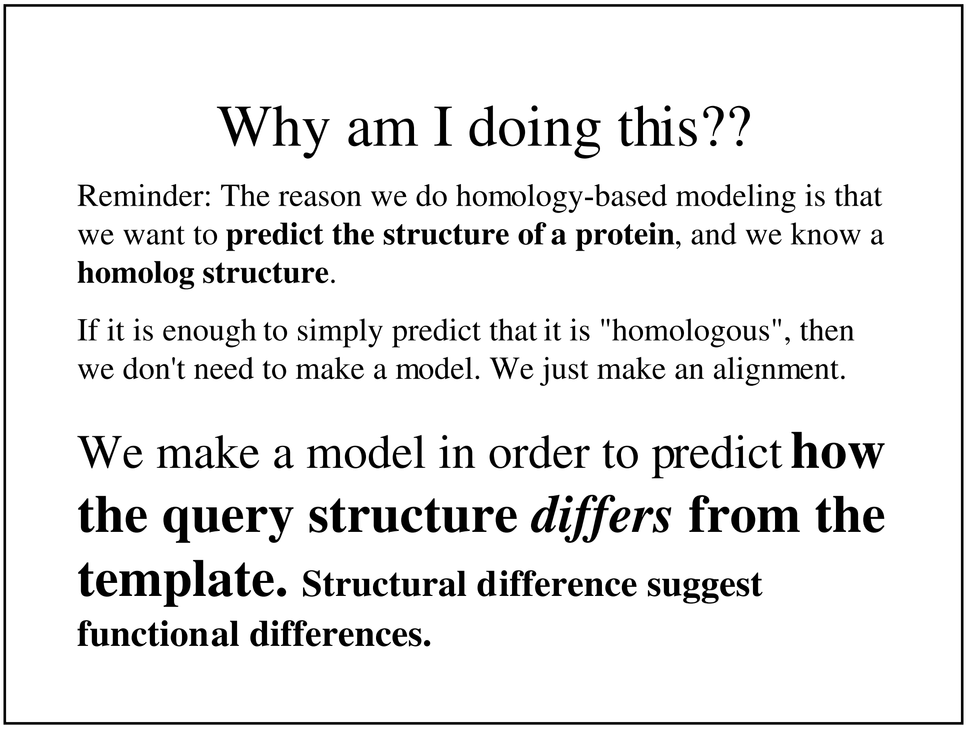
Bioinformatics 2 -- lecture 6

Loop modeling

Energy minimization

Steps in homology modeling

- Identify a sequence of interest.
- Search database for homologs of known structure.
- Align homologs with each other and with query.
- Add structural homologs, if necessary.
- Define SCRs and “Designated Loops”. Assign coordinates.
- Loop search or loop generate. Assign loop coordinates.
- End repair. Splice repair. Other repairs.
- Energy minimization. Analysis. Interpretation.

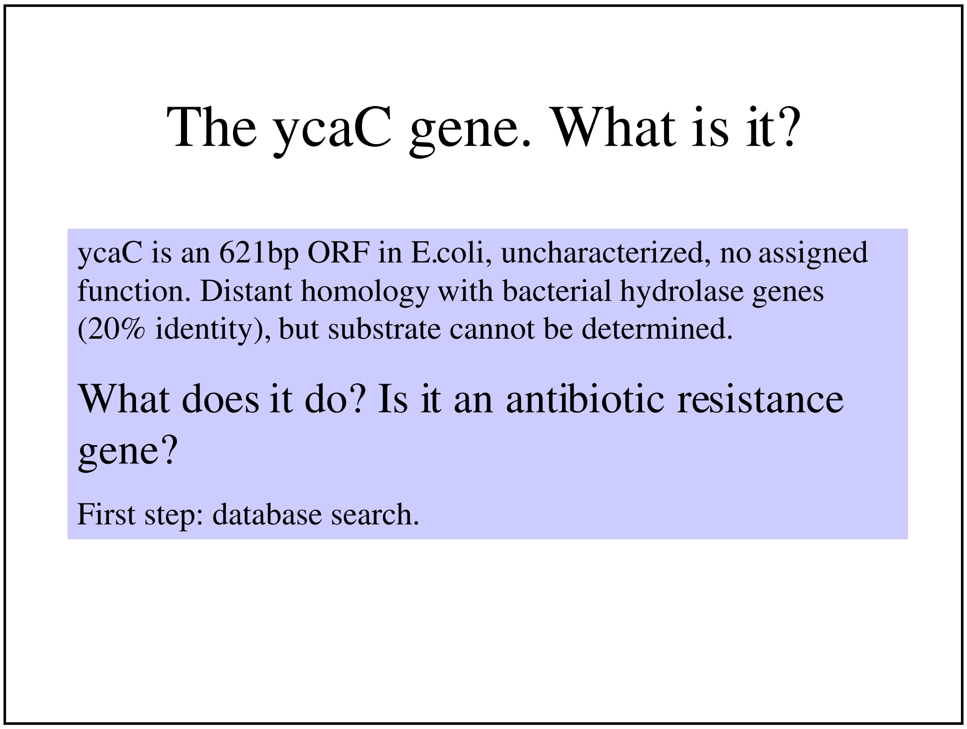


Why am I doing this??

Reminder: The reason we do homology-based modeling is that we want to **predict the structure of a protein**, and we know a **homolog structure**.

If it is enough to simply predict that it is "homologous", then we don't need to make a model. We just make an alignment.

We make a model in order to predict **how the query structure *differs* from the template. Structural difference suggest functional differences.**



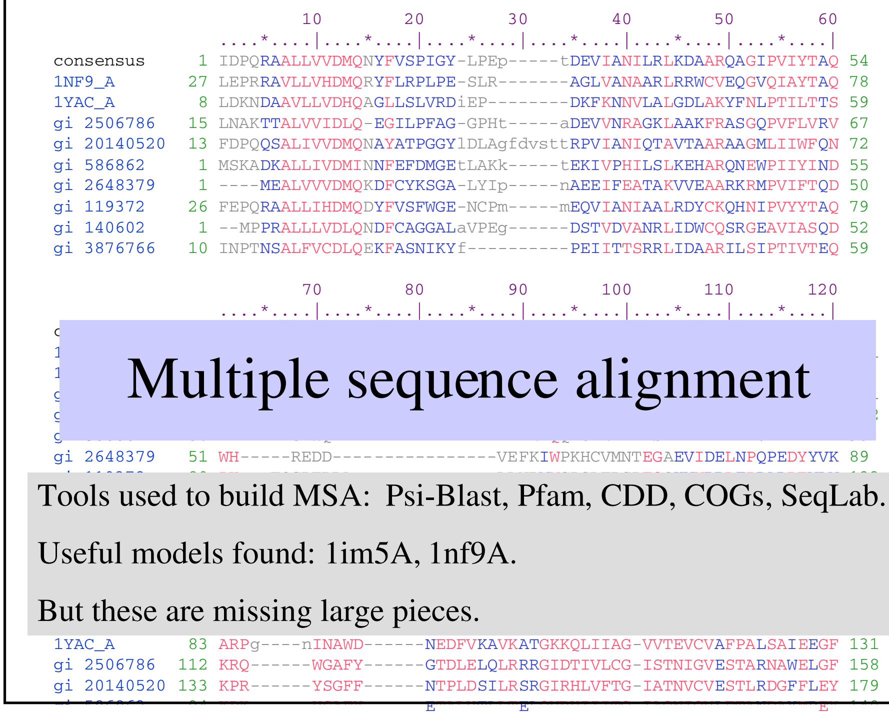
The ycaC gene. What is it?

ycaC is an 621bp ORF in E.coli, uncharacterized, no assigned function. Distant homology with bacterial hydrolase genes (20% identity), but substrate cannot be determined.

What does it do? Is it an antibiotic resistance gene?

First step: database search.

Multiple sequence alignment



```
                       10        20        30        40        50        60
                ....*....|....*....|....*....|....*....|....*....|....*....|
consensus     1 IDPQRAALLVVDMQNYFVSPIGY-LPEp-----tDEVIANILRLKDAARQAGIPVIYTAQ 54
1NF9_A       27 LEPRRAVLLVHDMQRYFLRPLPE-SLR-------AGLVANAARLRRWCVEQGVQIAYTAQ 78
1YAC_A        8 LDKNDAAVLLVDHQAGLLSLVRDiEP--------DKFKNNVLALGDLAKYFNLPTILTTS 59
gi 2506786   15 LNAKTTALVVIDLQ-EGILPFAG-GPHt-----aDEVVNRAGKLAAKFRASGQPVFLVRV 67
gi 20140520  13 FDPQQSALIVVDMQNAYATPGGYlDLAgfdvsttRPVIANIQTAVTAARAAGMLIIWFQN 72
gi 586862     1 MSKADKALLIVDMINNFEFDMGEtLAKk-----tEKIVPHILSLKEHARQNEWPIIYIND 55
gi 2648379    1 ----MEALVVVDMQKDFCYKSGA-LYIp-----nAEEIFEATAKVVEAARKRMPVIFTQD 50
gi 119372    26 FEPQRAALLIHDMQDYFVSFWGE-NCPm-----mEQVIANIAALRDYCKQHNIPVYYTAQ 79
gi 140602     1 --MPPRALLLVDLQNDFCAGGALaVPEg------DSTVDVANRLIDWCQSRGEAVIASQD 52
gi 3876766   10 INPTNSALFVCDLQEKFASNIKYf----------PEIITTSRRLIDAARILSIPTIVTEQ 59

                       70        80        90       100       110       120
                ....*....|....*....|....*....|....*....|....*....|....*....|
gi 2648379   51 WH-----REDD----------------VEFKIWPKHCVMNTEGAEVIDELNPQPEDYYVK 89
1YAC_A       83 ARPg----nINAWD------NEDFVKAVKATGKKQLIIAG-VVTEVCVAFPALSAIEEGF 131
gi 2506786  112 KRQ------WGAFY------GTDLELQLRRGIDTIVLCG-ISTNIGVESTARNAWELGF 158
gi 20140520 133 KPR------YSGFF------NTPLDSILRSRGIRHLVFTG-IATNVCVESTLRDGFFLEY 179
```

Tools used to build MSA: Psi-Blast, Pfam, CDD, COGs, SeqLab.

Useful models found: 1im5A, 1nf9A.

But these are missing large pieces.

Steps in homology modeling

√ Identify a sequence of interest.

√ Search database for homologs of known structure.

√ Align homologs with each other and with query.

- Add structural homologs, if necessary.
- Define SCRs and “Designated Loops”. Assign coordinates.
- Loop search or loop generate. Assign loop coordinates.
- End repair. Splice repair. Other repairs.
- Energy minimization. Analysis. Interpretation.

CE structural alignment

Since the basis set sequences have very low similarity, we want to collect some more structures of the same kind to ***enrich the basis set.*** Even distant homologs or structural homologs are better than building loops *from scratch*.

Insert CE web site demonstration here.
For those of you reading this at home, check out
cl.sdsc.edu/ce.html

Transferring an externally generated alignment into InsightII

Sadly, InsightII cannot import alignments, even of the most common formats. Instead, we must do the alignment from within. But InsightII's alignment tools are shoddy at best. Here's how to import an alignment from, say, a CE structural superposition:

(1) Load basis set structures and extract sequences in InsightII . Load query sequence.

(2) Open two windows on the same screen, one for the alignment (SeqLab, Netscape, whatever), one for the sequence window of InsightII. (If possible, show only the basis set sequences in SeqLab, etc.)

(3) For each basis set sequence, use **middle-mouse** to scroll sideways, **right-mouse** to insert gaps.

After the alignment.

We have already discussed the next two steps:

Defining the SCRs (these should be SSEs)

Assigning the "Designated Loops" (these are not necessarily *loops* in the secondary structure sense.)

Open and study **yac_assigned.psv**
Note the placement of the boxes in the structure by using
HOMOLOGY-->Sequences-->Color (and hit Color by: C-alpha)
Choose a color and a box (or an atom) and execute.
The associated structure will have that color.

When to use SCRs from multiple templates.

Normally it is best to pick SCRs from one template, the best template.

In this case, the best template (1nba) does not have a helix at position 70-76. And let's suppose it is *predicted to be a helix* by Psi-PRED, and there is a basis set member (1hso) that *has a helix* in the right place. So, we use it.

But most of the time, we would try to pick SCRs from **one template**.

Steps in homology modeling

√ Identify a sequence of interest.

√ Search database for homologs of known structure.

√ Align homologs with each other and with query.

√ Add structural homologs, if necessary.

√ Define SCRs and “Designated Loops”. Assign coordinates.

- Loop search or loop generate. Assign loop coordinates.
- End repair. Splice repair. Other repairs.
- Energy minimization. Analysis. Interpretation.

Loop Search

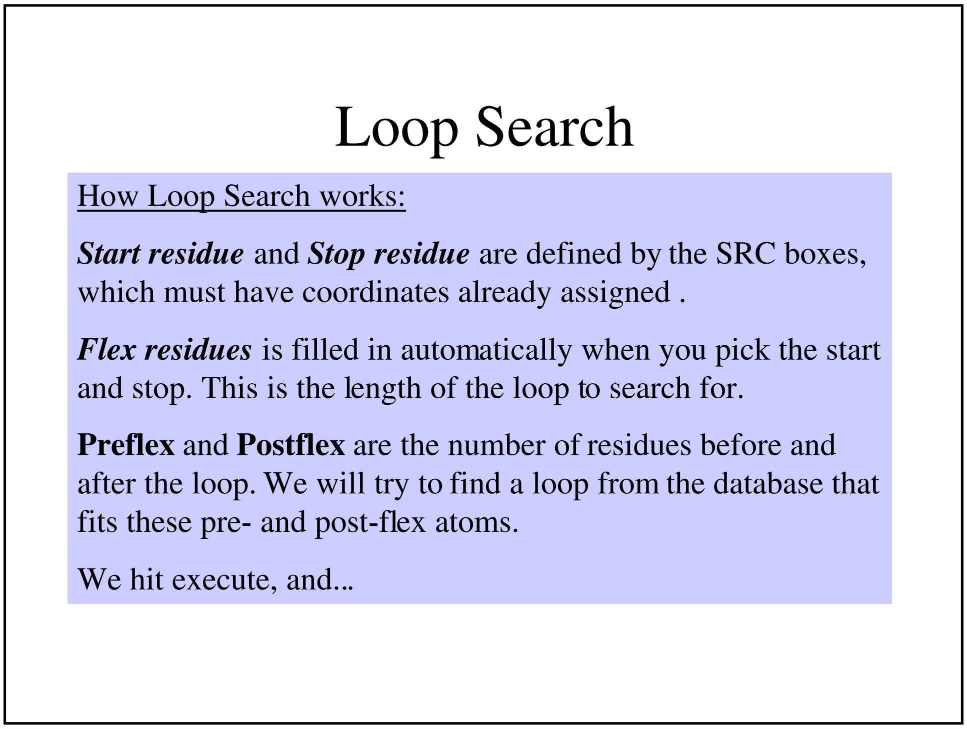
If any member of the basis set (not necessarily the primary source of your SCR boxes) has a loop of the right length, you should use it. Box is and assign coordinates as “**Designated Loop**” (to be explained later).

If no member of the basis set has a loop of the right length, then we try to find a loop of the right size in the database. Use

Homology-->Loops-->Search

Click on the two boxes bordering the loop. When you hit execute, the program will search the database.

(You cannot use this for the N and C termini! It will fail if the loop is very long, and it cannot be zero length.)



Loop Search

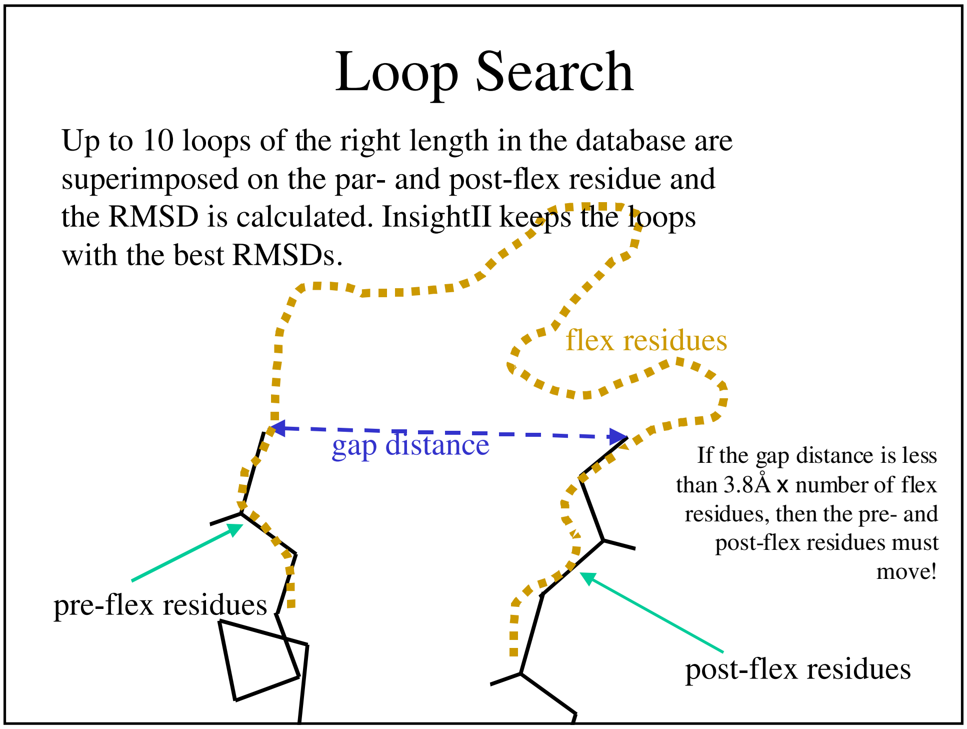
How Loop Search works:

Start residue and ***Stop residue*** are defined by the SRC boxes, which must have coordinates already assigned .

Flex residues is filled in automatically when you pick the start and stop. This is the length of the loop to search for.

Preflex and **Postflex** are the number of residues before and after the loop. We will try to find a loop from the database that fits these pre- and post-flex atoms.

We hit execute, and...



Loop Search

Up to 10 loops of the right length in the database are superimposed on the par- and post-flex residue and the RMSD is calculated. InsightII keeps the loops with the best RMSDs.

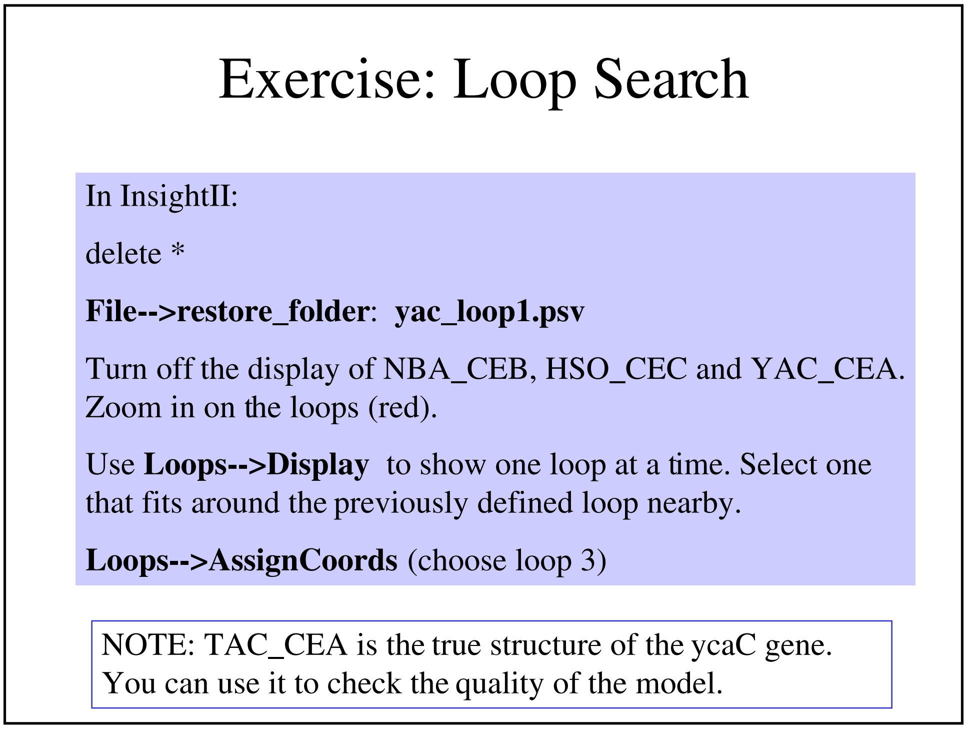
flex residues

gap distance

If the gap distance is less than 3.8Å x number of flex residues, then the pre- and post-flex residues must move!

pre-flex residues

post-flex residues



Exercise: Loop Search

In InsightII:

delete *

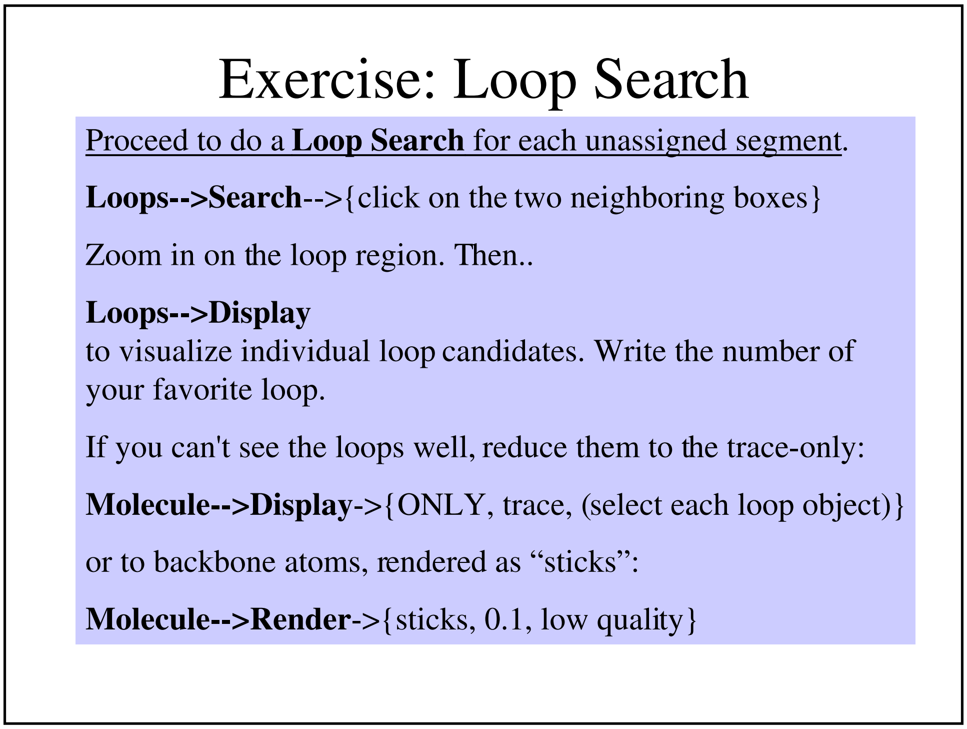
File-->restore_folder: **yac_loop1.psv**

Turn off the display of NBA_CEB, HSO_CEC and YAC_CEA. Zoom in on the loops (red).

Use **Loops-->Display** to show one loop at a time. Select one that fits around the previously defined loop nearby.

Loops-->AssignCoords (choose loop 3)

NOTE: TAC_CEA is the true structure of the ycaC gene. You can use it to check the quality of the model.



Exercise: Loop Search

Proceed to do a **Loop Search** for each unassigned segment.

Loops-->Search-->{click on the two neighboring boxes}

Zoom in on the loop region. Then..

Loops-->Display
to visualize individual loop candidates. Write the number of your favorite loop.

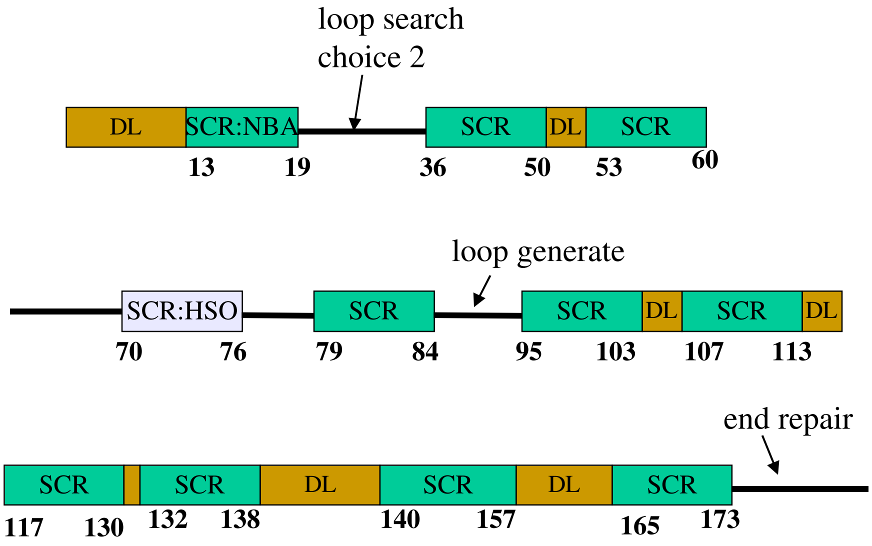
If you can't see the loops well, reduce them to the trace-only:

Molecule-->Display->{ONLY, trace, (select each loop object)}

or to backbone atoms, rendered as “sticks”:

Molecule-->Render->{sticks, 0.1, low quality}

Keep notes on the growing model



What makes a good loop?

•Aligns well with pre- and post-flex residues.

•Does not collide with other loops, or the backbone.

•Fills space. No big voids.

•Has polar sidechains *out*, non-polar sidechains *in*.

•Travels over non-polar surfaces (mostly). i.e. Does not bury charged residues.

•Has positive phi-angles at Glycines (mostly).

Loop Generate

Occassionally, InsightII cannot find a good loop in the database. A "last resort" option for building coordinates into a loop region is **Loops-->Generate**

It works just like Loops-->Search, but using a different algorithm: Levinthal's *random tweak method*.

Levinthal should have known better...

Loop quality and info level

Using more structural information improves the quality of loop prediction.

Method	Info source
Designated Loops	Global structural homology
Loop Search	Realistic local structure
Loop generate	Correct stereochemistry, only

Q: Why are randomly generated loops likely to be *worse* than database-derived loops?

Exercise: Loop Generate

In InsightII: Use your current model, or, optionally:

delete *

File-->restore_folder: **yac_loop2.psv**

Turn off the display of NBA_CEB, HSO_CEC and YAC_CEA.

Loops-->Generate

Loops-->Display (write the number of your favorite loop)

Loops-->AssignCoords

Loops-->assigncoords (choose loop 3)

Steps in homology modeling

√ Identify a sequence of interest.

√ Search database for homologs of known structure.

√ Align homologs with each other and with query.

√ Add structural homologs, if necessary.

√ Define SCRs and “Designated Loops”. Assign coordinates.

√ Loop search or loop generate. Assign loop coordinates.

- End repair. Splice repair. Other repairs.
- Energy minimization. Analysis. Interpretation.

Done modeling loops. End repair.

Homology:Refine-->End_repair

...is an automatic function. It fills in any missing coordinates at the N and C termini.

If there is a large piece of the structure missing in the alignment, we need external tools to model it. [to be discussed later]

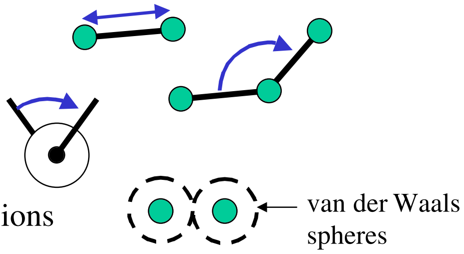
For today, we can ignore the C-terminal part. You may un-display it if you like.

constraint/restraint

Constrained energy minimization

Energy minimization using molecular mechanics *repairs* the following:

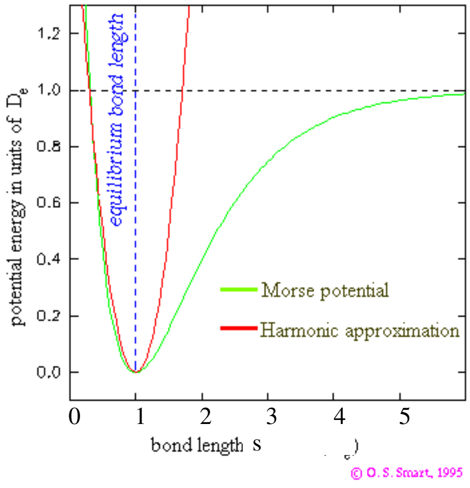
- bond lengths
- bond angles
- torsion angles
- non-bonded collisions



Constrained energy minimization: energy minimization in the context of constraints. `For example, in splice repair all atoms except the splice site atoms are` **`constrained`** `to their current positions. They are used in the energy calculations, but they cannot move.`

Harmonic potentials and Morse potentials

Harmonic and Morse potentials are **restraint** functions.



A force is applied to move the atoms to their ideal distances/angles.

See Orengo p. 129

Why do a *simulation*?

Why not just put every atom at its ideal position?
Can't we solve for this position using Calculus?

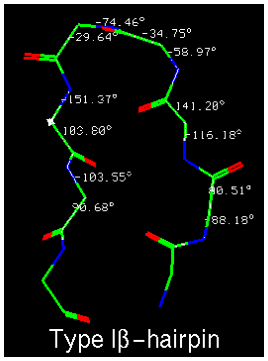
Yes, it is possible to build a molecule with exactly ideal bond lengths and angles. [Start at one end and connect one amino acid at a time]. But this would not produce the same 3D coordinates as a simulation.

Why? (see next slides)

Internal coordinates for proteins

		ϕ	ψ	ω	χ_1	χ_2
ALA	1~~~	0.000	127.140	180.000		
VAL	2~~~	-148.378	111.409	180.000	-179.551	
GLY	3~~~	-72.763	39.684	180.000		
HIS	4~~~	-73.084	122.882	180.000	-87.256	-62.962
THR	5~~~	-73.735	116.210	180.000	49.292	

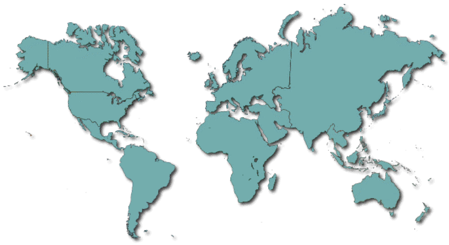
Ideal bond lengths, angles, plus torsion angles are enough to build the 3D structure.



Type Iβ-hairpin

The strange properties of internal coordinates when linked in a chain.

Cartographers before 1733 used internal coordinates



Errors in internal coordinates accumulate along the chain.

A global reference point (the stars and Harrison's perfect clock in 1733) removed the error accumulation problem.

Splice repair, a constrained simulation

The bonds between modeled segments (SCRs, designated loops, other loops), may be distorted.

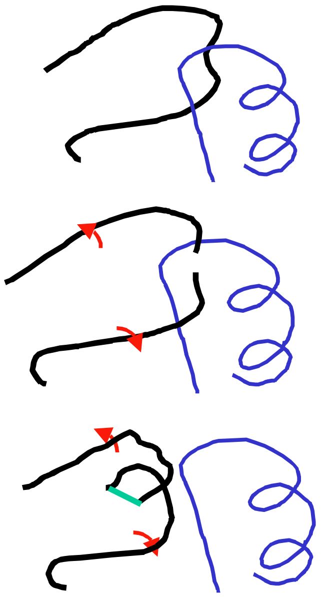
Refine-->Splice repair

...does a limited simulation to fix the stereochemistry around these splice points.

Run Splice repair and watch the stereochemistry fix itself.
Refine-->Splice_repair-->{add,all, (execute),
end_definition-->{steepest, 5., 100, (execute)}
end_definition-->{conjugate, 5., 100, (execute)}

If splice repair fails, manual repair may be necessary.

Manual repositioning



Crossing chains can never be repaired by energy minimization, since atoms would have to cross a very high energy barrier.

To move the backbone to a new position, we first make a cut, then rotate the backbone around selected **torsion angles**. Then repair the cut.

Exercise: manual repositioning using transform-->torsion

delete *

file-->restore_folder: yac_torsion.psv

Find the crossed loops. Cut peptide bond between 26, 27. Torsion around 20 (phi,psi) and 30 (psi,phi)

BIOPOLYMER:modify-->bond-->{break, select atoms}

Transform-->torsion-->{add,select 20:N,CA,C}
Do the same thng, adding torsions for **30:C,CA,N**. ***Be sure to select atoms in the direction of the movable part!!*** If you choose atoms C, CA, N, then the N-terminal side will move. If you choose N, CA, C, then the C-terminal side will move.

Move the chain using **middle-mouse**. Toggle torsions using **F7**. Move the chain to where it does cross any chains.

Exercise: manual repositioning using transform-->torsion

Put up a distance monitor. Try to make this distance short (<3Å)

Measure-->Distance-->{select C26 , N27}

When the loop is untangled, save the new torsion angles.

Transform-->torsion-->{clear, keep}

Re-create the broken bond.

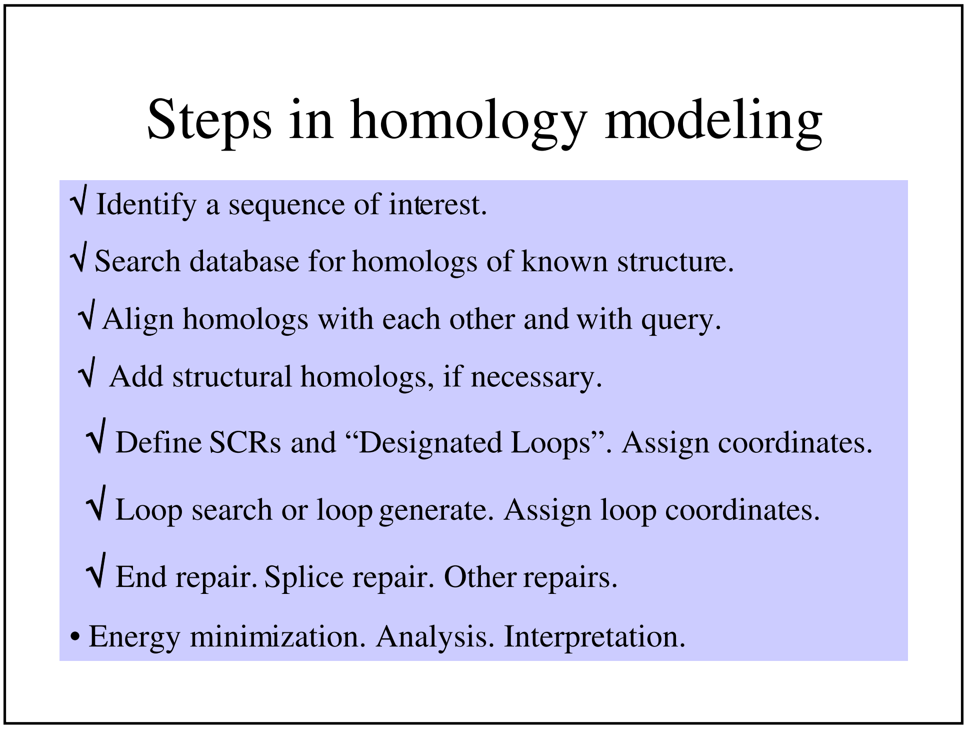
BIOPOLYMER:modify-->bond-->{create, select atoms}

Exercise: relaxation

Relaxation, as performed by Insight, is constrained energy minimization. You may choose the unconstrained atoms (the ones that are allowed to move), then InsightII will move these atoms subject to restraints and molecular mechanics forces.

(Van der Waals repulsion and torsion angles are molecular mechanics forces, bond lengths and bond angles are restraints.)

If necessary, delete * and restore yac_relax.psv
Refine-->relax-->{add,loop sides, (execute),
end_definition-->{steepest, 5., 100, (execute)}
end_definition-->{conjugate, 5., 100, (execute)}
Refine-->relax-->{add,loop backbone, (execute),
end_definition (as before). Then add SCR sides.



Steps in homology modeling

√ Identify a sequence of interest.

√ Search database for homologs of known structure.

√ Align homologs with each other and with query.

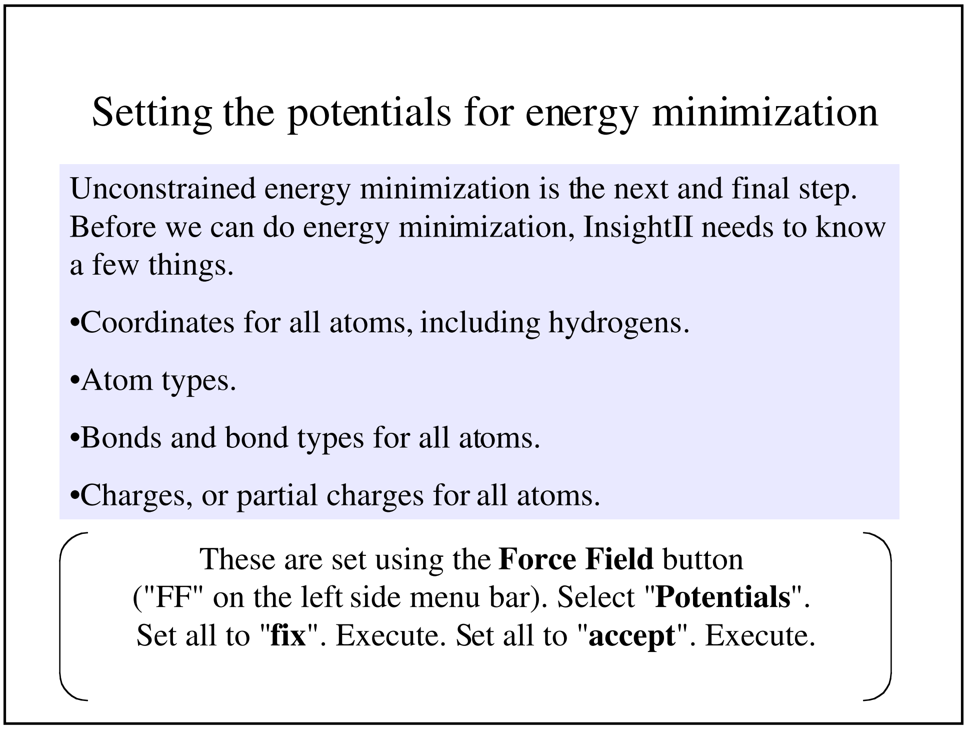
√ Add structural homologs, if necessary.

√ Define SCRs and “Designated Loops”. Assign coordinates.

√ Loop search or loop generate. Assign loop coordinates.

√ End repair. Splice repair. Other repairs.

- Energy minimization. Analysis. Interpretation.



Setting the potentials for energy minimization

Unconstrained energy minimization is the next and final step. Before we can do energy minimization, InsightII needs to know a few things.

•Coordinates for all atoms, including hydrogens.

•Atom types.

•Bonds and bond types for all atoms.

•Charges, or partial charges for all atoms.

These are set using the **Force Field** button ("FF" on the left side menu bar). Select "**Potentials**". Set all to "**fix**". Execute. Set all to "**accept**". Execute.

Energy minimization using Discover

Discover_3:Strategy-->Simple Minimize

Discover will not let you minimize unless the potentials have been set. You may have to add hydrogens (Biopolymer module) and/or set the potentials (FF menu).

Review

X-ray crystallography
    R-factor, resolution, Bragg's law, B-factor
NMR
    Ensemble, NOESY, TOCSY, spin system, distance geometry
PDB website
    How to find. How to download.
Rotation, superimposition, RMSD. Dali method. Contact maps.
Secondary structure propensity/prediction, GOR, Psi-Pred, Chou-Fasman, Q3 score, Ch, Ce score.
Protein structure classification (SCOP, CATH databases), class, fold, architecture, topology, analog, homolog.
Levinthal's Paradox. Anfinsen's thermodynamic hypothesis.
Molecular machanics, force field, simulation, constraint/restraint.
InsightII: basis set, structure alignment, SCR, designated loop, loop search, etc etc.
Spelunking. Secondary structure. H-bonds, torsion angles, sidechains (memorize them!). TOPS diagrams.