Things I could present at the lab meeting

* PyMOL tricks
* Preliminary gapsearch work: how many needed before you would expect to see 180 degree gaps?
  + Yes, but a long time before you see a 180 degree gap is *also* predicted by various models of correlation between structures. The only two options are not "depth dependence" and "depth dependence + spacing rule". There are also variants upon "depth dependence + correlation between structures". That is, if seeing one sequence gives us information about the next sequence, then you have more expectation of extreme values.
    - One kind of correlation is evolutionary intertia. The probability of Y(pos) in sequence B given Y(pos) in sequence A is (1-chance of mutation per year)^(years since divergence). Though I don't know the mutation rate, you could at least find the mutation rate required to give us at least a probability of X of Y(pos) in sequence B, and if it's crazy rapid, rule that out. But we don't know the time since divergence either.
    - Another kind of correlation is functional correlation. Maybe some residue is important in binding a signaling molecule. Then, *regardless* of the time since divergence, we expect to see the same residue at some position in sequence B if we learn that it has the same function as sequence A.
  + From Ez-β: the probability of an aromatic in a strand, given the depth dependenc emodel. From HHOMP: for a maximally diverse dataset of 8-stranded barrels, the number of barrels with 180 degree gaps.
* Homology modeling, mapping step: how are HHOMP alignments made?
* **Homology modeling, template step: how good of models do you get from two proteins in an HHOMP cluster?**
  + So my presentation would be: BBTM40 zdiff results. Gonnet zzdiff results. HHOMP alignment zdiff results.
* Loop length

Okay. Selecting template step, and mapping step because it's complementary. How do I *find* a pair of sequences of known structure?

* BLAST search, using all the hundred sequences of known structure? Can I do a BLAST search of the HHOMP database?
  + If I can, then I can check the number of high-scoring BLAST hits against the number of *annotated* structures in the HHOMP database. (there are 23)
* The sequences have labels. Are there annotations that also contain links to the structures?
  + HHOMP assembled its sequences from the NCBI nr database of bacterial protein sequences, and the sequences in the NCBI environmental database. What are these?
* Align sequence to one member of a cluster, and then compare sequence identity with each member of the cluster: number of alignment is clusternumber \* 100.
* BLAST each sequence in the HHOMP database against the PDB. Check for two hits.
  + that's a lot of fucking searches

Under what assumptions will this actually work?

* In the last three years,
  + someone has determined a structure of a protein that was already previously studied enough for HHOMP to find the sequence
    - this is certain. I have no doubt of this.
  + *furthermore*, this structure is in a cluster with another protein of known structure.
    - *maaybe*, but maybe not.
* In their removal of redundant PDB structures, the HHOMP people removed a structure that corresponds to a sequence in one of the other protein's clusters
  + Pretty sure of this. Suggests a different appraoch: search for similar PDB sequences. When you find unincluded PDB structure B that is similar to included PDB structure A, search HHOMP for a sequence with 95%+ identity with B.
    - Problem with that is, what I really want is two structures that are in the same cluster but have low sequence identity to each other. If I find a structure in the same cluster, but it has 90% sequence identity, that's pretty useless.

If I have a script to calculate pairwise sequence identity, I can

* Find pairs of similar structures from the OPM annotated β barrels (10,000 comparisons)
  + Will also need a script that will read them and compare each pair.
* Compare OPM annotated β barrels with each member of a cluster (50-1000 comparisons)
  + Will also need a script to read the barrel, read the cluster, and loop through the comparison.

Ideally:

* HHOMP locate all 106 PDB solved structures
* Do a pairwise identity check against each of the sequences of the cluster they've been matched to, *only if* that cluster contains a protein of known structure
* Find a sequence in the cluster with high seq identity to the structure (>90%) but low seq identity to the protein of known structure that is annotated in HHOMP (<30%)

Essentially this is an optimization idea.

Or, I can just write a seq id checker for a cluster, align sequences to clusters, and bam! One alignment per cluster per sequence, so like... oh geez. 100\*number of clusters.... thousands of alignments, each of which take like a minute.

~~Program of action.~~

1. ~~Find the clusters. Copy one to this folder.~~
2. ~~Realign it according to the Gonnet series matrices using ClustalW through Python.~~
3. ~~Find or download the OPM 106. Copy one into this folder.~~
4. ~~Append the sequence to the cluster using ClustalW through Python.~~
5. ~~Write a script that~~

Things to try.

1. Write a pairwise seq id comparison, using ClustalX by way of Python. See how long it takes to run one sequence against one cluster.
2. See if aligning to the wrong sequence in a cluster still gives appreciable seq id for identical sequences. If it does, then you don't need to align to clusters, you can just align to a single sequence *within* a cluster.

CHANGE OF PLANS.

Make templates using PDB structures from NEIGHBORING clusters.

1. If comparing two sequences within a clsuter works, then you are reasoning that *similar CLANS map distances have similar structure divergences*. When you look for a low seq id lower bound, you are reasoning that *closer CLANS map differences have closer structures.*
2. In that case, you can get a lower bound by comparing to a structure in a *neighboring cluster*. And besides, the cluster separations are fairly arbitrary. Sequences *within* that cluster are even closer to each other than the two you compared.

New program of action.

1. Rewrite the zdiff calculator
2. Make (partial) list of nearby pairs of structurees on the HHOMP cluster map
3. Make Swiss-PDBViewer ideal alignments for nearby pairs
4. Make max seq id calculator - gotta learn clustalw command line. Or do it by hand.
5. Align downloaded PDB structures to structures that Daniel aligned
6. Modify zdiff calculator so that it can map bulk input to bulk output of zdiff reports and structures with b-factors and seq identity
   1. make the b-factors a really narrow range - like, 3. Or 1. So that instead of having to get average zdiff over the right selections you can look at it.
7. Run the pairs through the script with the Swiss-PDBViewer alignments
   1. glob and itertools.combinations to make the filename pairs
8. Pratice talk, write script of talk
9. Practice talk w/ dummy slides
10. Make figures
11. Go to SLEEP!