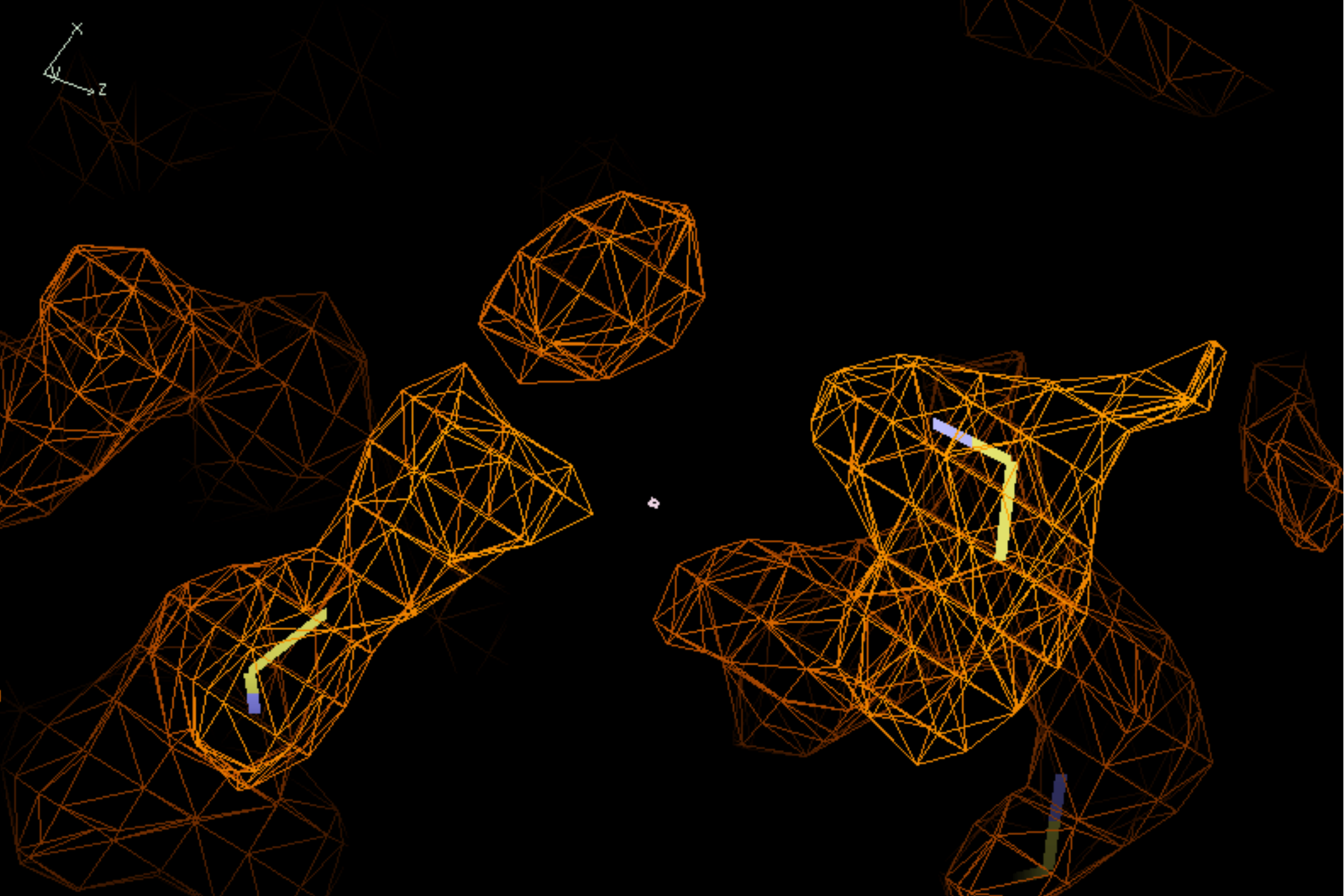
10 stages:

- **Find** candidate C-alpha positions
- **Grow** them into chain fragments
- **Join** and merge the fragments, resolving branches
- **Link** nearby N and C terminii (if possible)
- **Sequence** the chains (i.e. dock sequence)
- **Correct** insertions/deletions
- **Filter** based on poor density
- **NCS Rebuild** to complete NCS copies of chains
- **Prune** any remaining clashing chains
- **Rebuild** side chains

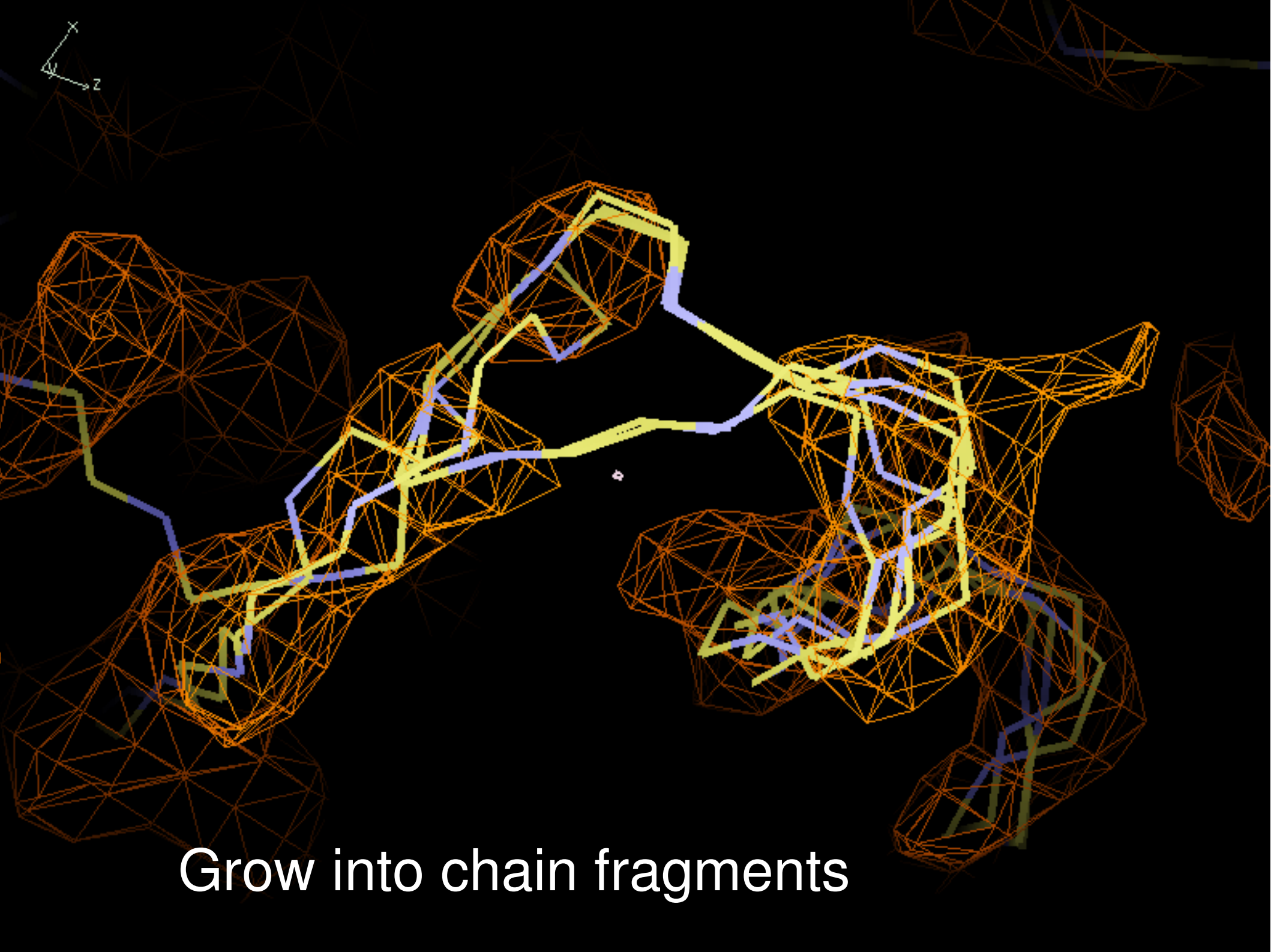
Buccaneer

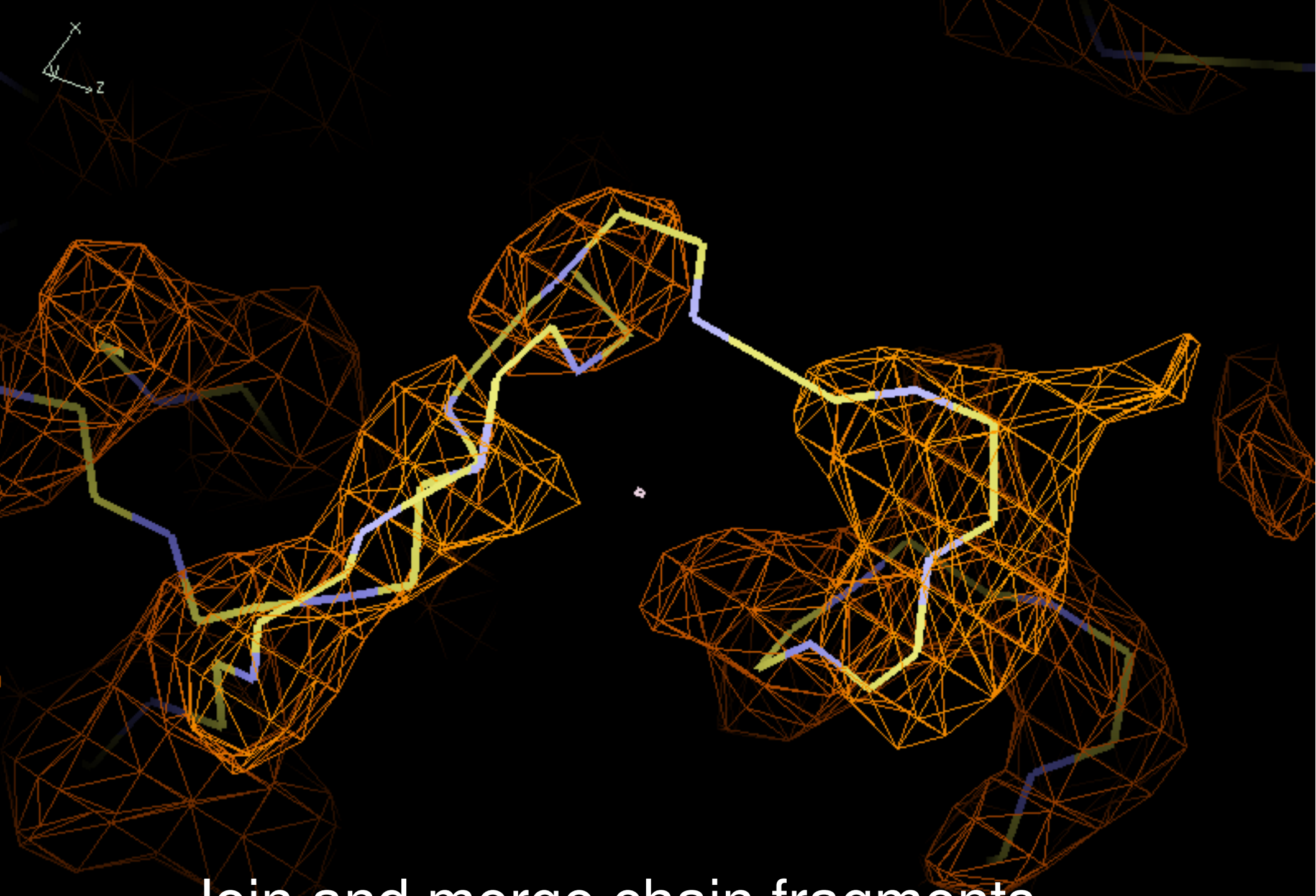
Case Study:

A difficult loop in a 2.9A map, calculated using real data from the JCSG.

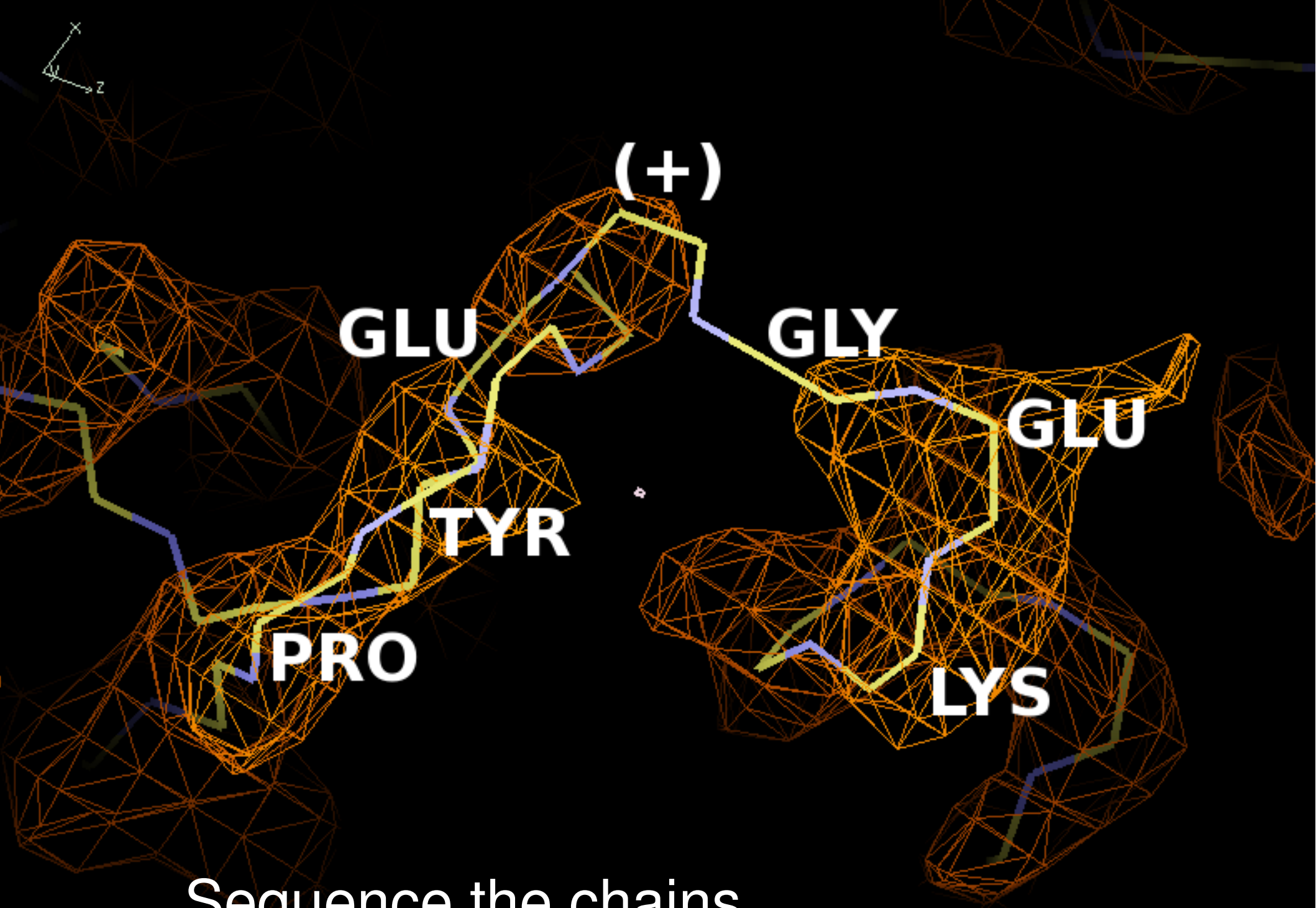


Find candidate C-alpha positions

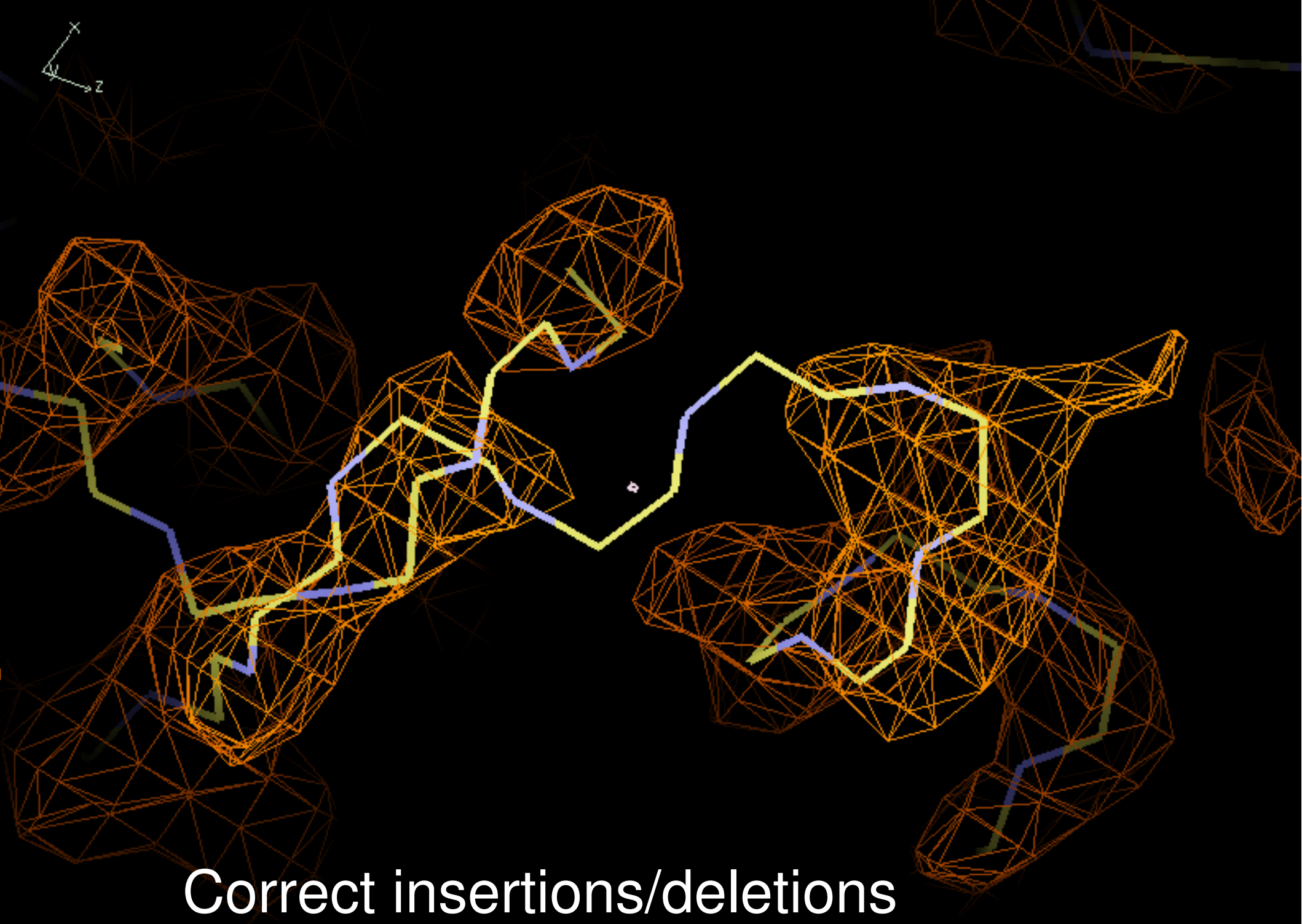

Grow into chain fragments



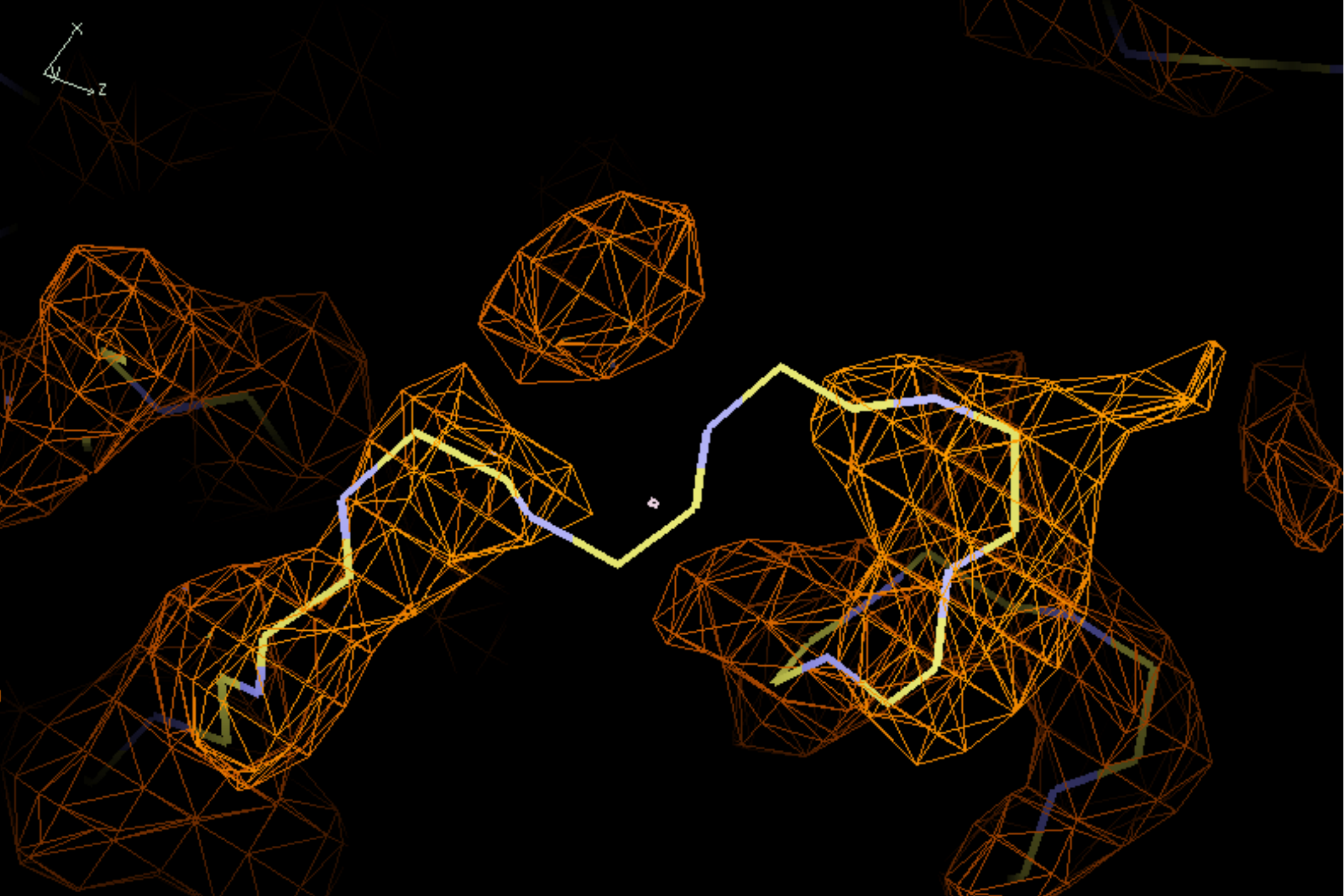
Join and merge chain fragments



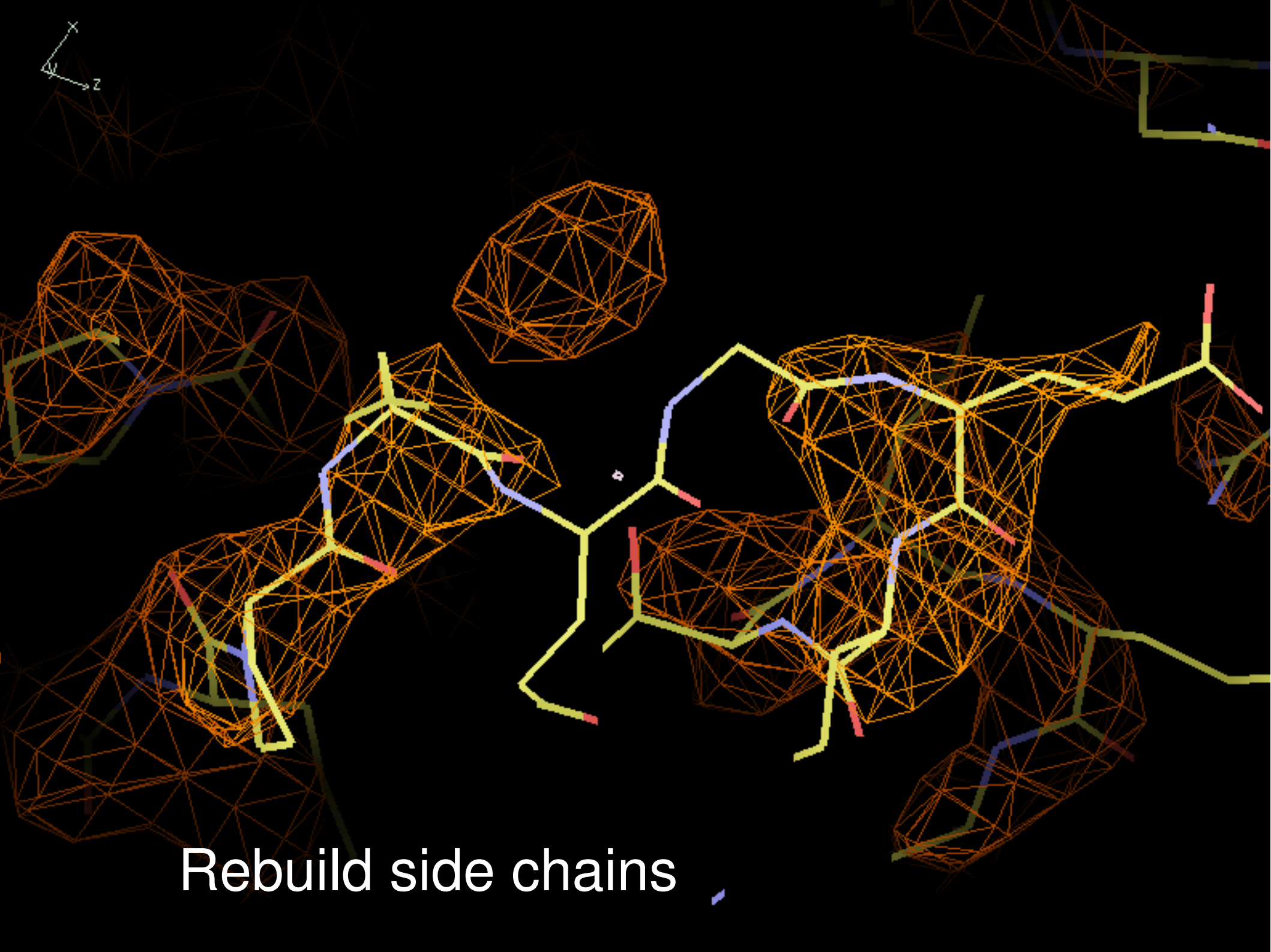
Sequence the chains

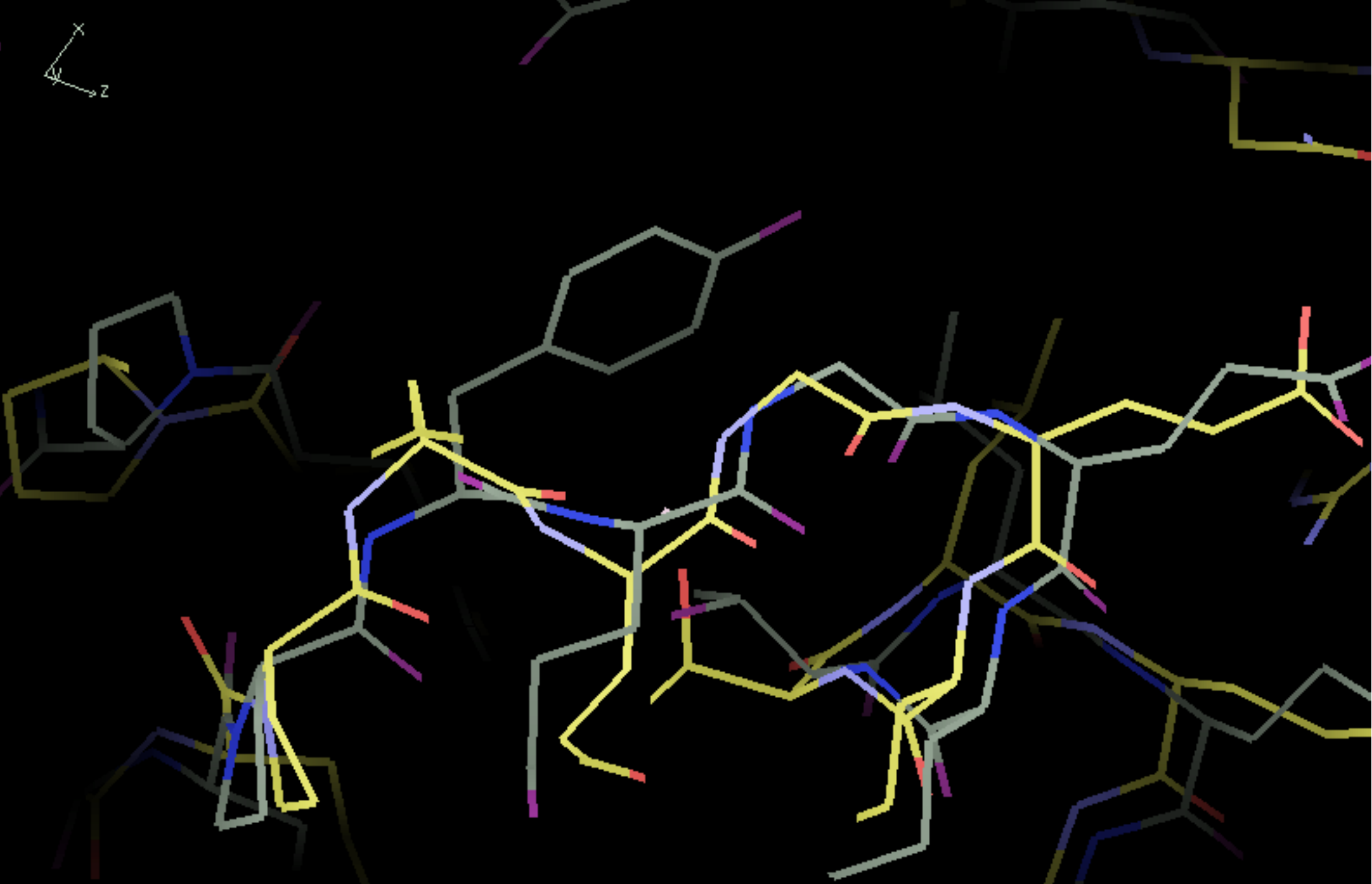


Correct insertions/deletions



Prune any remaining clashing chains


X
Z
Rebuild side chains

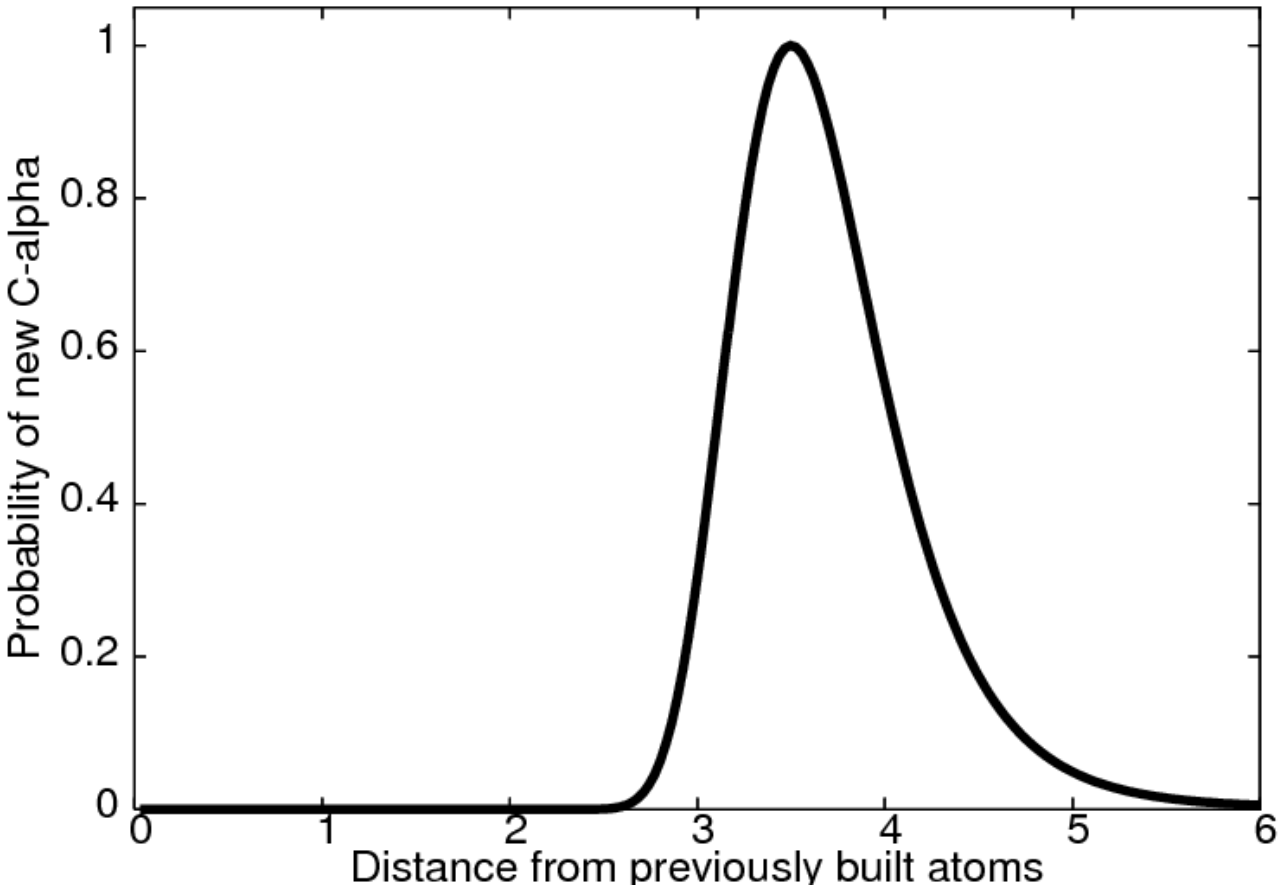


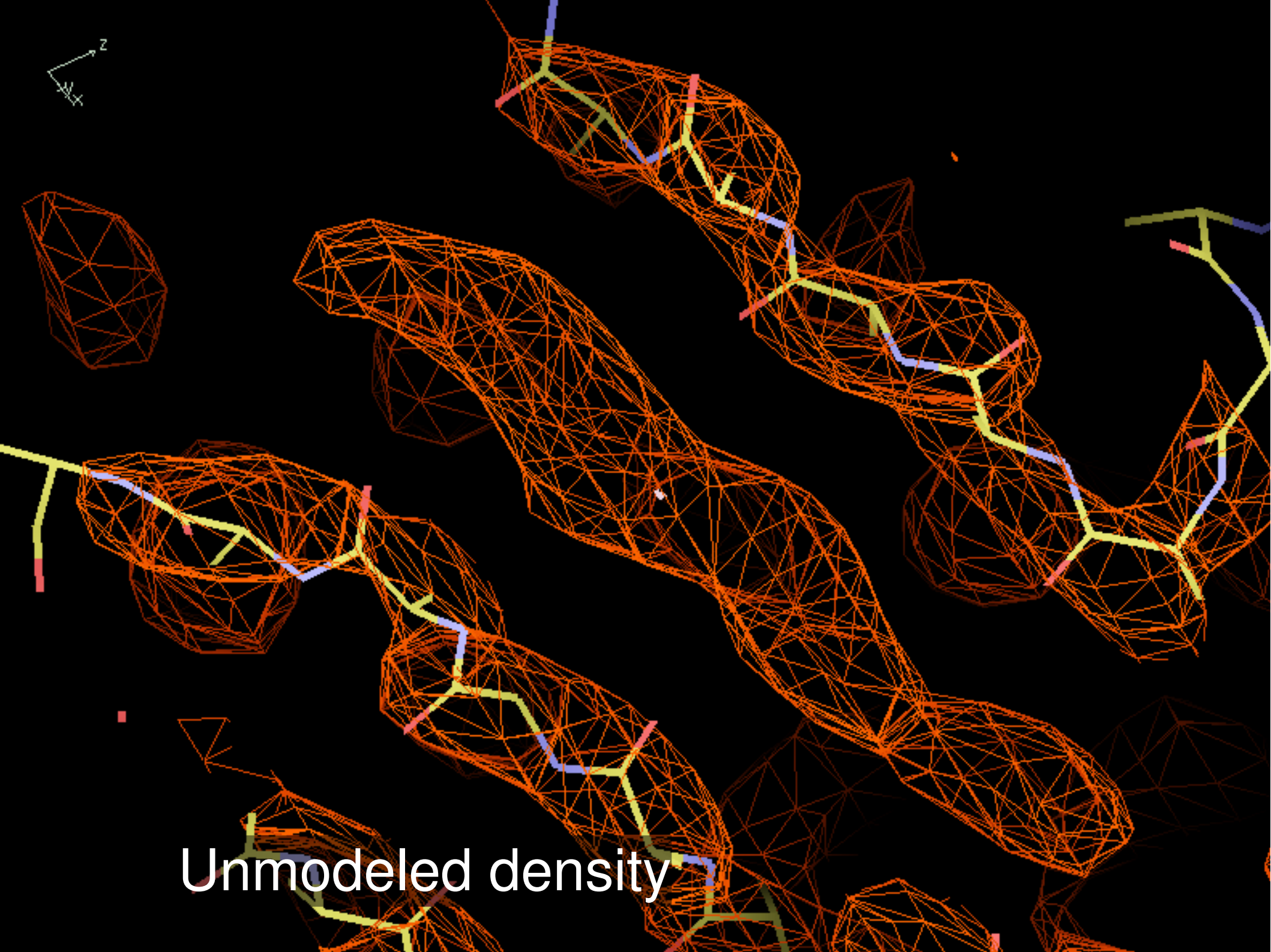
Comparison to the final model

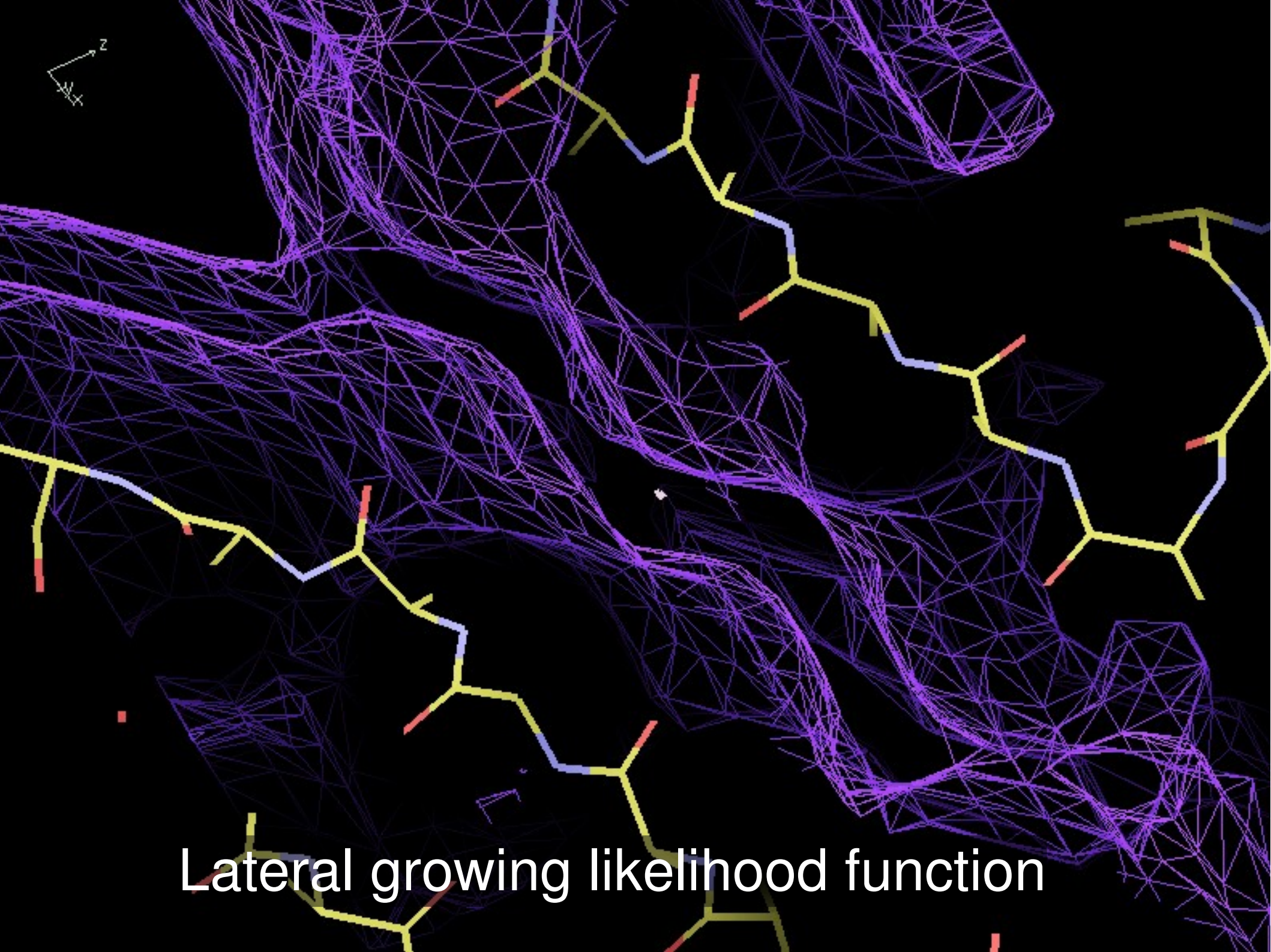
Buccaneer

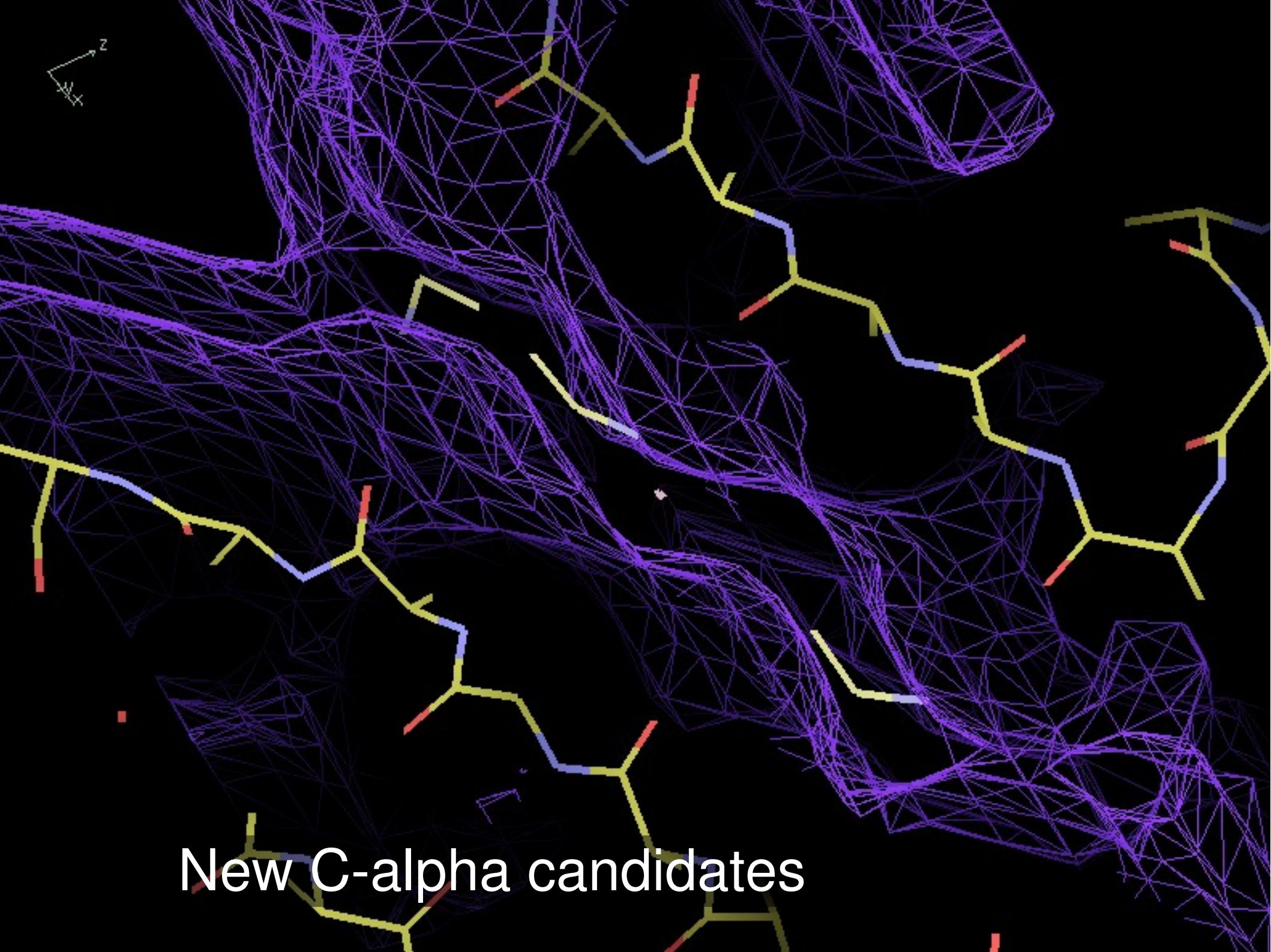
Model completion uses "**Lateral growing**":

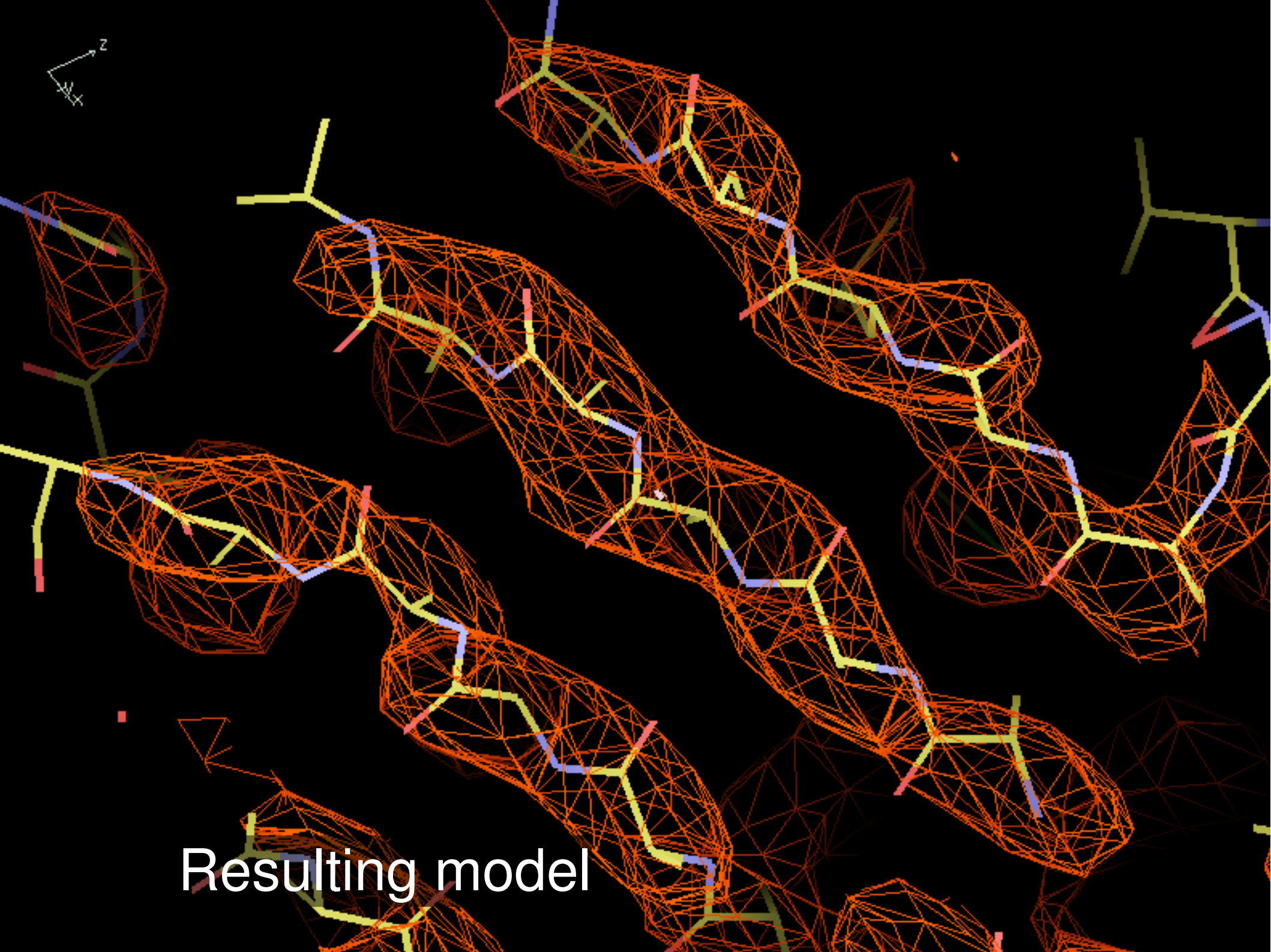
Grow sideways from existing chain fragments by looking for new C-alphas at an appropriate distance "sideways" from the existing chain:




Unmodeled density

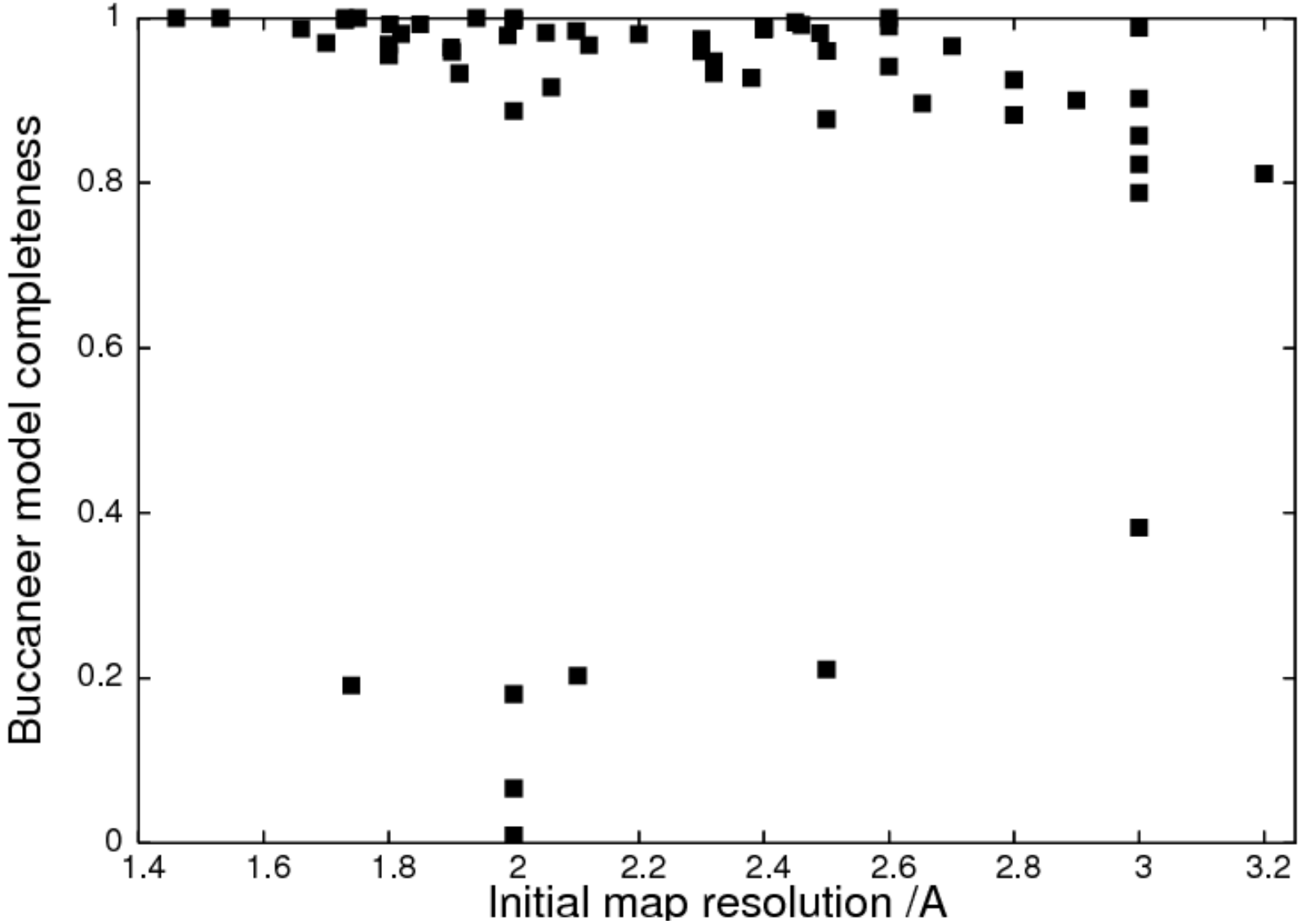

Lateral growing likelihood function


New C-alpha candidates


Z
X
Resulting model

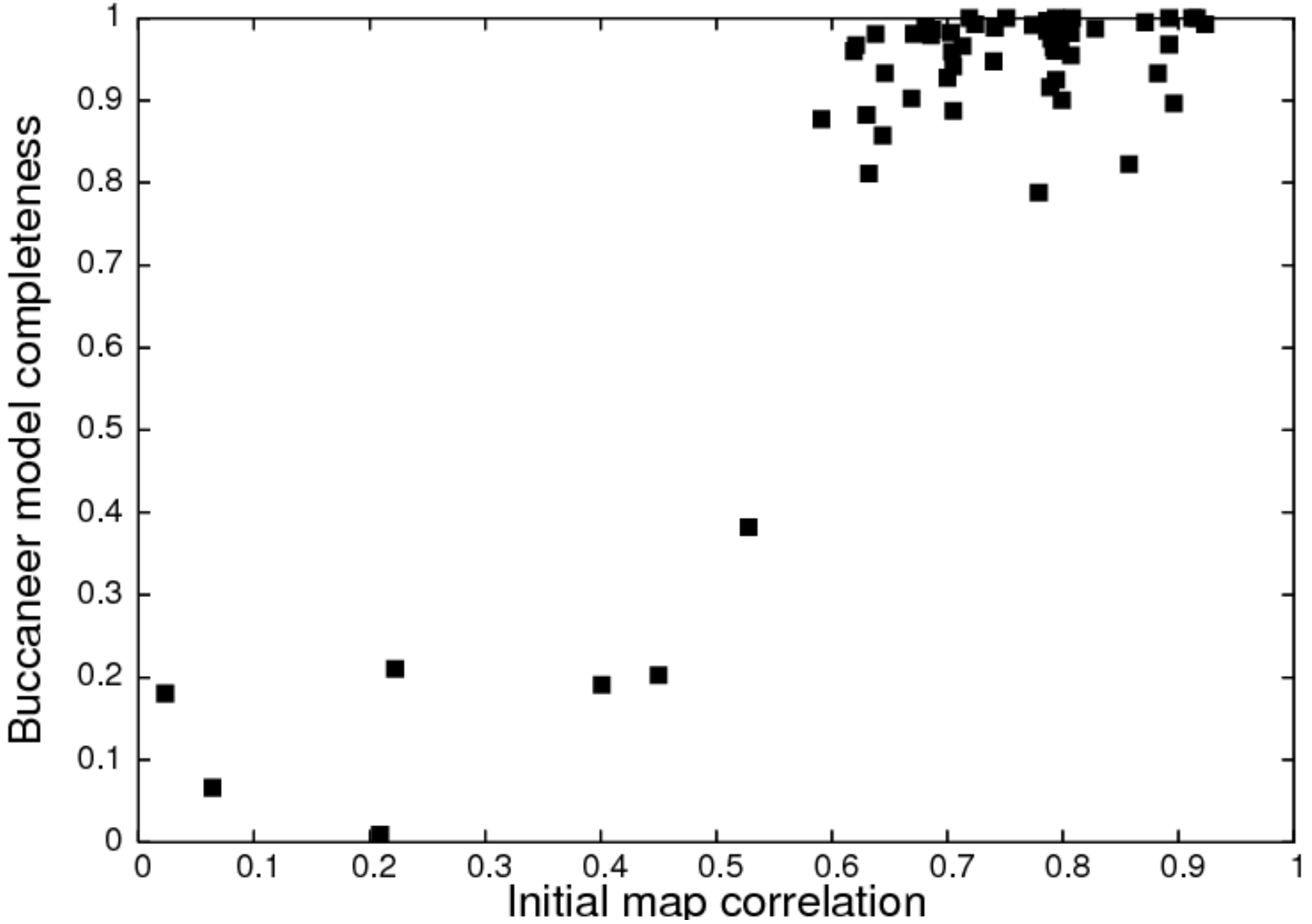
Buccaneer: Results

Model completeness not very dependent on resolution:

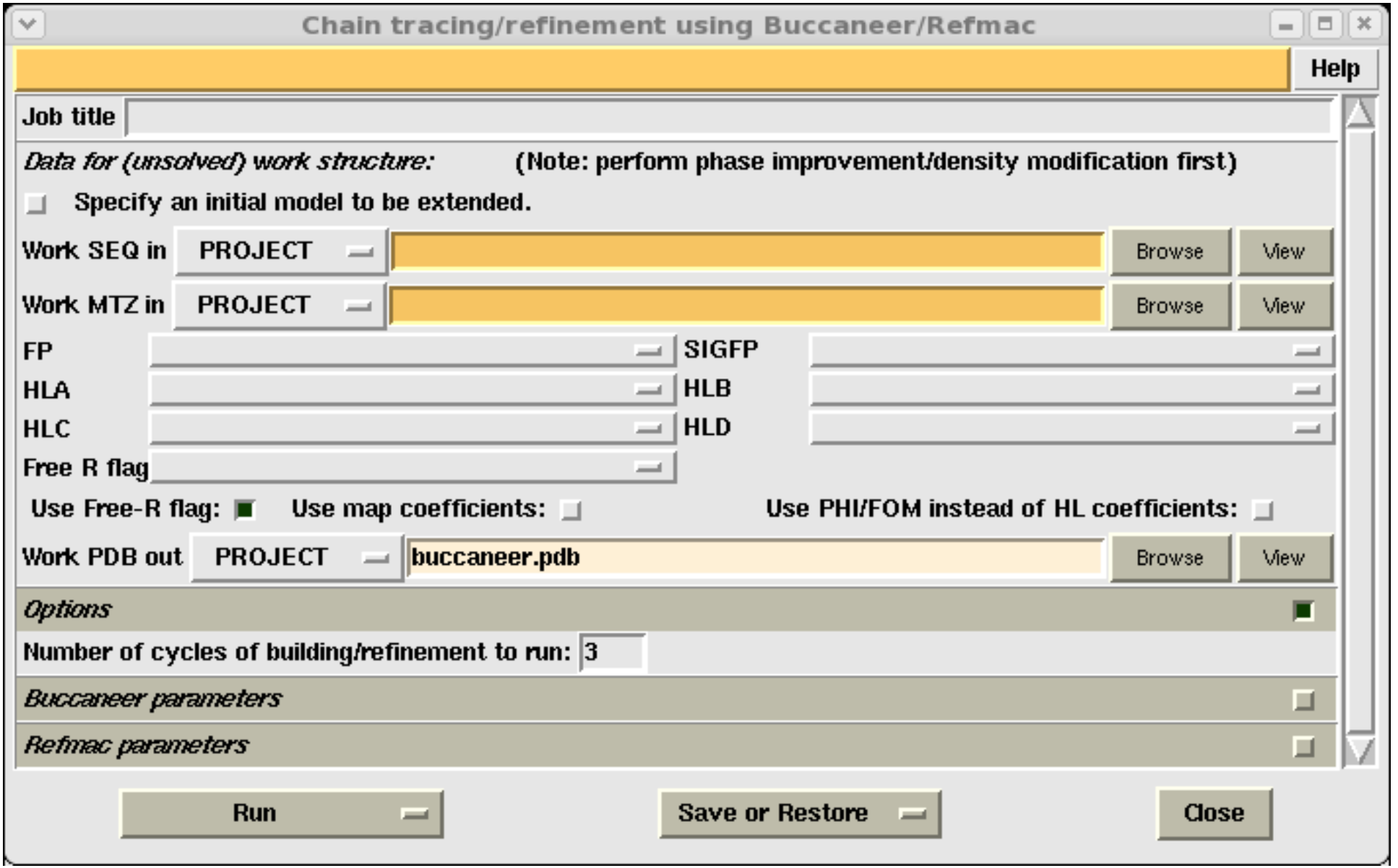


Buccaneer: Results

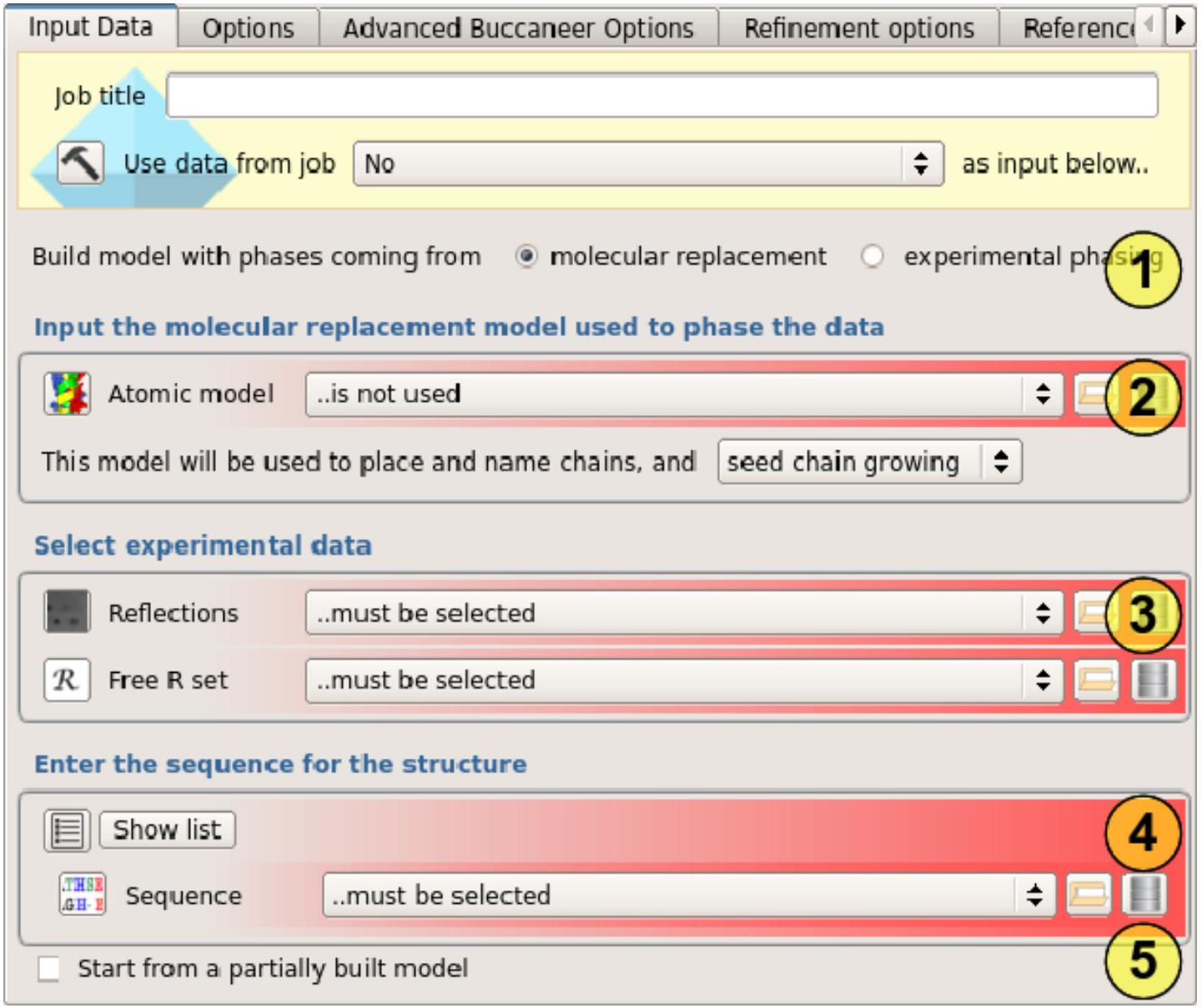
Model completeness dependent on initial phases:



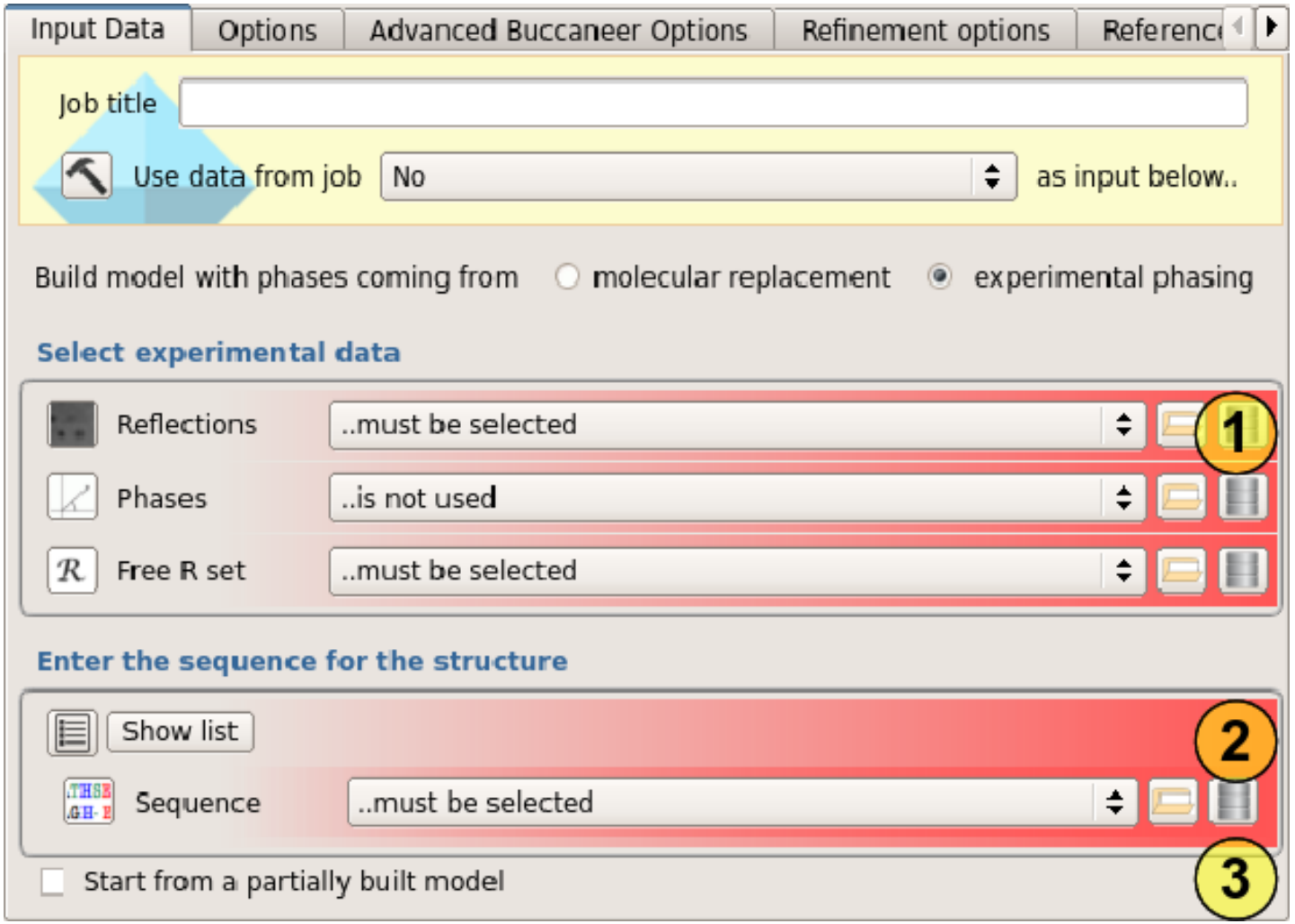
Buccaneer


Chain tracing/refinement using Buccaneer/Refmac
Help
Job title
Data for (unsolved) work structure:
(Note: perform phase improvement/density modification first)
Specify an initial model to be extended.
Work SEQ in
PROJECT
Browse
View
Work MTZ in
PROJECT
Browse
View
FP
SIGFP
HLA
HLB
HLC
HLD
Free R flag
Use Free-R flag:
Use map coefficients:
Use PHI/FOM instead of HL coefficients:
Work PDB out
PROJECT
buccaneer.pdb
Browse
View
Options
Number of cycles of building/refinement to run:
3
Buccaneer parameters
Refmac parameters
Run
Save or Restore
Close

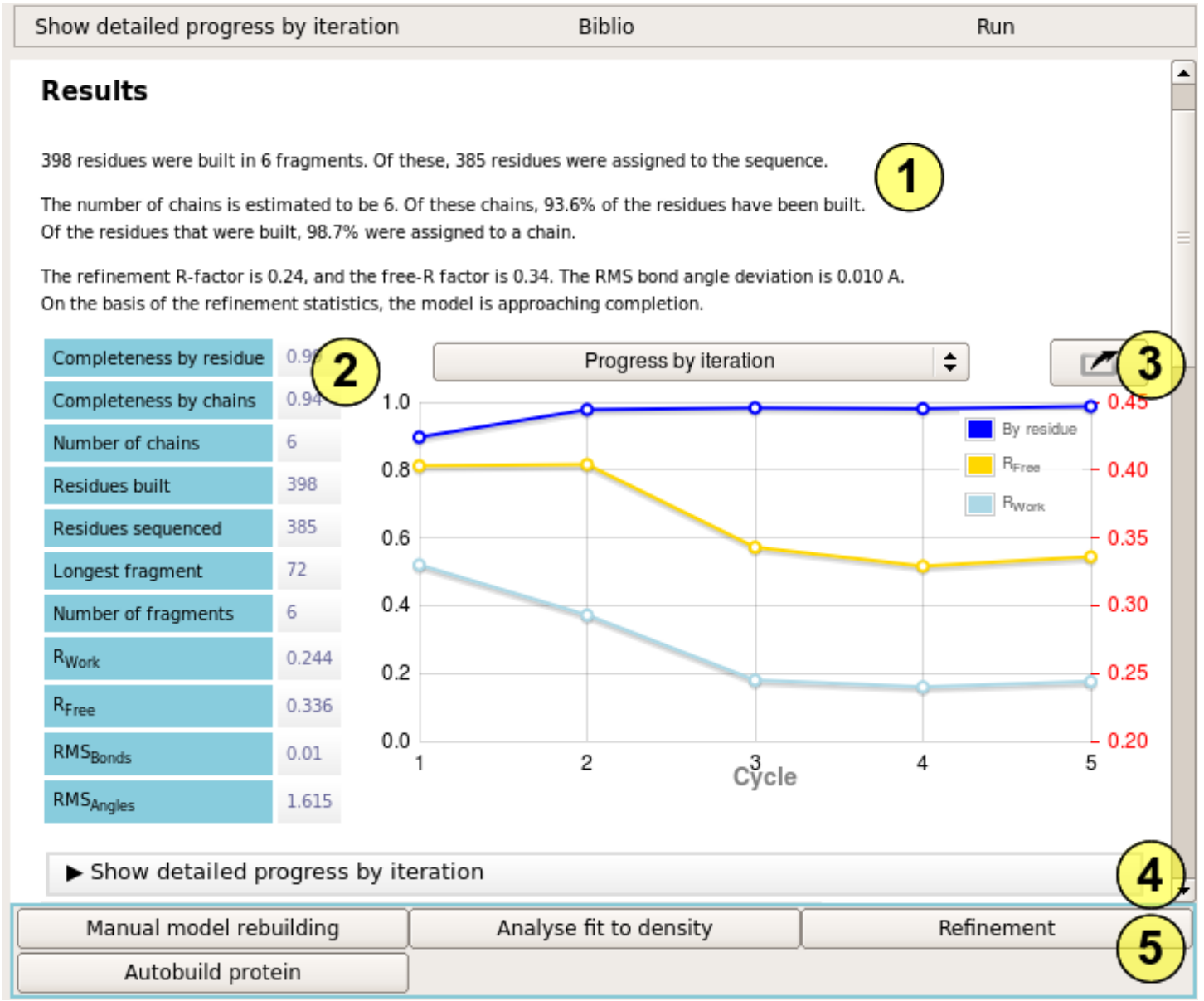
Buccaneer


Input Data
Options
Advanced Buccaneer Options
Refinement options
Reference
Job title
Use data from job
No
as input below..
Build model with phases coming from
molecular replacement
experimental phasing
1
Input the molecular replacement model used to phase the data
Atomic model
..is not used
2
This model will be used to place and name chains, and
seed chain growing
Select experimental data
Reflections
..must be selected
3
Free R set
..must be selected
Enter the sequence for the structure
Show list
4
Sequence
..must be selected
5
Start from a partially built model

Buccaneer


Input Data
Options
Advanced Buccaneer Options
Refinement options
Job title
Use data from job
No
as input below..
Build model with phases coming from
molecular replacement
experimental phasing
Select experimental data
Reflections
..must be selected
1
Phases
..is not used
Free R set
..must be selected
Enter the sequence for the structure
Show list
2
Sequence
..must be selected
Start from a partially built model
3

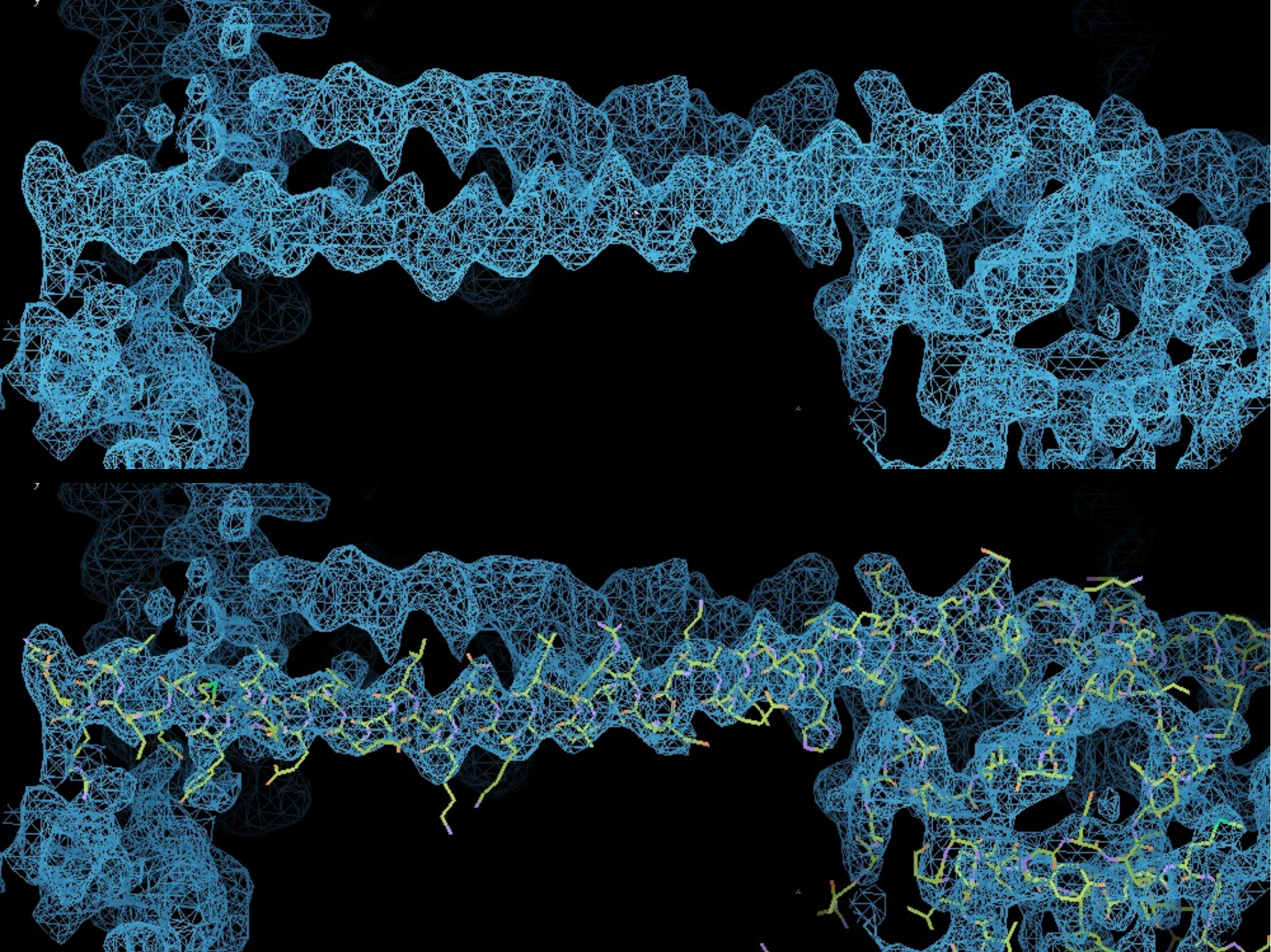
Buccaneer


Show detailed progress by iteration
Biblio
Run
Results
398 residues were built in 6 fragments. Of these, 385 residues were assigned to the sequence.
1
The number of chains is estimated to be 6. Of these chains, 93.6% of the residues have been built.
Of the residues that were built, 98.7% were assigned to a chain.
The refinement R-factor is 0.24, and the free-R factor is 0.34. The RMS bond angle deviation is 0.010 A.
On the basis of the refinement statistics, the model is approaching completion.
Completeness by residue
Completeness by chains 0.94
Number of chains 6
Residues built 398
Residues sequenced 385
Longest fragment 72
Number of fragments 6
RWork 0.244
RFree 0.336
RMSBonds 0.01
RMSAngles 1.615
2
Progress by iteration
3
By residue
RFree
RWork
1.0
0.8
0.6
0.4
0.2
0.0
0.45
0.40
0.35
0.30
0.25
0.20
1
2
3
4
5
Cycle
▶ Show detailed progress by iteration
4
Manual model rebuilding
Analyse fit to density
Refinement
Autobuild protein
5

Buccaneer

What you need to do afterwards:

- Tidy up with Coot.
  - Or ARP/wARP when resolution is good.
  - Buccaneer+ARP/wARP better+faster than ARP/wARP.
- Typical Coot steps:
  - Connect up any broken chains.
  - Use density fit and rotamer analysis to check rotamers.
  - Check Ramachandran, molprobity, etc.
  - Add waters, ligands, check un-modeled blobs..
  - Re-refine, examine difference maps.



Buccaneer: Summary

A simple, (i.e. MTZ and sequence), very fast method of model building which is robust against resolution.

User reports for structures down to 3.7A when phasing is good.

Results can be further improved by iterating with refinement in refmac (and in future, density modification).

Proven on real world problems.

Use it when resolution is poor or you are in a hurry. If resolution is good and phases are poor, then ARP/wARP may do better. Best approach: Run both!

Nucleic Acid Building

Nautilus:

- A new tool for nucleic acid model building
- Automated (CCP4i) or interactive (Coot)
- Starting from:
  - Experimental phasing
  - Molecular replacement
  - Protein complexes

Nautilus

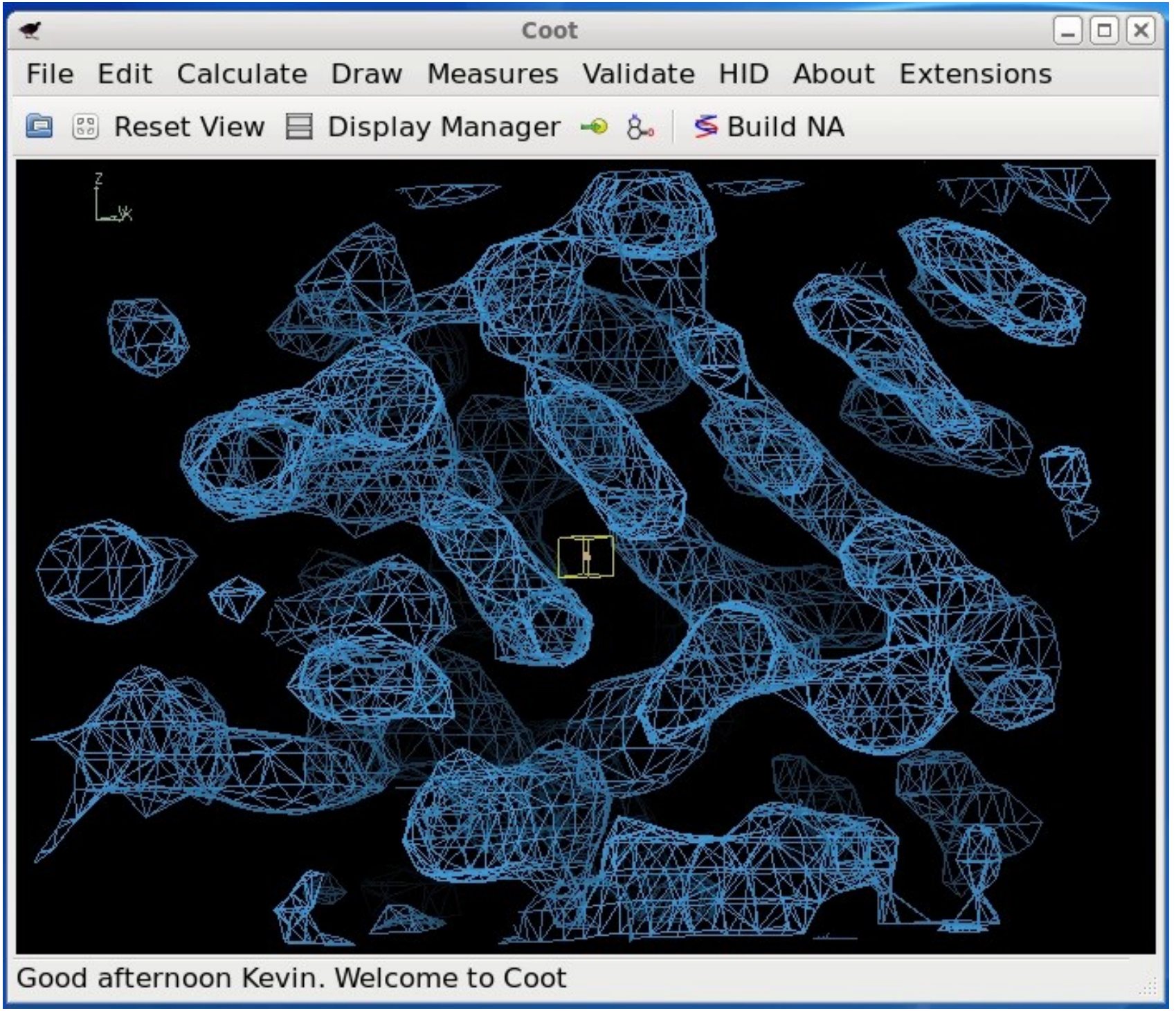
The task:

- To build continuous nucleic acid chains into electron density.
- To assign sequence to those chains.
- To allow addition of nucleotide chains to non-nucleotide structures.

Nautilus

'Fingerprint' detection:

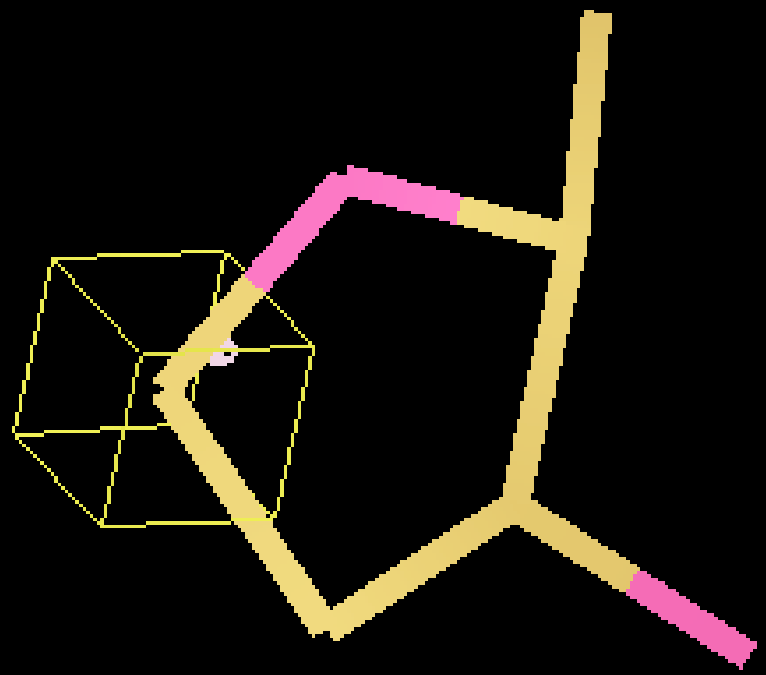
- Identify high and low density features consistent with the presence of nucleic acid features.
- Very fast.
- Related to 'Essens' (Kleywegt and Jones), but with looks at both ridges and troughs.

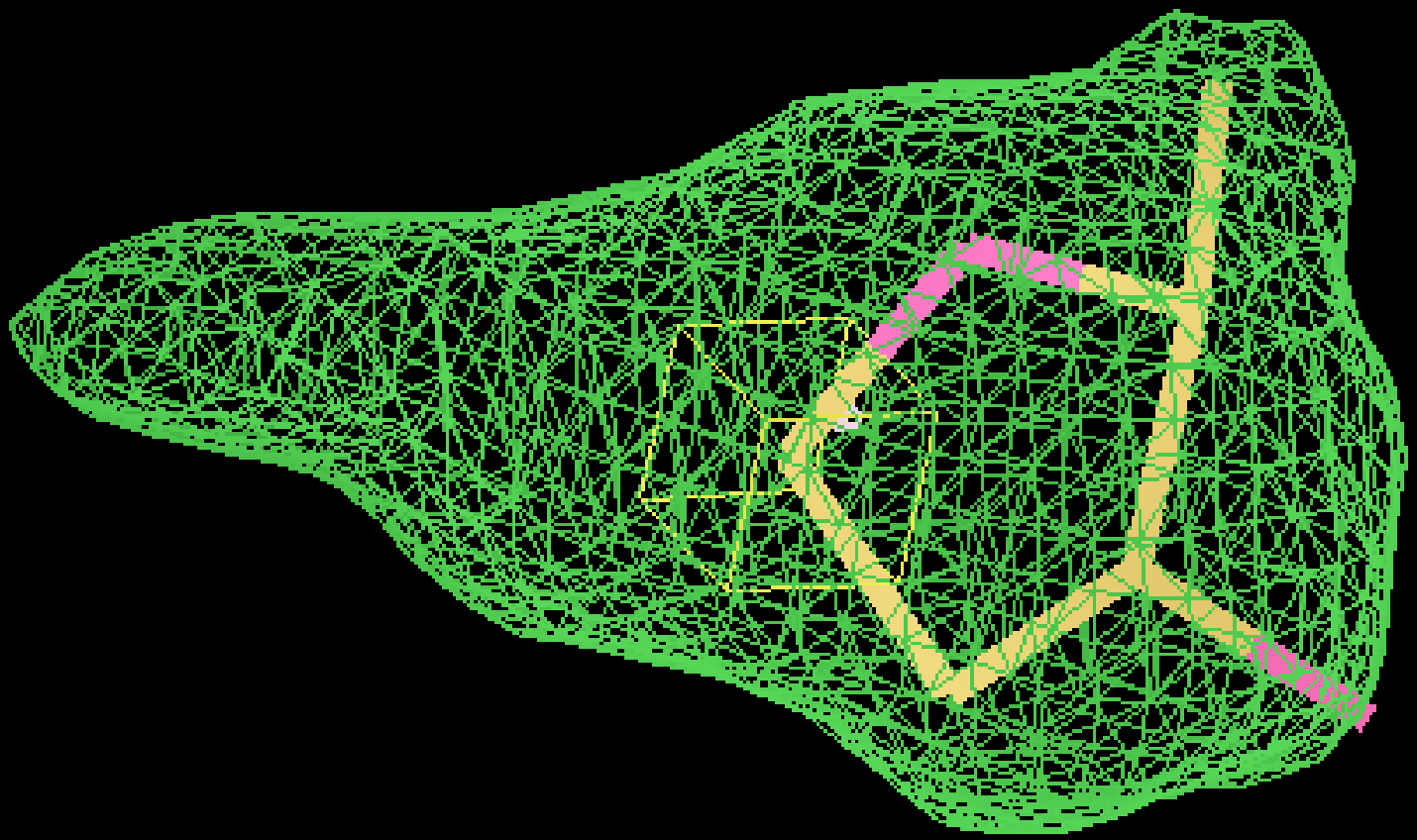

Coot
File Edit Calculate Draw Measures Validate HID About Extensions
Reset View
Display Manager
Build NA
Good afternoon Kevin. Welcome to Coot

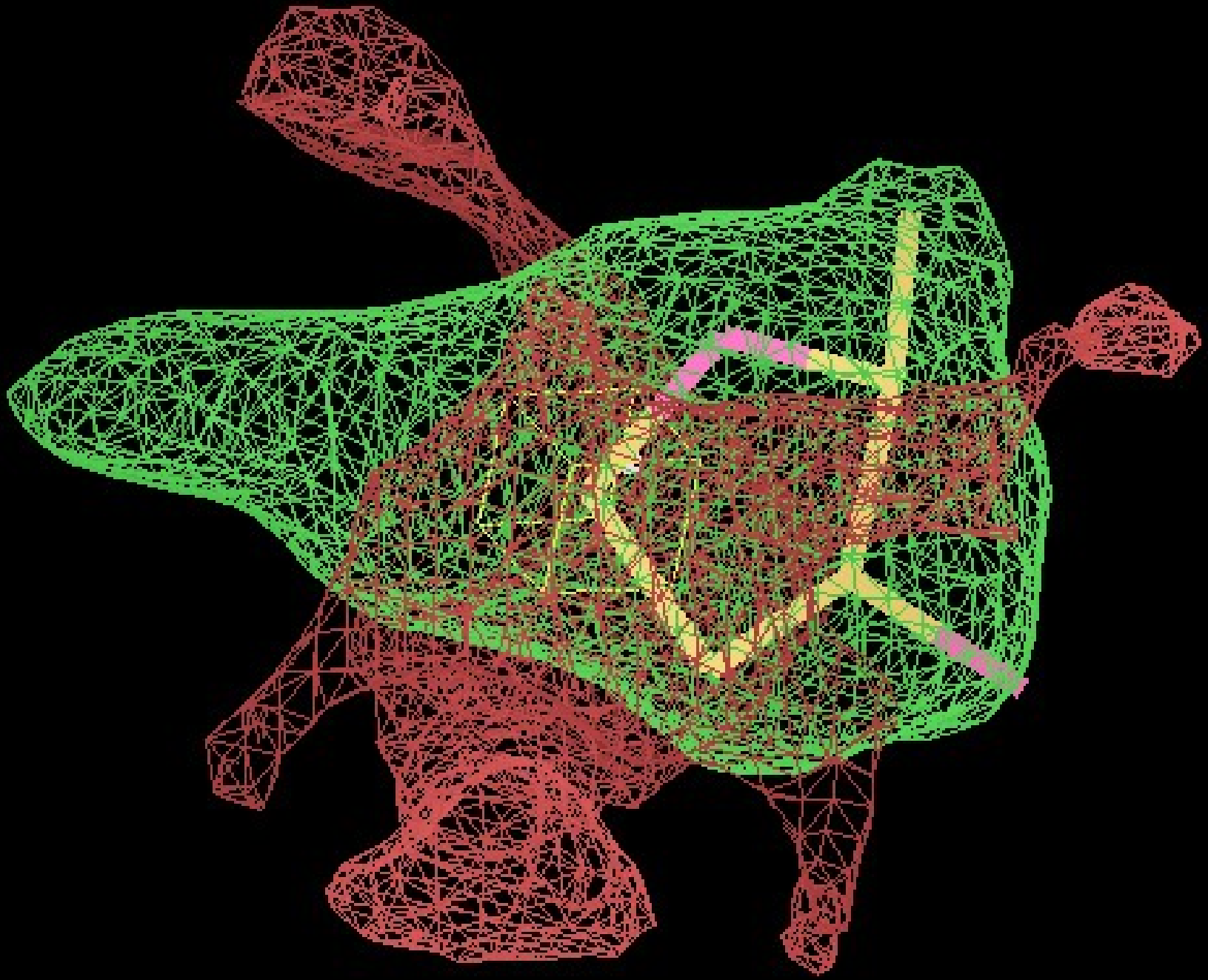
Nautilus

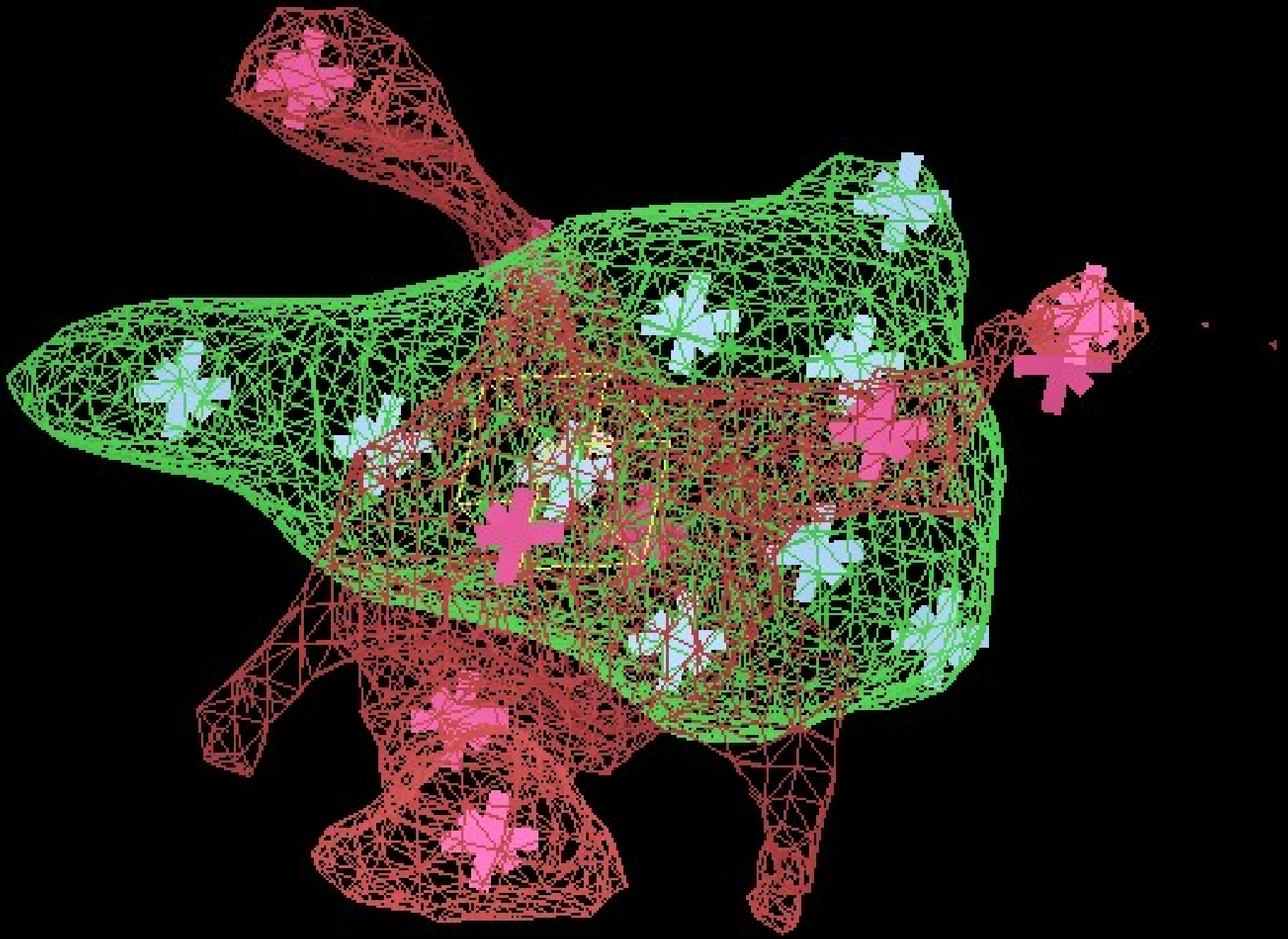
Types of fingerprint:

- Sugar
- Phosphate
- Base type (A/C/G/U)

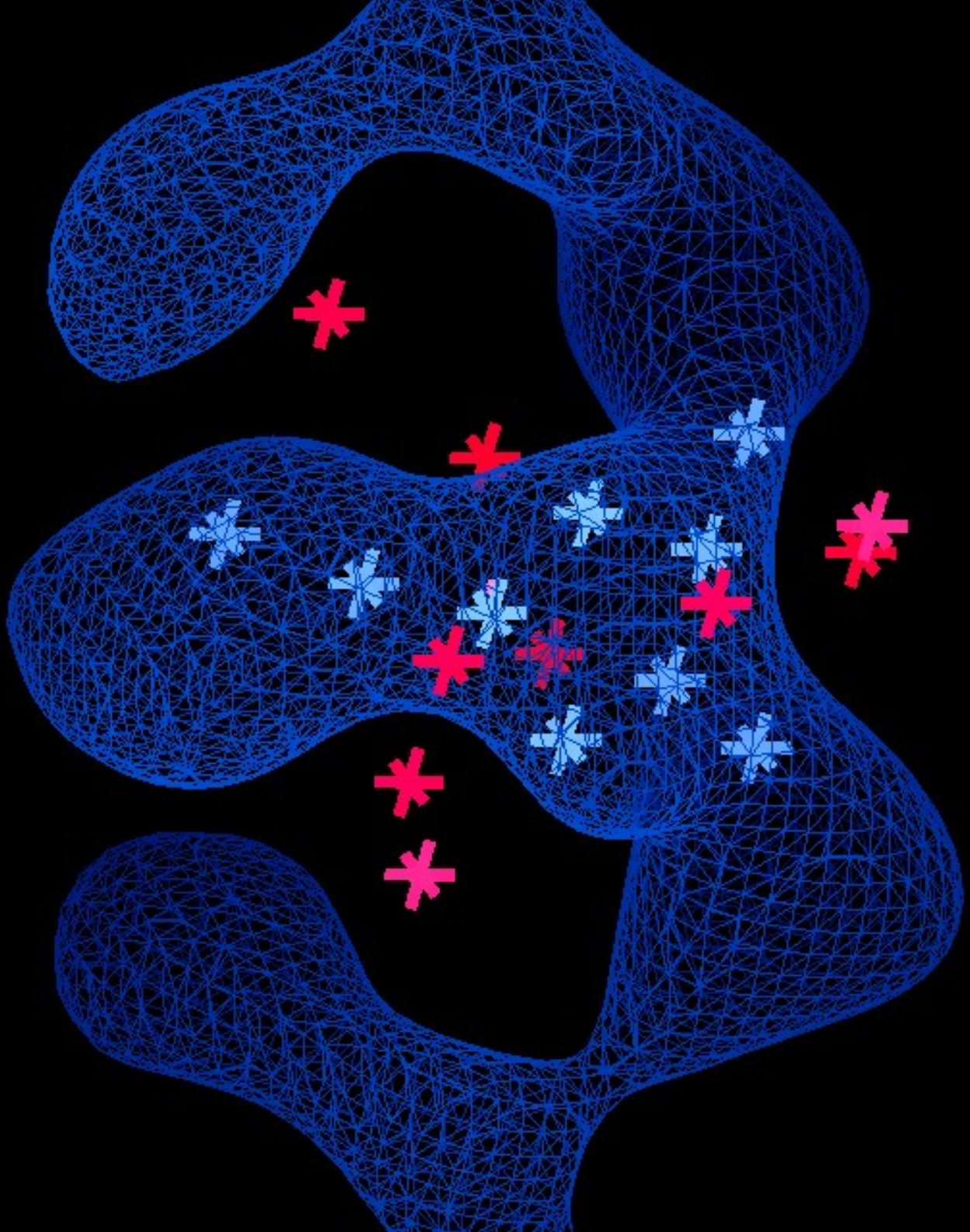




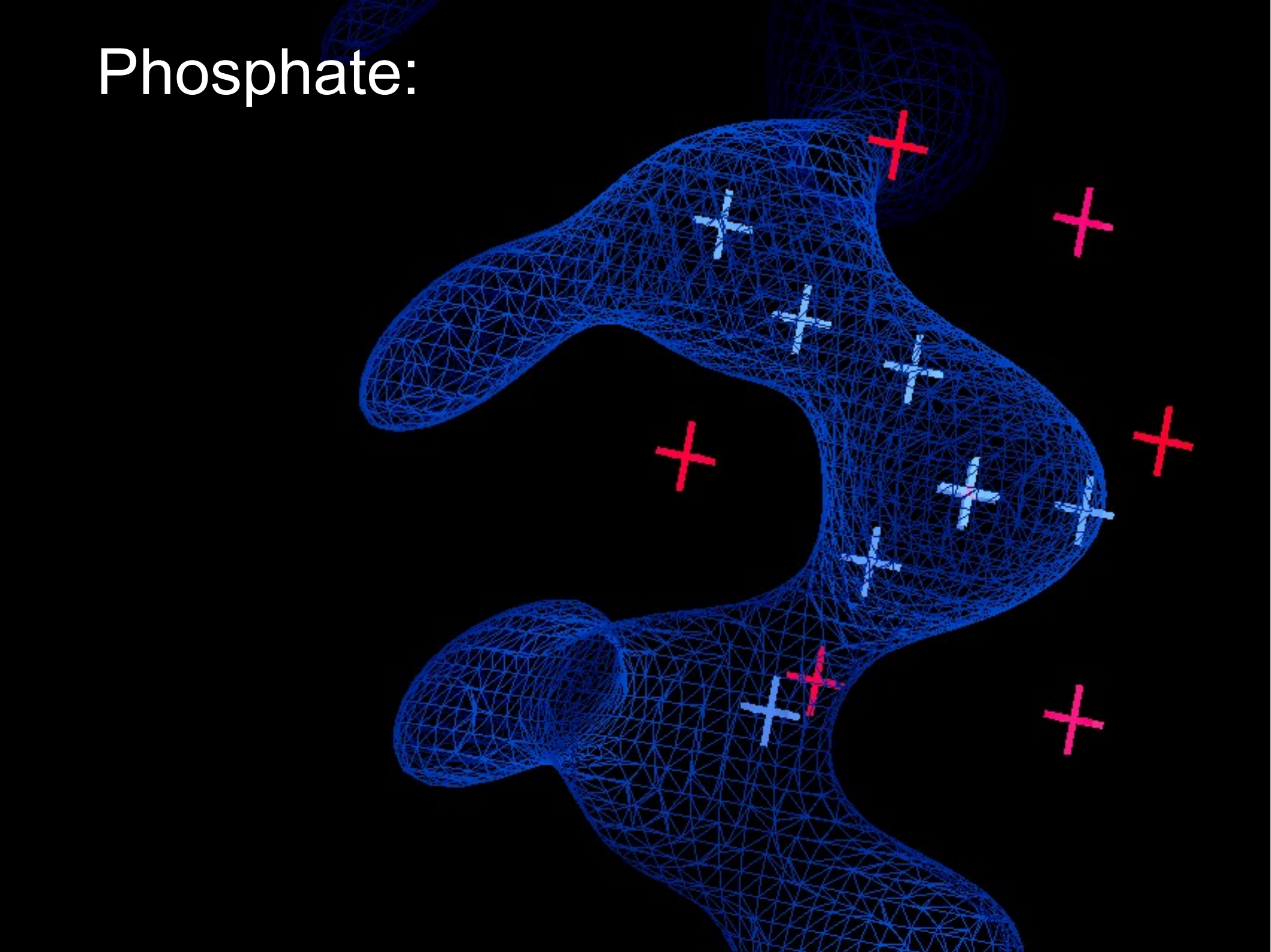




Sugar:



Phosphate:



Nautilus

Use the difference between the mean of the 'high' points and the mean of the 'low' points as a score indicating how likely it is the given group is present at a given position and orientation.

Need to search positions and orientations – a more optimized version of the same target uses the minimum of the highs minus the maximum of the lows – can often stop the calculation before testing all the sample points.

Nautilus

Steps:

- Find chain seeds
- Grow into chains
- Join overlapping chains
- Link nearby chains
- Prune clashing chains
- Rebuild chains to ensure connectivity
- Assign sequence
- Build bases

Nautilus

Find:

- Optimised 6-d rotation-translation using the sugar or phosphate fingerprint.
  - ~5 seconds for whole ASU
- Sugar:
  - Build a single nucleic acid using the best matching equivalent from the database, scored by 1 x sugar + 2 x phosphate fingerprints
- Phosphate:
  - Build a pair of nucleic acids using the best matching equivalent from the database, scored by 1 x phosphate + 2 x sugar fingerprints

Nautilus

Grow:

- Try adding additional nucleic acids to either end of each fragment, scored by the sugar fingerprint and the intermediate phosphate fingerprint.
  - ~1-2 seconds

Join:

- Merge overlapping fragments into longer fragments
  - <0.1 second

Link:

- Join fragments with nearby 3' and 5' terminii
  - ~0.5 second

Nautilus

Prune:

- Eliminate clashing regions
  - <0.1 second

Rebuild chains:

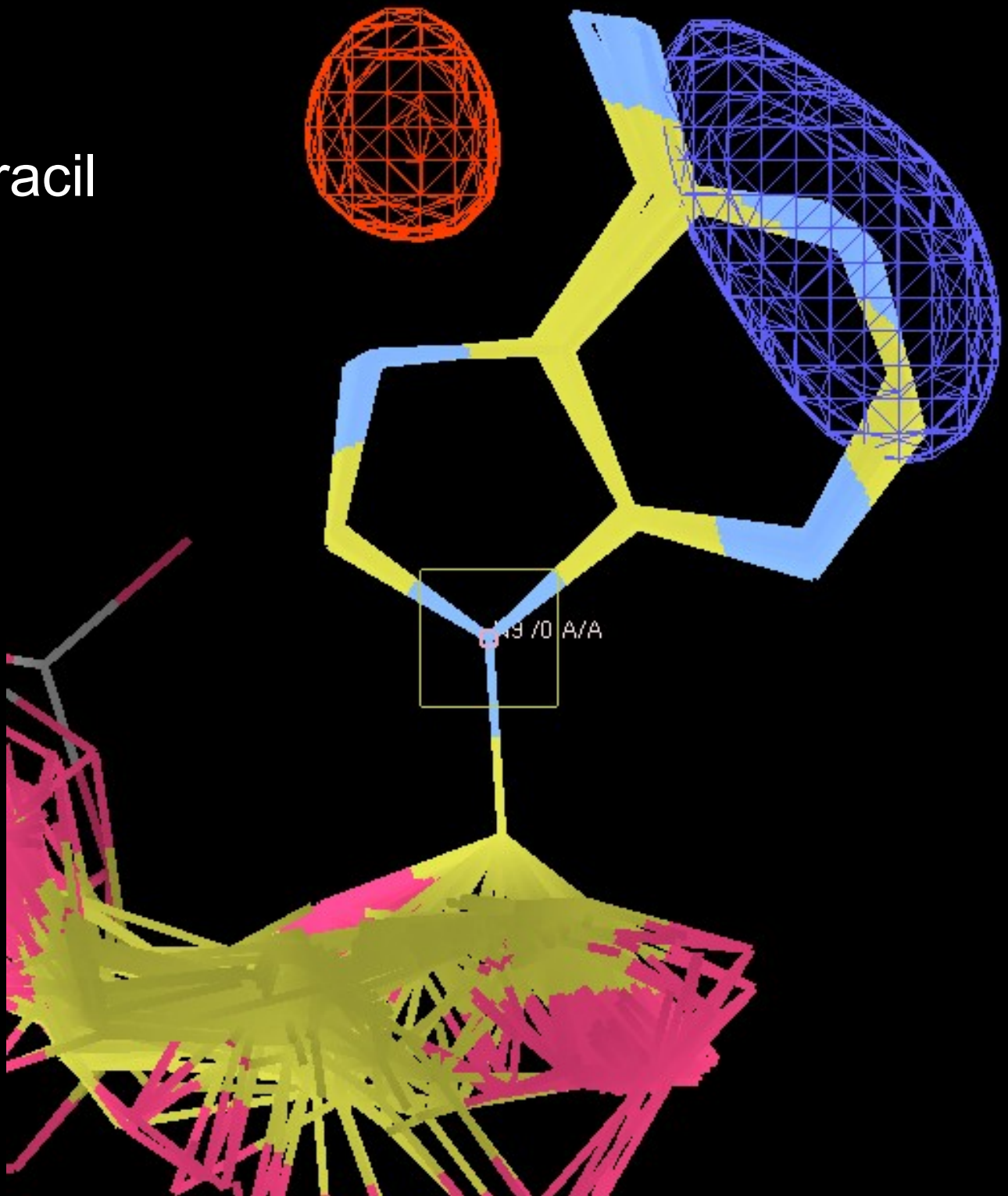
- Rebuild each sugar-sugar link using a fragment from the database
  - ~0.3 seconds

Sequence:

- Score base-type fingerprints at each position and assign sequence
  - <0.1 second

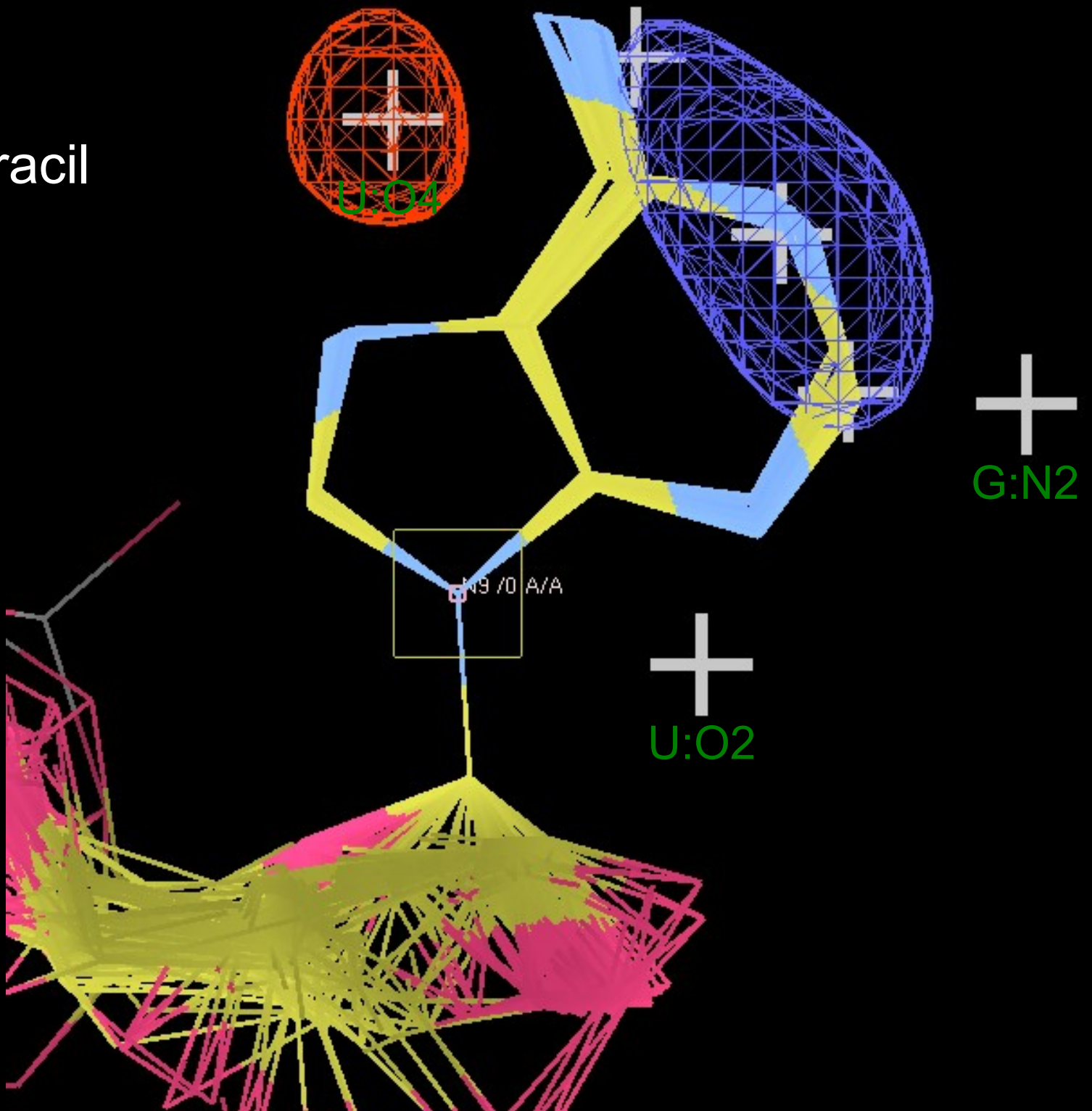
Base:

Adenine-Uracil



Base:

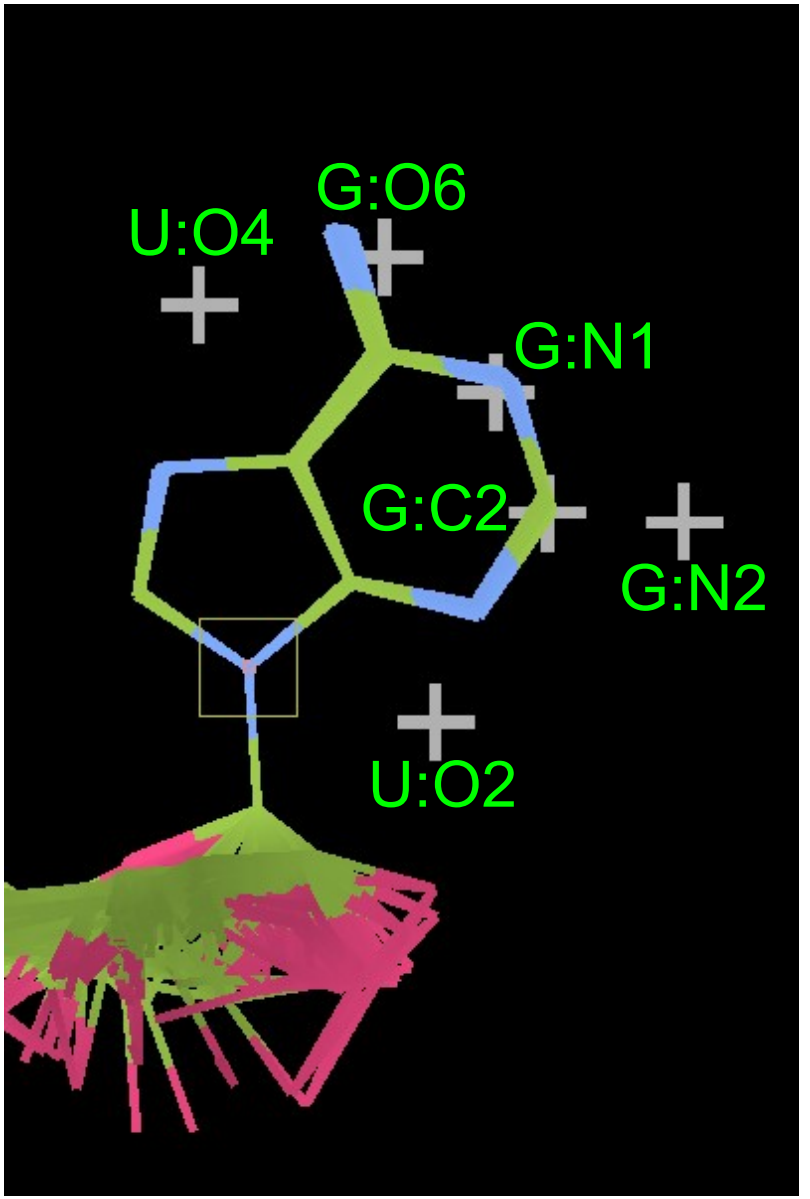
Adenine-Uracil



Nautilus

Adenine:

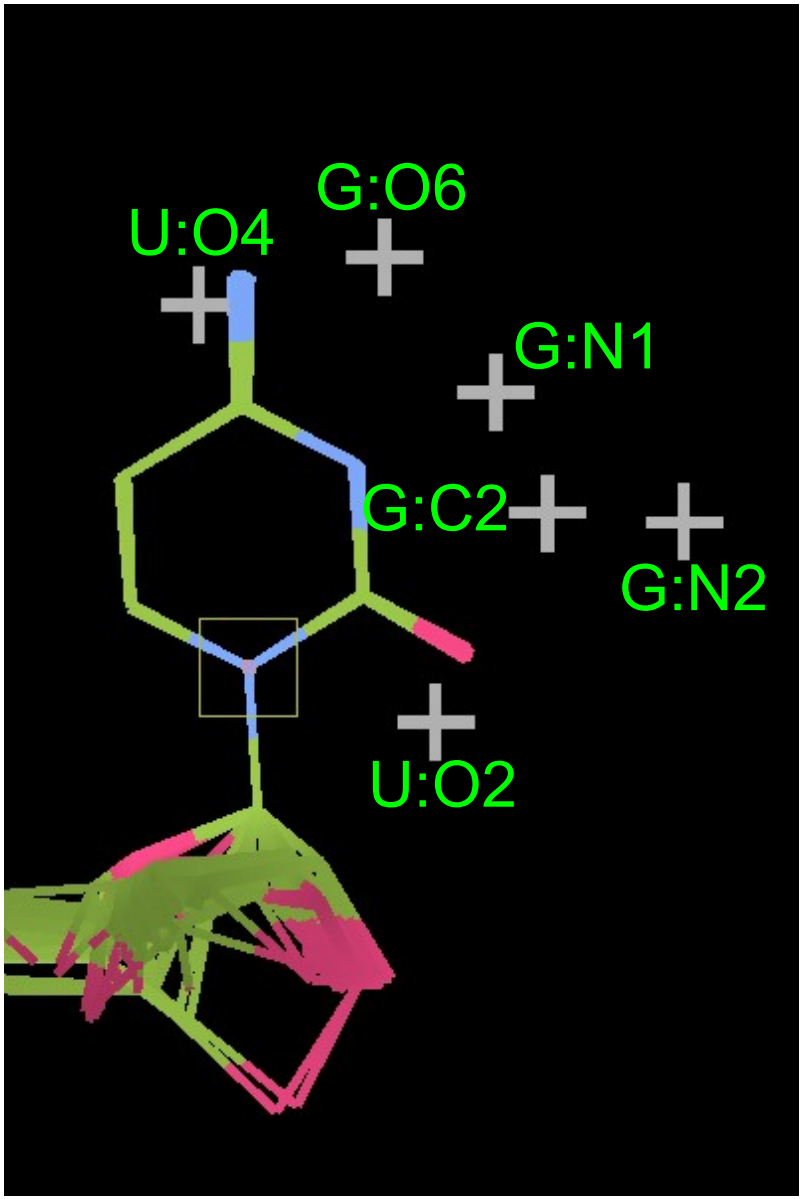
	U: O4	U: O2	G: O6	G: N1	G: C2	G: N2
A	-	-	+	+	+	-
C						
G						
U						



Nautilus

Cytosine:

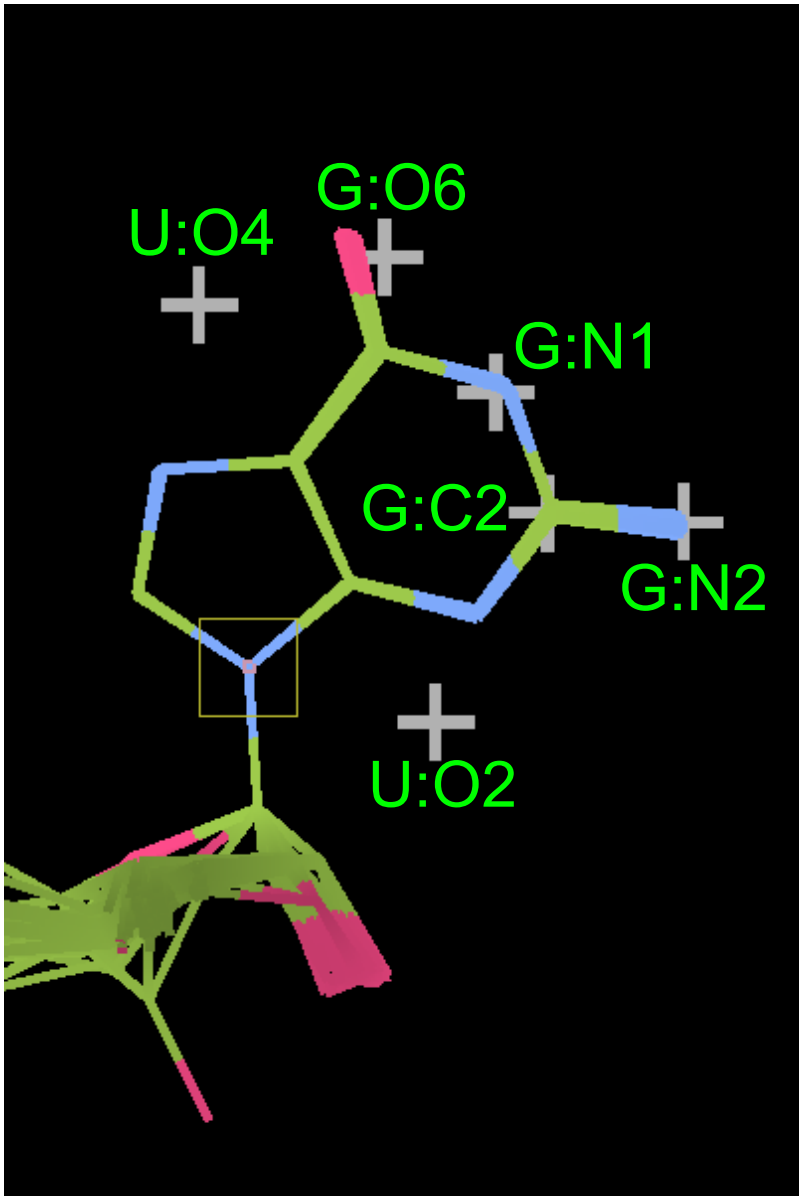
	U: O4	U: O2	G: O6	G: N1	G: C2	G: N2
A	-	-	+	+	+	-
C	+	+	-	-	-	-
G						
U						



Nautilus

Guanine:

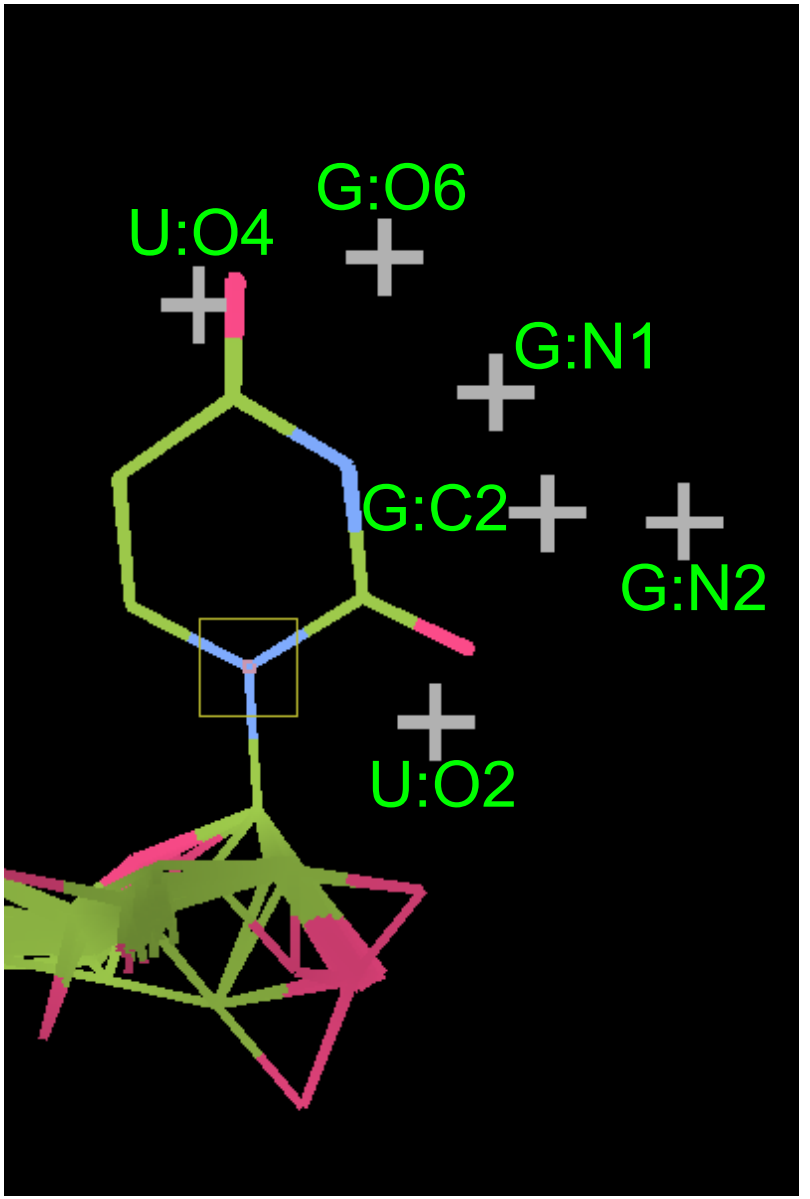
	U: O4	U: O2	G: O6	G: N1	G: C2	G: N2
A	-	-	+	+	+	-
C	+	+	-	-	-	-
G	-	-	+	+	+	+
U						



Nautilus

Guanine:

	U: O4	U: O2	G: O6	G: N1	G: C2	G: N2
A	-	-	+	+	+	-
C	+	+	-	-	-	-
G	-	-	+	+	+	+
U	+	+	-	-	-	-



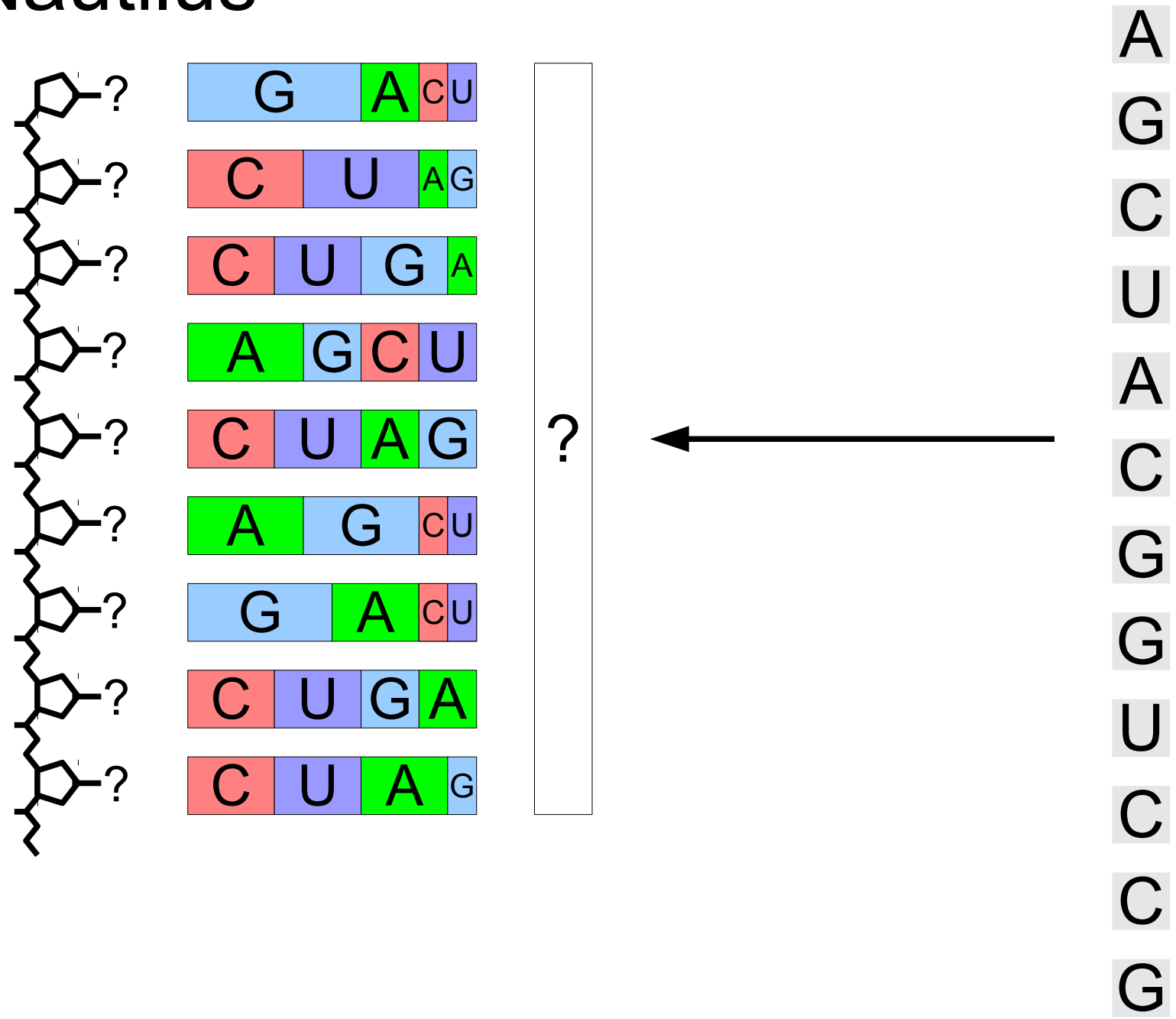
Nautilus

But the real world isn't black and white. Ideally we want a probability of a base being of a particular type.

- Calculate z-scored densities for the density at each of the 6 sample positions for 200 bases (50 of each type), to form a sample database.
- Calculate z-scored densities for the 6 sample positions of the unknown base.
- Find the 50 closest matches to the unknown base from the database.
- Assign probability of being A/C/G/U on the basis of the proportion of of the 50 closest matches being of each type (+ an error term).

Google: k-NN (k-Nearest Neighbour)

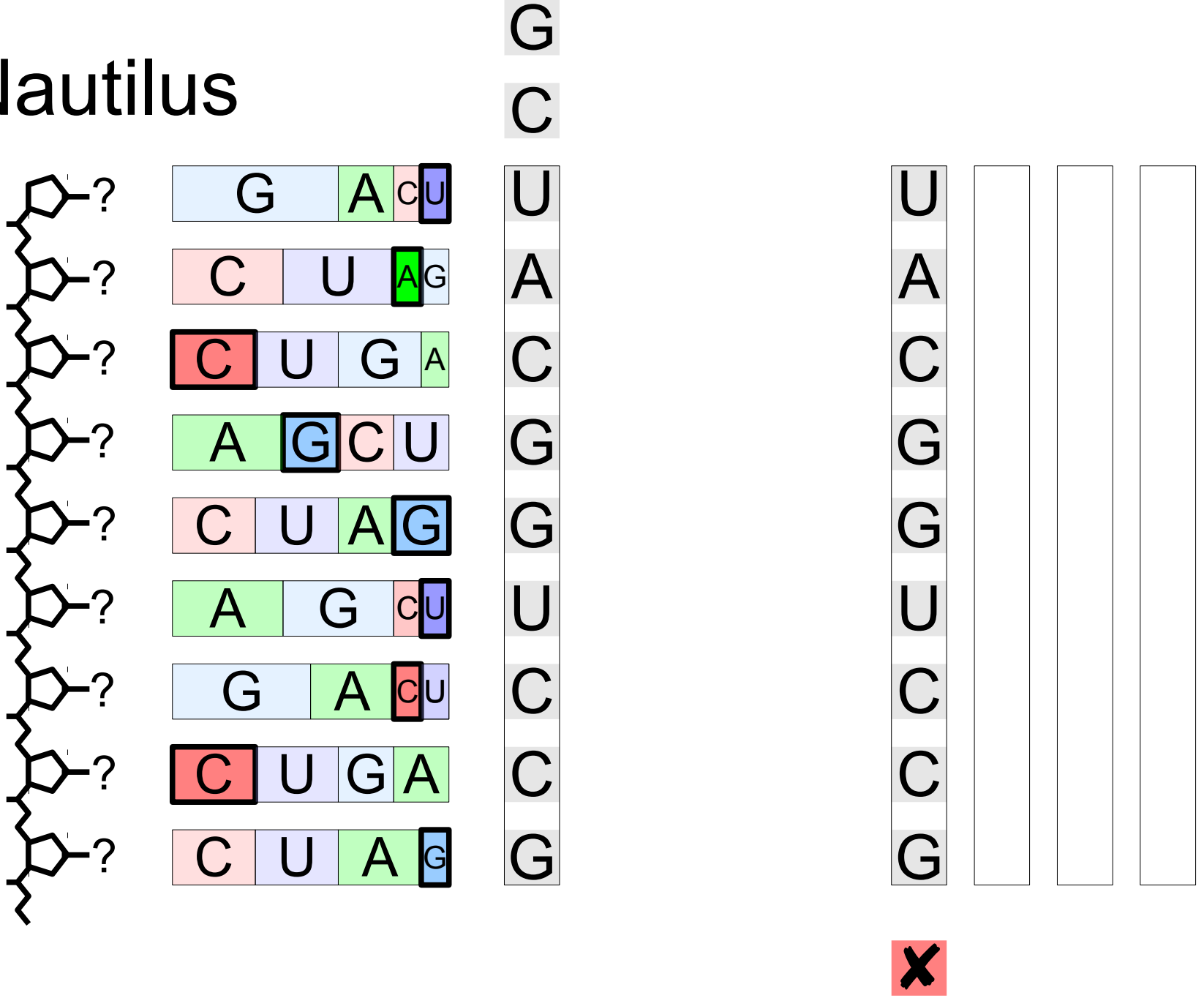
Nautilus



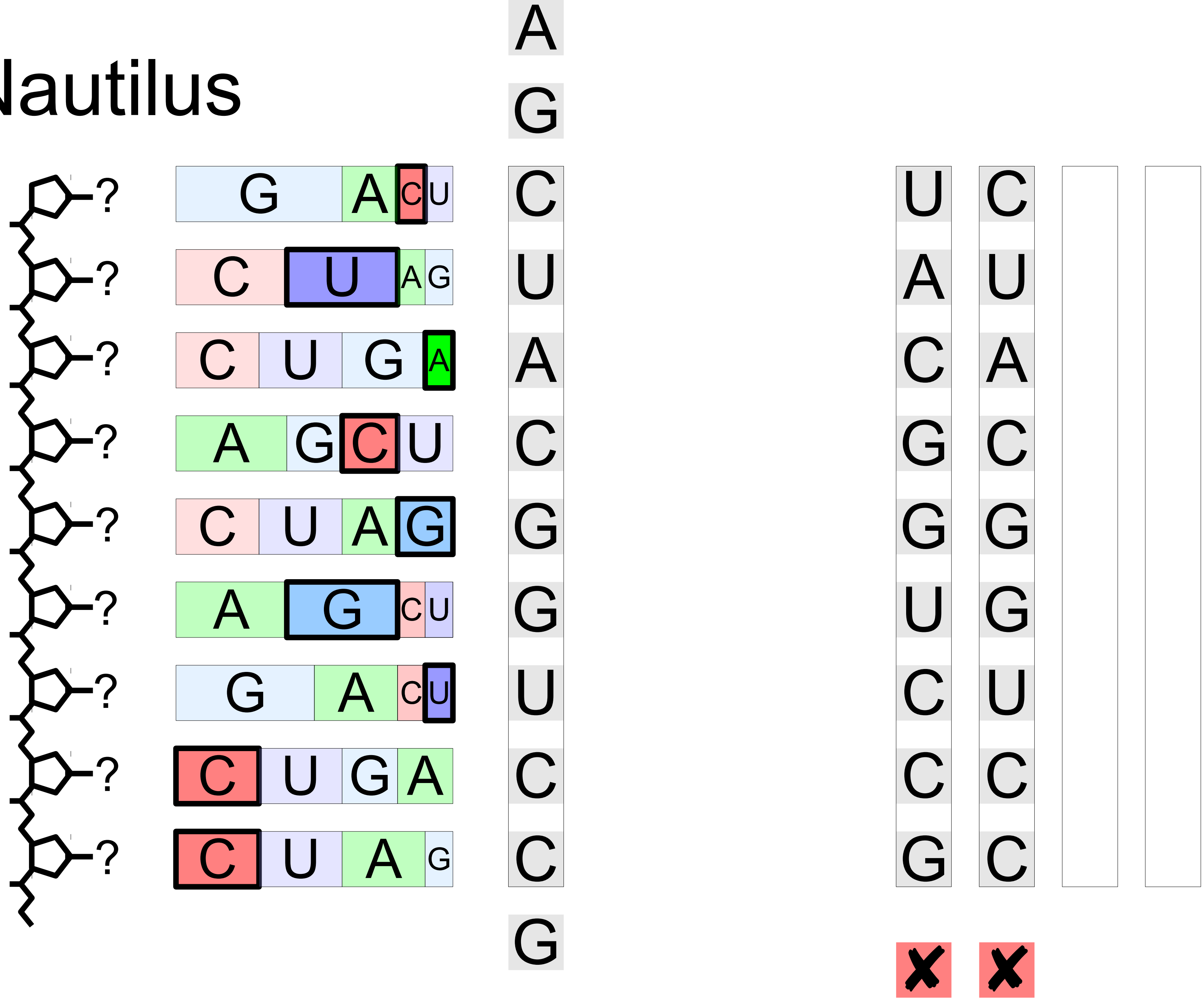
Kevin Cowtan, kevin.cowtan@york.ac.uk

2015

Nautilus

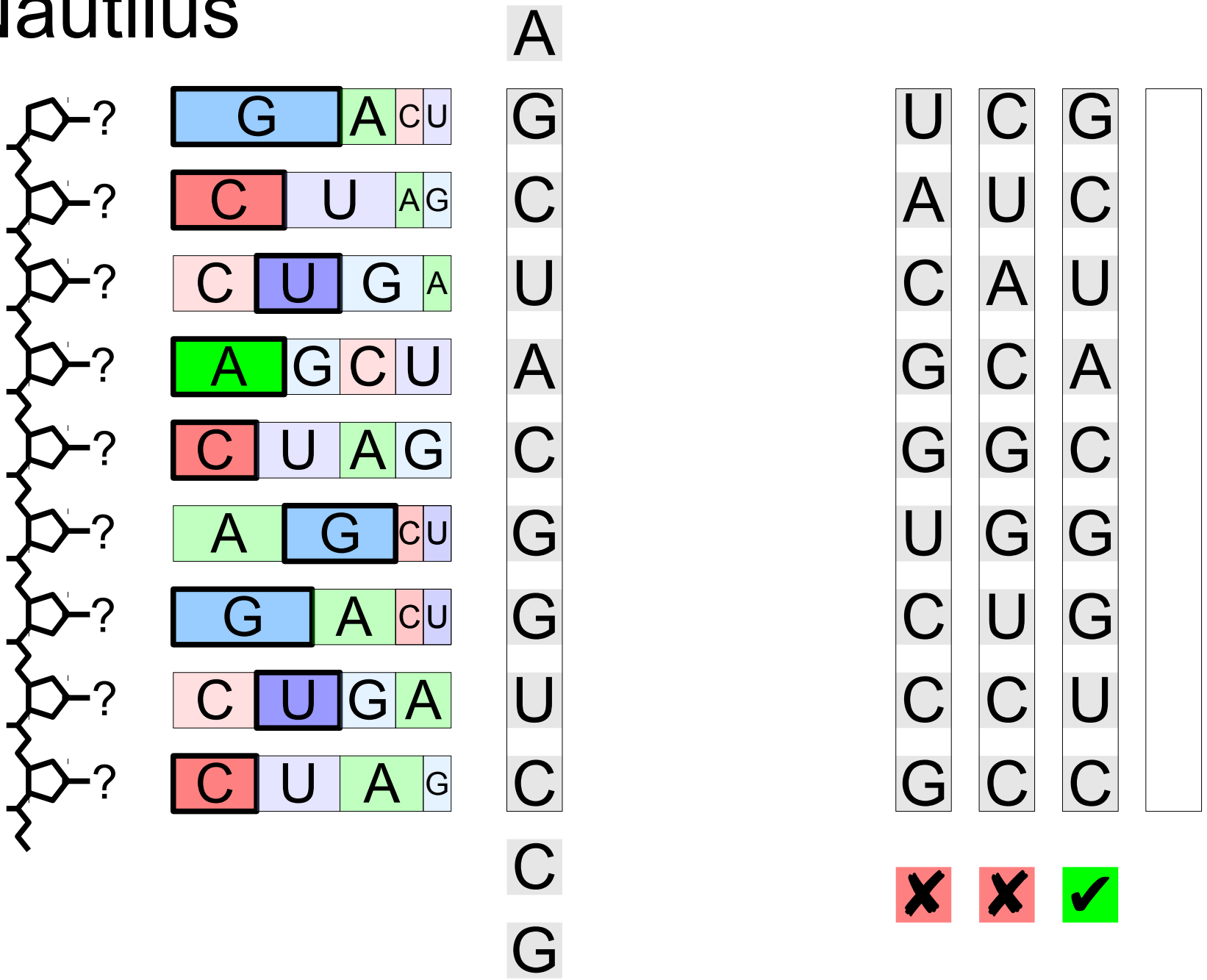

G
C
U
A
C
G
G
U
C
C
G
U
A
C
G
G
U
C
C
G
?
?
?
?
?
?
?
?
?
?
G A C U
C U A G
C U G A
A G C U
C U A G
A G C U
G A C U
C U G A
C U A G

Nautilus



Kevin Cowtan, kevin.cowtan@york.ac.uk 2015

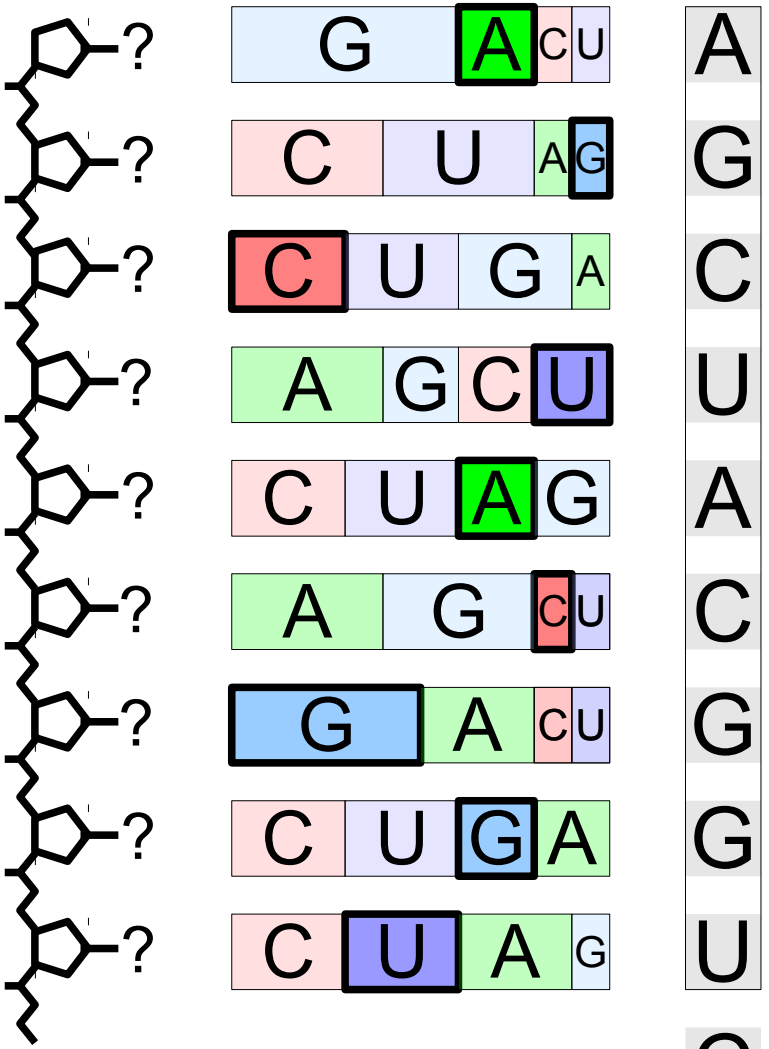
Nautilus



Kevin Cowtan, kevin.cowtan@york.ac.uk

2015

Nautilus



Kevin Cowtan, kevin.cowtan@york.ac.uk

2015

Nautilus

Results:

- Good results on synthetic noisy data at 3.5A and user reports on real data at 3.8A.
  - Need more data
- Like '*buccaneer*', phases are more important than resolution.
- Failed on a quadruplex structure with good phases.
  - Try a different database?

Achnowledgements

Help:

- JCSG data archive: www.jcsg.org
- Garib Murshudov, Raj Pannu, Pavol Skubak
- Eleanor Dodson, Paul Emsley, Randy Read, Clemens Vonrhein

Funding:

- The Royal Society, BBSRC