

Molecular Modeling 2020 -- Lecture 19. Fri Apr 3 Homology Modeling

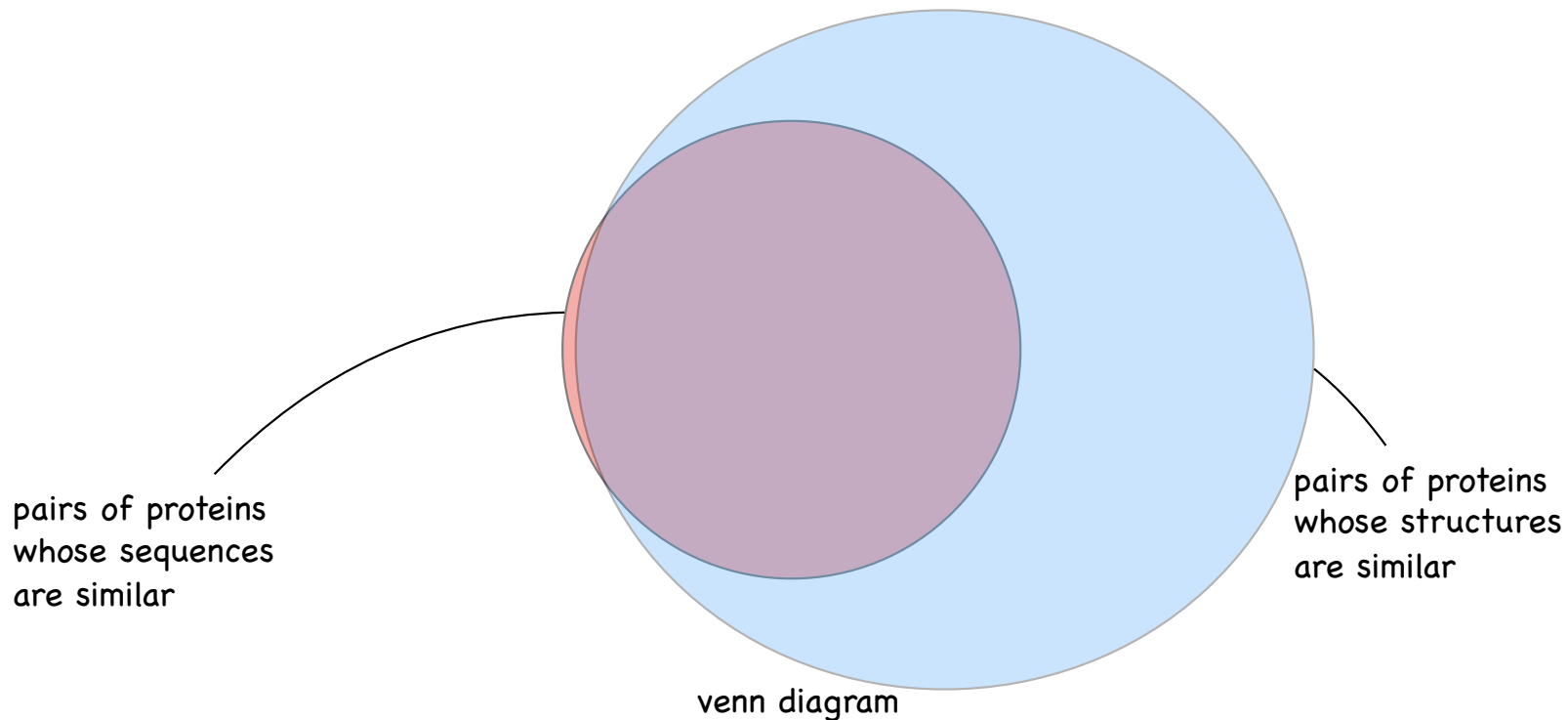
Database search

Alignment

Automated modeling

Indels

If the sequence is similar, then the structure is similar.



Homology modeling

also called “comparative modeling”

- Sequences that have similar sequence have similar structure.
- Therefore we can model a sequence of unknown structure based on a homolog of known structure.
- Differences between homolog functions can be inferred.



Searching for a homolog of known structure.

Download "Sequence 1" from <http://www.bioinfo.rpi.edu/bystrc/courses/biol4550/biol4550.html>. Name it "**strepto.fasta**"

Open it in **MOE**

SEQ: Protein > Search > PDB

In the database search window, **Load chain** (select strepto)

Search

Choose the hit with the best e-value

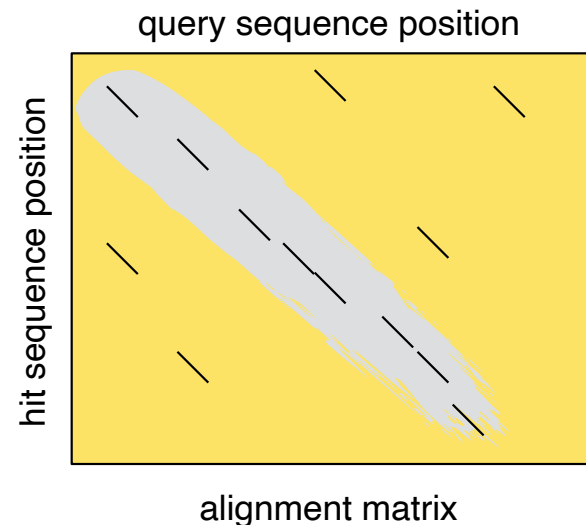
Inspect the alignment

Load the alignment. **Close** the search window,

In **SEQ** window, color residues by similarity (bottom bar **Residues > Similarity**)

Sequence Database search -- what's going on.

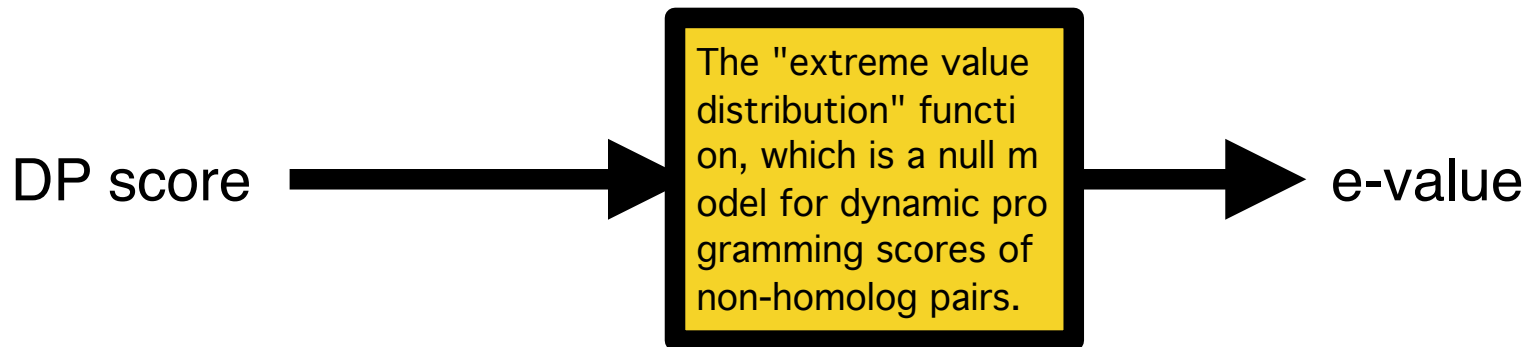
- Your sequence (**query**) is chopped up into 3-tuples.
- Every 3-tuple (there are exactly 8000) has its own look-up table, or **index**, of database locations (pdbcode, chain ID, position)
- **Hits** are chains with the most 3-tuples arranged along a diagonal on the **alignment matrix**, query vs hit.
- Hits are aligned to query using the **dynamic programming algorithm** (Smith-Waterman)
- The Dynamic Programming score is converted to a statistic, called the **e-value**.



e-value

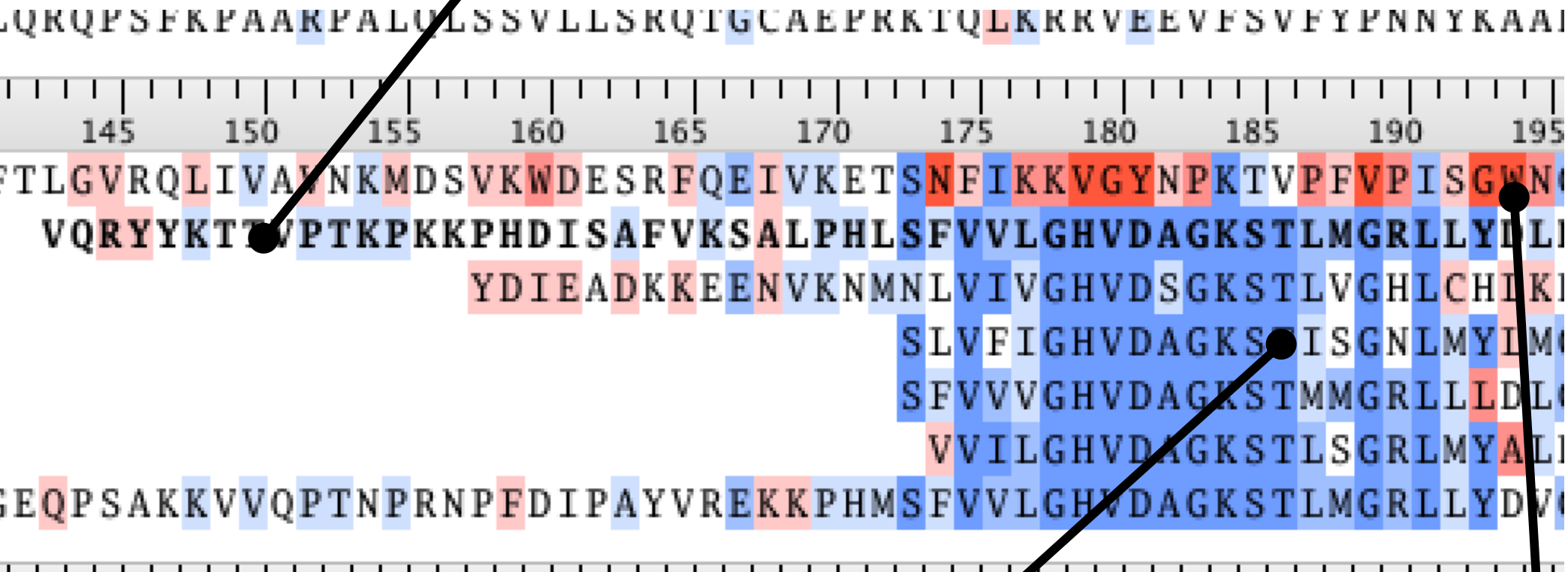
The significance test for a database search

- The number of times in a database search that you will get a random, non-homologous hit with the same score or better.



After loading alignment.

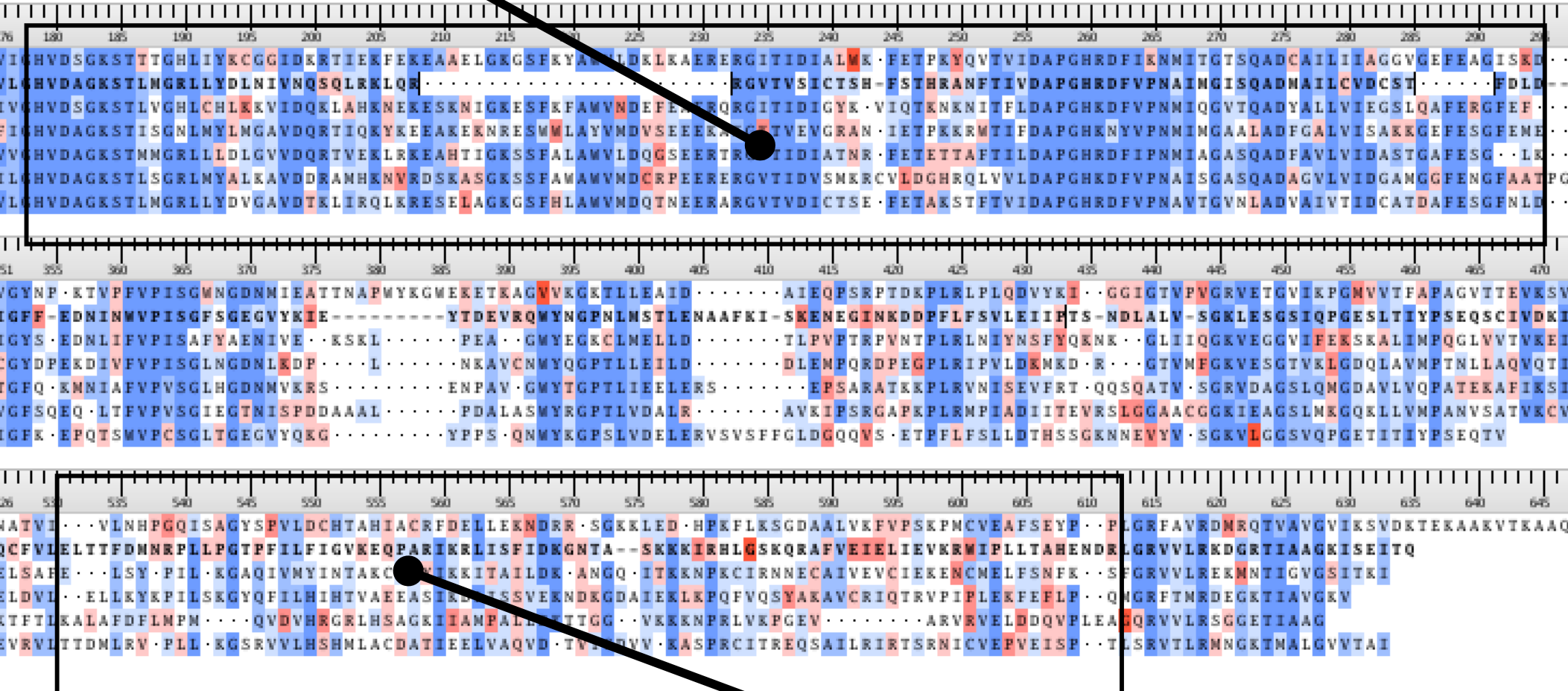
bold sequence, has coordinates



not bold, does not have coordinates

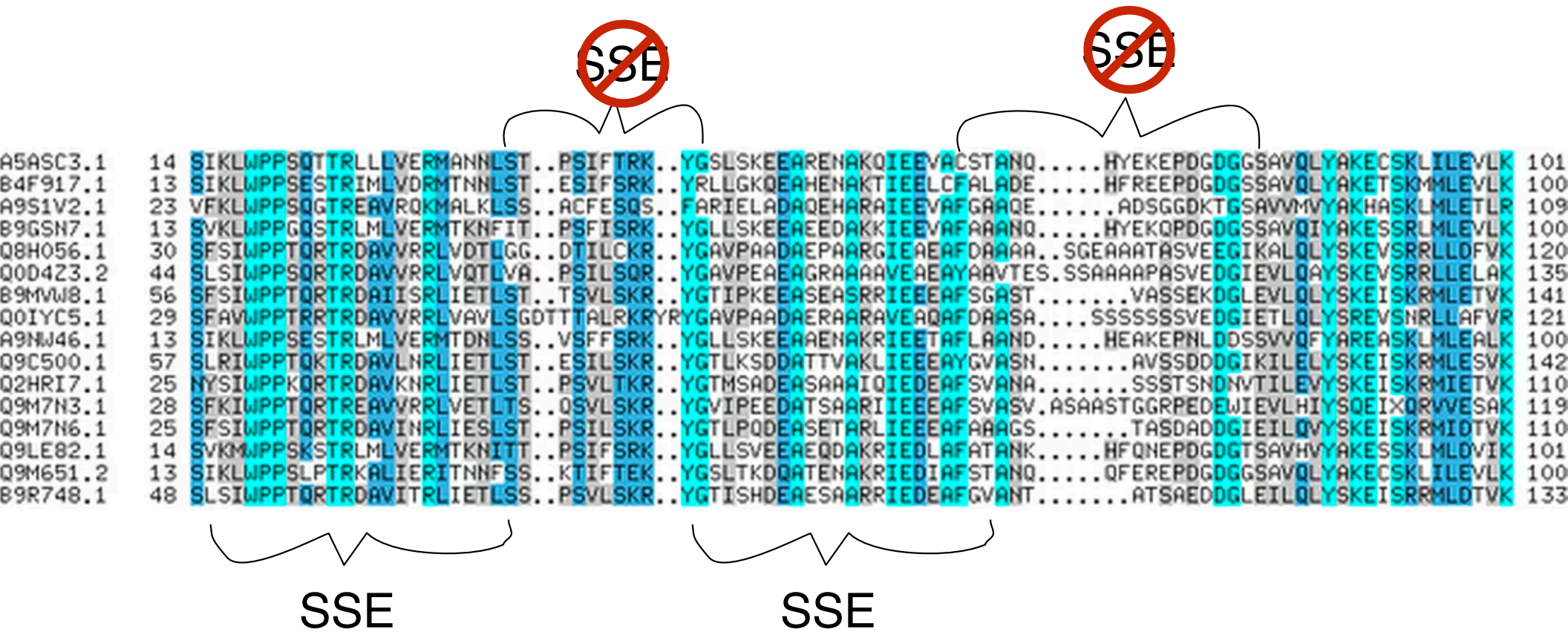
Residues>Similarity colors well-aligned regions blue, poorly aligned red.

Well aligned region. Bluish in *Similarity* coloring. Not so many indels.



Poorly aligned region. Reddish in *Similarity* coloring. Gaps all over the place.

Secondary structure elements (SSE) should not have Indels



A rare exception to the *no-insertions-in-helix* rule

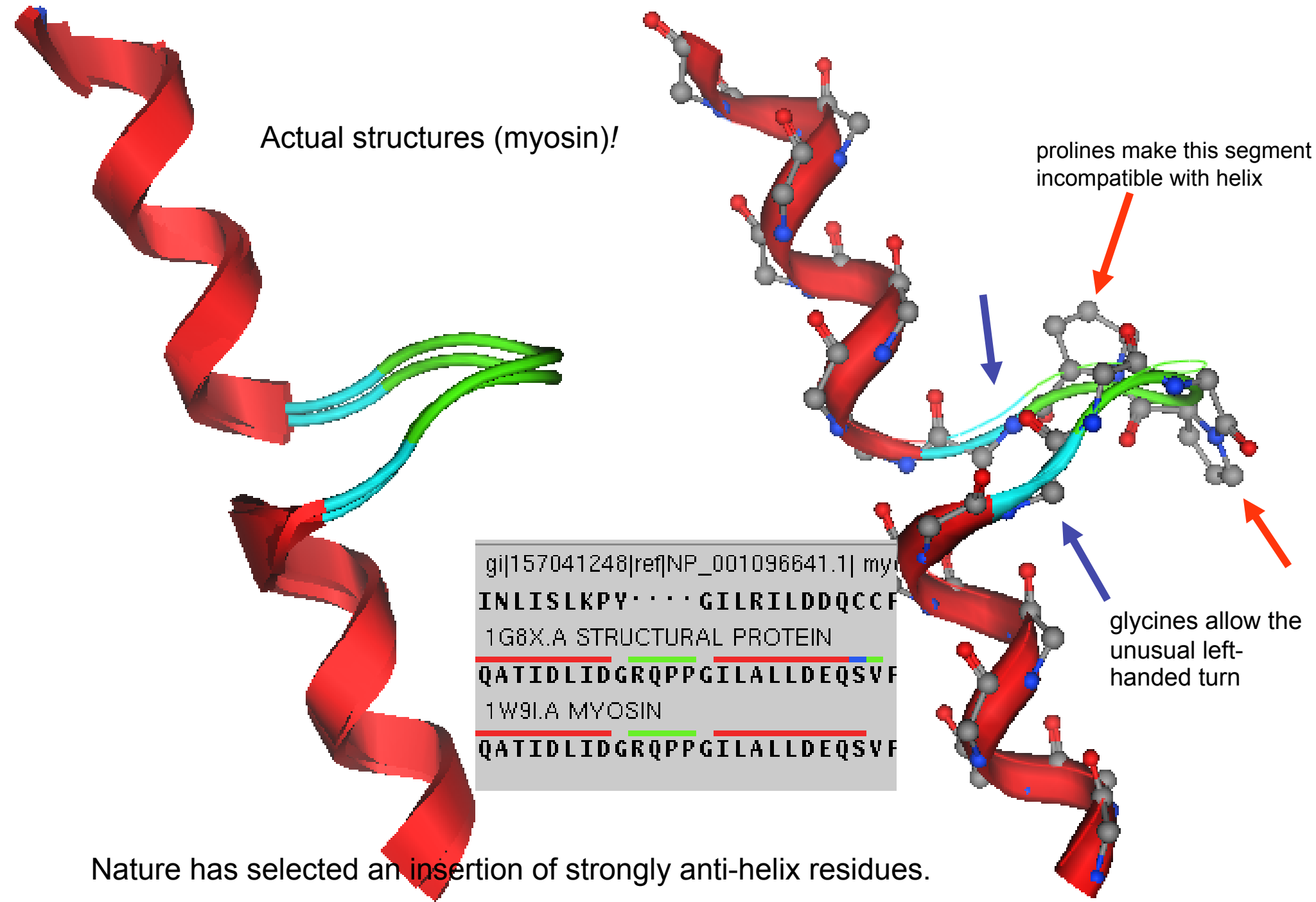
Actual structures (myosin)!

prolines make this segment incompatible with helix

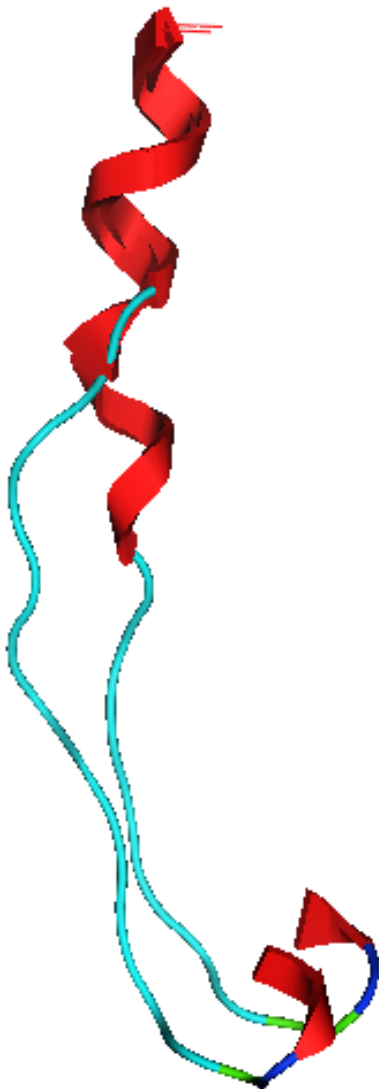
glycines allow the unusual left-handed turn

```
gi|157041248|ref|NP_001096641.1| myo  
INLISLKPY· · · GILRILDDQCCF  
1G8X.A STRUCTURAL PROTEIN  
QATIDLIDGRQPPGILALLDEQSVF  
1W9I.A MYOSIN  
QATIDLIDGRQPPGILALLDEQSVF
```

Nature has selected an insertion of strongly anti-helix residues.



A deletion within a helix actually leads to the destruction or truncation of the helix.



Deletion in the sequence leads to shorter helix, not a helix with a "*deletion*" in it.

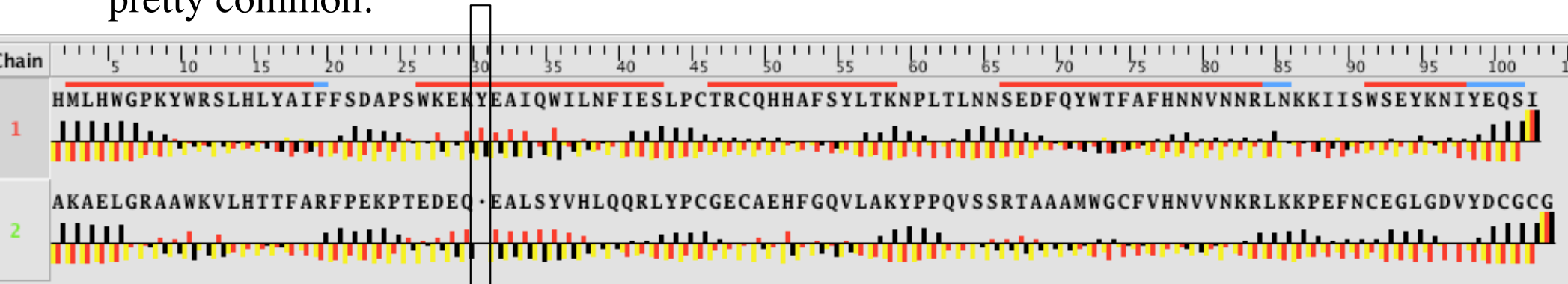
This shows two actual structures. The one with the deletion is more extended, to span the distance.

```
AFADNQ·PCINLIS  
TFIDFG LDSQATID
```

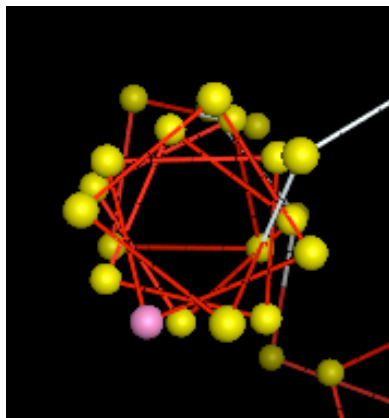
Alignment

Deletion falls within a helix. What happens?

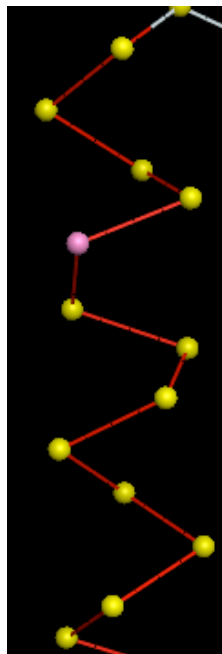
True indels within a contiguous alpha helix are rare, but **alignment errors** are pretty common.



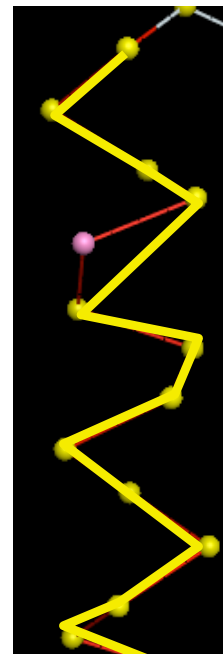
Deletion in the middle of a helix



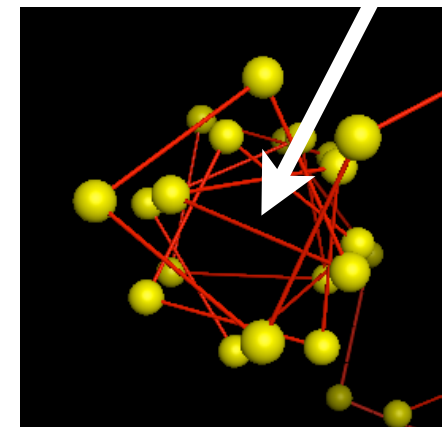
before deletion



after

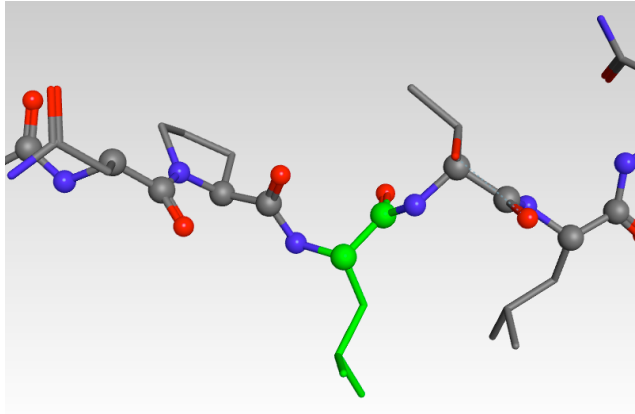


chain crosses through middle of helix!?

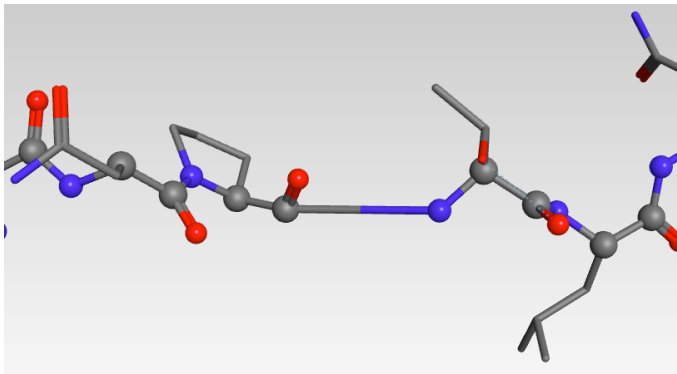


Deletion within strand is virtually impossible.

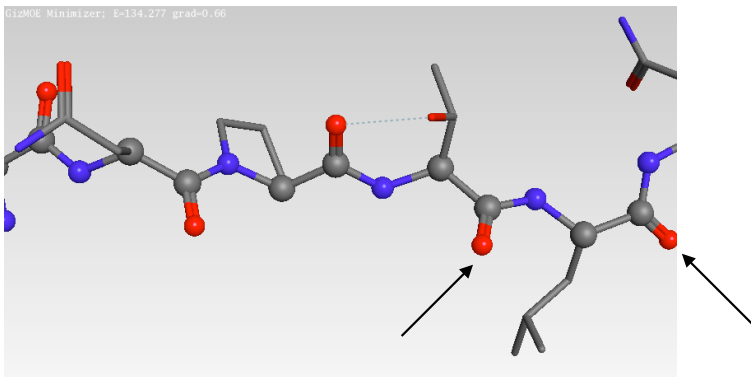
What happens if you try to delete one residue?



Before deletion. Residue in green will be deleted.



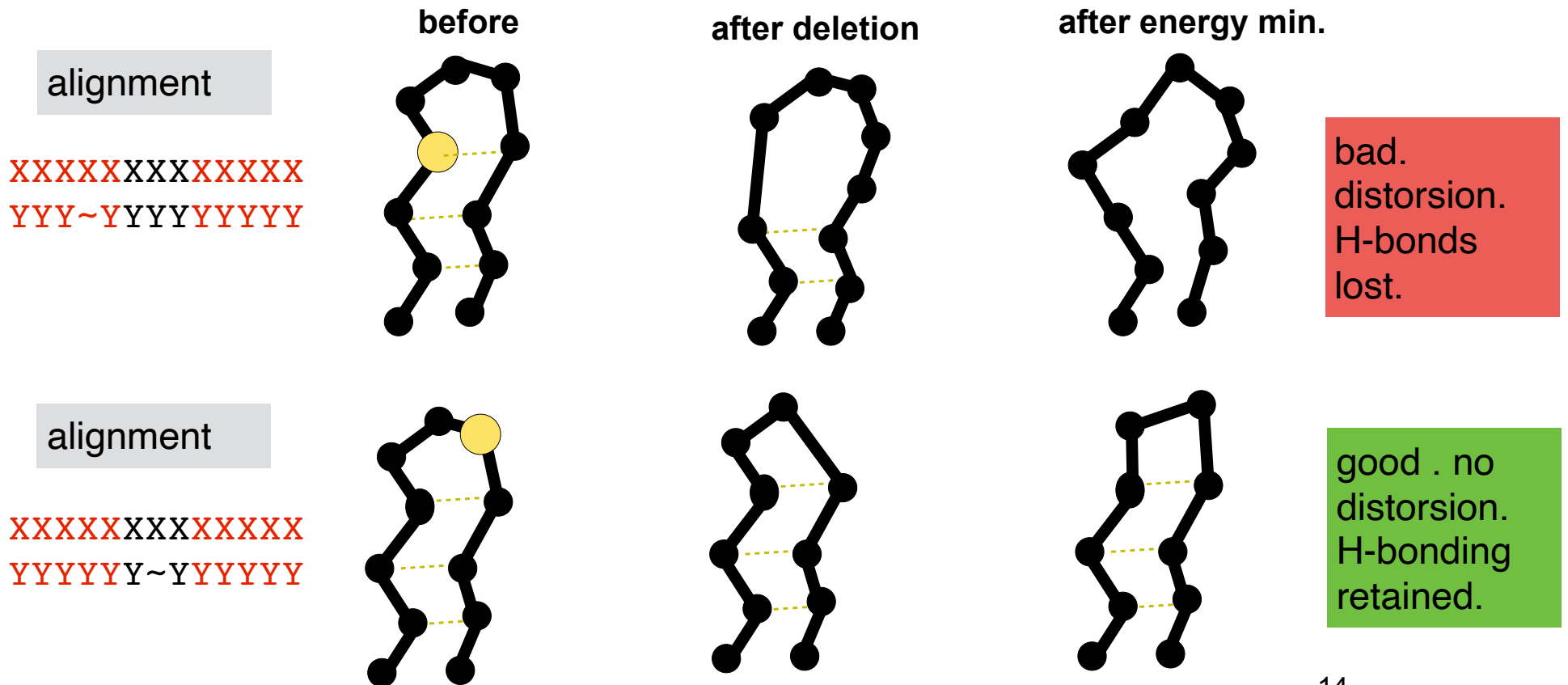
After deletion. Before energy minimization. Long peptide bond.



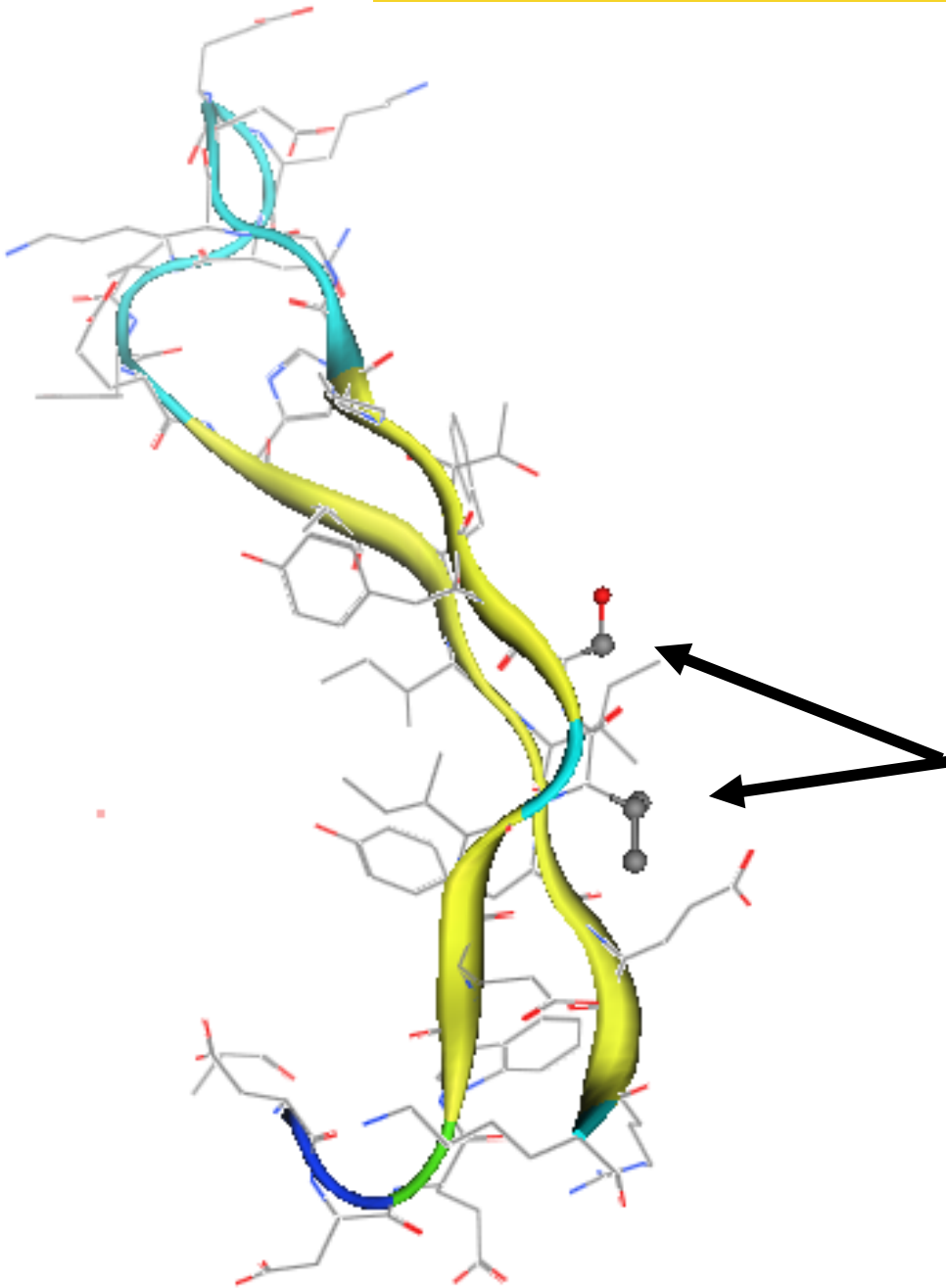
After energy minimization. Long peptide bond. Bad energy. **Hyper-extended. H-bonds pointing the wrong way.**

Move beta strand indel to turn

- Moving a deletion out of a helix or strand avoids disruption of the SSE caused by the energy minimization (which is incapable of saving a bad alignment).
- Moving a deletion move the disruption to where it can be tolerated.



Beta-bulges



Beta-bulge :

Two sidechains (usually polar) point to one side of the sheet, instead of just one. This causes a kink in an otherwise continuous pairing.

A beta-bulge is almost always a 1-residue insertion, here, usually a polar amino acid, most frequently D.

Occasionally, a many residue insertion occurs here. In that case it is called a “**beta-blowout.**”

Evolutionary significance of an alignment

Given this alignment...

target

ACDEFGHIKLMNPQRSTVWY

| | : | | | | : | | | | | :
| | : | | | | : | | | | | :

.CDDFGACDGHIIYIM . .QQSTVWF

template

Biologist infers...

- The gene was extended by one residue at the N-terminus.
- The Phe is conserved.
- Four residue deletion occurred between G to H.
- A non-similar mutation Y->K occurred.
- A two-residue insertion occurred between M and Q.
- A similar mutation F->Y occurred.

Aligned positions share a common ancestral position.

An alignment as modeling instructions

Given this alignment...



•Modeler program should...

- Add Ala to the N-terminal Cys using energy minimization.
 - Keep the conserved Phe sidechain and backbone.
 - Cut out the four residue insertion and connect G to H.
 - Switch non-similar sidechains Y->K. Possibly move backbone.
- Possibly pick another alignment.
- Cut at M-Q, insert two residues, Asn-Pro.
 - Switch similar sidechains F->Y. Keep backbone fixed.

Aligned positions share a common spatial position.

The alignment tells MOE how to do the loop search

target

ACDEFG HIKLMNPQRSTVWY

| | : | | | | : | | | | | | :
| | : | | | | : | | | | | | :

template

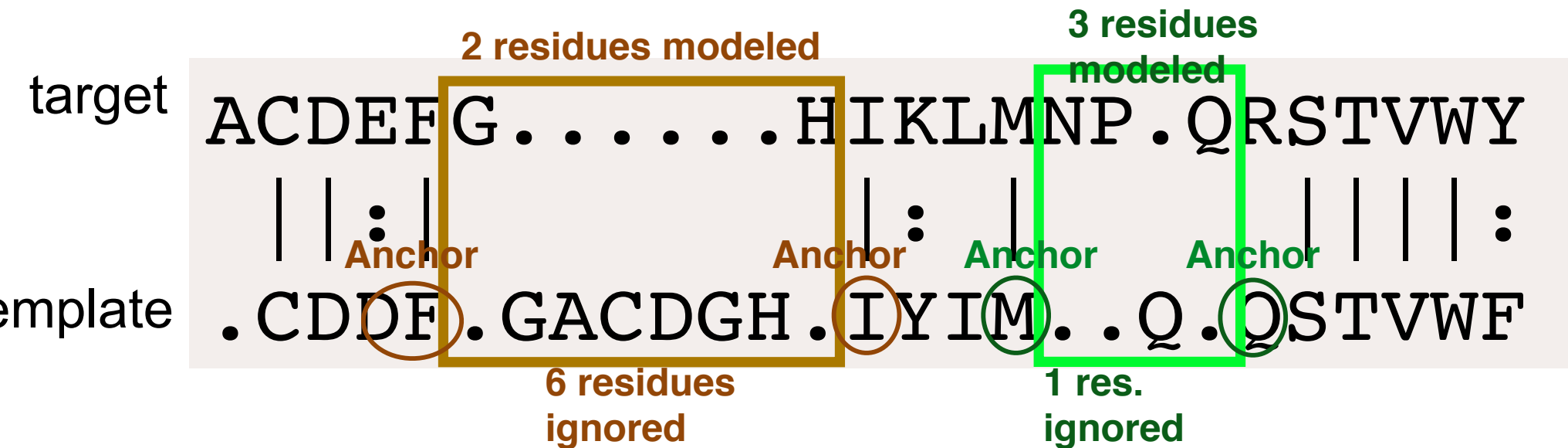
.CDDFGACDGHYIM . . QQSTVWF

MOE will align G to G, H to H, delete ACDG and attempt to connect G to H using energy minimization. No loop search is done.

MOE will perform a loop search for a two-residue insertion between positions M and Q. However, M and Q are adjacent, so no good loops will be found.

Both of these cases are non-optimal.

Optional expansion of loop search region (*n*-for-*m*)



"2-for-6" loop search

"3-for-1" loop search

Align F to F, I to I, delete GACDGH and add 2-residue loop GH from a loop search.

Align M to M, R to Q₍₂₎, delete Q and add a 3-residue loop NPQ from a loop search.

This approach general gives better results, especially if the indel is longer than 1 residue.

Exercise 19.1 -- Automated homology modeling

- At NCBI, search Protein database for **WP_006418556.1**
- Send to: File. FASTA. Hit Create file.
- Find the saved file and rename it "ex19.fa"
- In MOE:
 - Open ex19.fa
 - **Protein | Search | PDB**
 - Select the hit with the fewest gaps in the alignment. Load.
 - Load it.

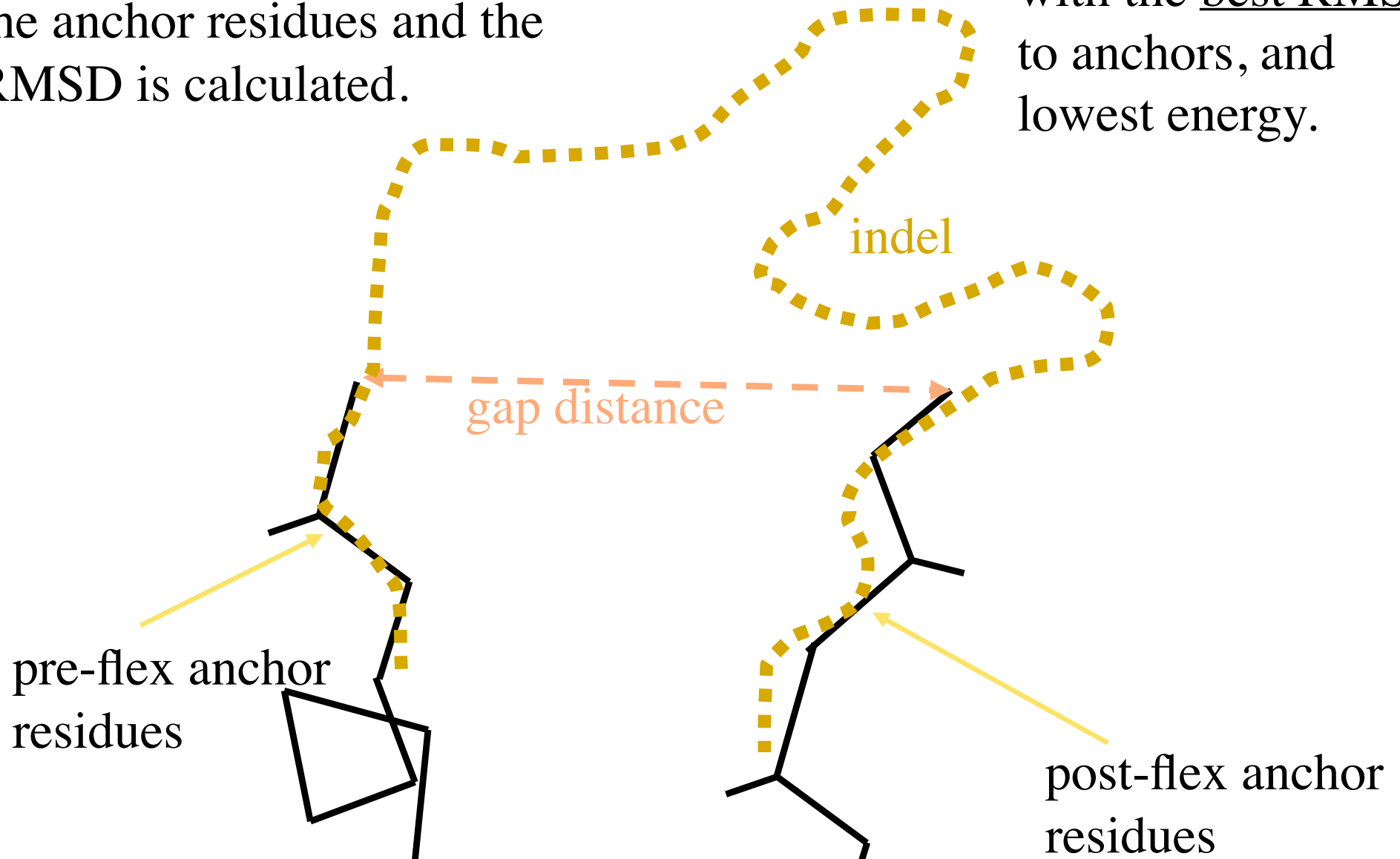
Exercise 19.1 -- p 2

- **Alignment | Align**
- **Protein | Homology model**
 - uncheck 'open database'. Check 'Load final'.
 - Select the template chain
 - **Edit | Potential | Inert**
 - **Edit | Potential | Fix**
 - Make the template atoms thin lines, light gray.
- Select new model. Rename the tag "model1"
 - Make atom ball-and-stick. Color by element
 - Optionally add silhouette. White background. Make model atoms stand out.
 - Adjust z-clip, fog, with middle-clicked atom centered.
- Walk-through. (stereo viewing really helps here...)
 - see **Walk Through** slide

Indels become Loop Searches

Loops of the right length in the database are superimposed on the anchor residues and the RMSD is calculated.

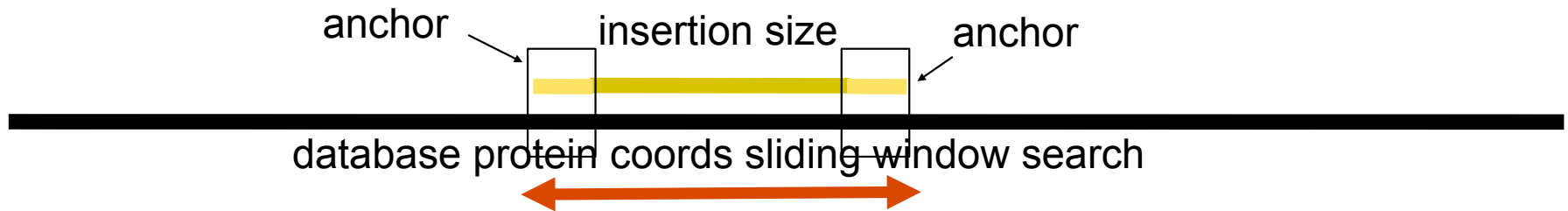
MOE keeps the loops with the best RMSDs to anchors, and lowest energy.



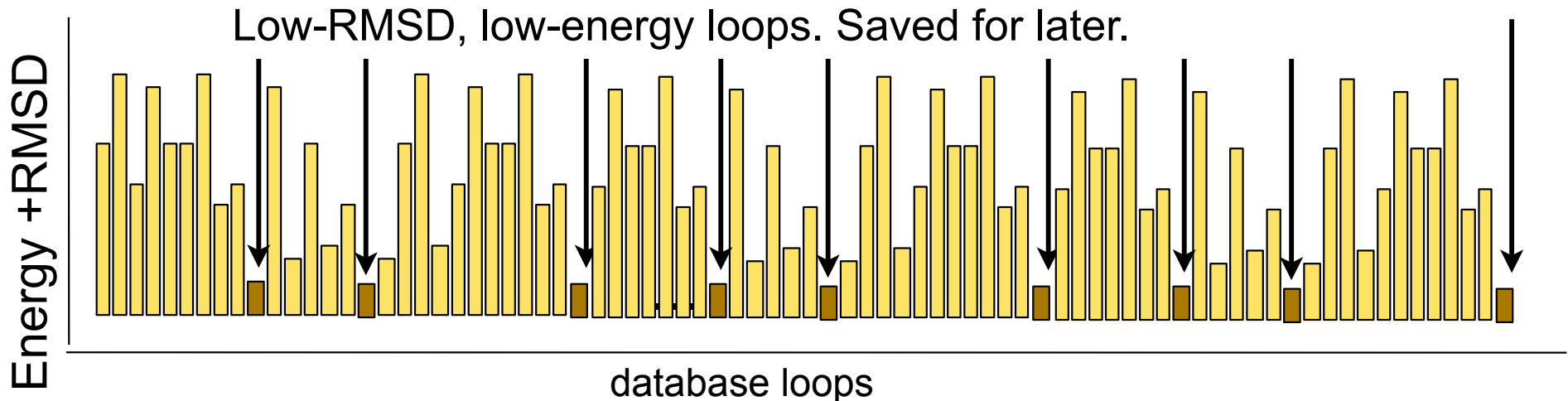
Automated loop search

database search for compatible anchors

PDB structures are scanned with "sliding window"

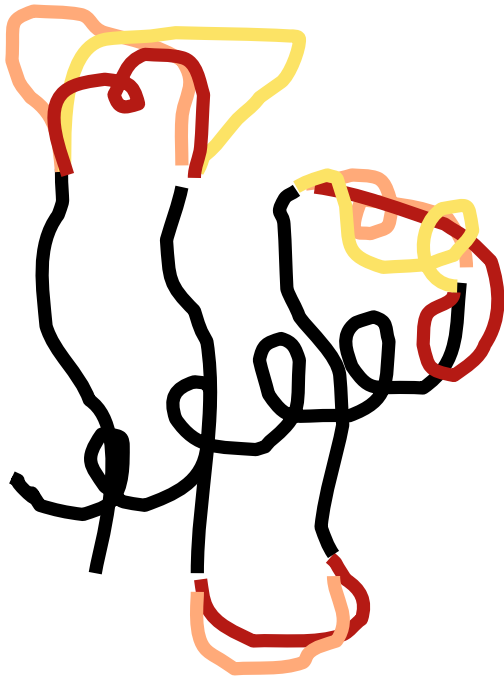


Anchor coords are superimposed on the database protein, and the RMSD is calculated. Low RMSD fragments are superimposed on the model to check the energy, after quick-n-dirty energy minimization.



Automated loop search

Finding multiple compatible loops.



Each indel elicits a loop search.

Loops from each search are swapped in randomly. Energy is minimized, and the result is recorded.

Swapping stops when energy stops improving.

Keeping an inert copy

Open another copy of the homology modeling template.

Select it.

Edit | Potential | Inert

Energy calculations will ignore inert atoms.

Display the atoms as thin lines, colored light gray, like a ghost.

Use your inert copy to see how far your model has deviated from the original template.

A walk-through



left-mouse select side chain atom
shift-middle-mouse drag to middle of screen.
shift-alt-middle-mouse wiggle, inspect :
H-bonds.
Hydrophobic packing.
Electrostatics.
Go to next residue, repeat,
until you have seen
every residue.

Walk-through

Hint: be very careful when you have the alt key depressed as this moves atoms.

In general, move one atom.

Watch how the model behaves when you wiggle an atom about this much. \longleftrightarrow Does it snap back? Does it settle somewhere else? Does it keep going back to the same place? In general, move just one atom.

Hint: If you move an atom too far, use the inert template as a guide to put it back where it started. \downarrow

inert atoms

Rules

- ALL BURIED H-BOND DONORS AND ACCEPTORS MUST BE PAIRED.
- ALL BURIED CHARGES MUST BE PAIRED.
- NO LARGE BURIED HYDROPHOBIC VOIDS.
- NO COLLISIONS OR OVER-PACKING.
- MOST SIDENCHAINS SHOULD BE IN GOOD ROTAMERS (SEE LATER LECTURE)

The best structure is
the one that breaks
the fewest rules.

Review

ACDEFGHIKLMNP . . . QRSTVY
ACDEFGHIKLMNPQRS . . . TVY

1. Where are the loop anchors?
2. What metric is used in the loop database search to judge the appropriateness of a loop?
3. What other metric(s) is/are used?
4. In the sequence editor, what keys are used to move a block of sequence?
5. What is the evolutionary meaning of an insertion?
6. What does MOE's homology model script do when the alignment has an insertion?
7. If the space between the anchors is too small for the insertion, what do you do?
8. If the space between the anchors is too large for a deletion, what do you do?
9. How likely is a deletion in the middle of a helix?
10. ...deletion in the middle of a sheet?
11. ...insertion in the middle of a helix?

Exercise 19.2 : Sequence editing practice

- Create a new sequence:
- SEQ: New / Text Editor
- Type ACDEFGHIKLMNPQRSTVY
- Save: aa.seq (close Text Editor window)
- Open: aa.seq (type: raw_seq) , twice -->

```
ACDEFGHIKLMNPQRSTVY
ACDEFGHIKLMNPQRSTVY
```

- Try “abacus” mode: **middle-mouse**

- *Make this alignment*—->

```
ACDEFGHIKLMNP . . . QRSTVY
ACDEFGHIKLMNPQRS . . . TVY
```

- Use “side” move: **shift-middle-mouse**

- *Make this alignment*—->

```
ACDEFGH . . . IKLMN . . . PQRSTVY
ACDEFGHIKL . . . MNPQR . . . STVY
```

- Use “slide” move **ctrl-middle-mouse**

- *Make this alignment*—->

```
ACDEFGH . . . IKLMN . . . PQRSTVY
ACDEFGHIKL . . . MNPQR . . . STVY
```

- Use “block” move. **left-mouse** select, **option-middle-mouse**

- *Make this alignment*—->

```
ACDEFGH . . . IK . . . LMNP . . . QRSTVY
ACDEFGHIKL . . . M . . . NPQRS . . . TVY
```

-

Recommended: do Homology Model Tutorial

Help/Tutorials/Homology Modeling of Proteins

Run the tutorial *verbatim*, exactly as stated, the first time.

Make sure you understand each step.

If you see something unexpected, go back.