An e-mail I ended up never sending to Daniel.

Hi,

Is everything all right with the alignments? I'm working on a couple improvements right now.

1. Removing sequences distant from the template

There are sequences in the clusters with that have as many as 96% of positions different from the template. The zdiff results indicate the possibility of fairly good performance all the way down to 13% identity, but 4% is extreme. Vik recommends removing all sequences that have less than 15% identity with the template from the final alignment. This is easy to do and seems like a definite benefit, and I will have it done very soon.

Then there are some other things that I may or may not do. Vik thinks that the zdiff results indicate that the alignments are more or less good, so now I'm in a tradeoff situation where I can spend my time improving the alignments, or doing science *with*the alignments. But here're the things I'm considering.

2. BBTM matrices

I'm still not sure what the strongest differences between the BBTM matrices and the Gonnet matrices are, and if we'll really see significant improvements by trading good alignment of loops for good alignment of strands. Worth a try perhaps.

3. Align with Praline

Praline will use different substitution matrices depending upon the predicted secondary structure. By default it uses PHAT, not BBTM, but honestly I doubt that PHAT and BBTM are that different: something to check up upon.

4. Align with a hidden Markov model based program

Like Praline, this will probably acknowledge the differences between loop and strand. They'll end up represented as two distinct groups of states in the Markov chain, I think.

5. Restore missing loops

In the alignments that you have, the sequence of the template is the one extracted from the PDB structure files that we use. I'm guessing your backend requires this - how else will it know what positions to substitute out?

However, it would be better if when I'm *making* the alignments, I include the *full* sequence of the template structure, including any loops they couldn't get any decent structural data on. That way the loops can help the strands get lined up.

This one would take me longer than the above two. I'm not entirely sure it's worth the time, and once I have a better idea of what it's going to require I'm going to ask Vik if he thinks I should still go through with it. I'm thinking yes, though: especially since we're using the Gonnet matrix, which I would expect to do better on loops than strands, it seems like this could help a lot. But, looking at the zdiff results, the one with lots of missing loops (3EFM) didn't do significantly worse as a template than the other two of equivalent sequence distance from their respective targets. So maybe it doesn't really matter that much.

Best,

Alex

Why would I use BBTM if I can already use PHAT without the work? (bbecause BBTM will align the aromatic girdles better)

Will missing loops screw up the HMM? Depends - does it end up recognizing the upward-strand long-loop downward-strand short-loop pattern? If it does it would mess it up a *lot*. This isn't hard to test. Get the structural alignment, then do the HMM alignment with and without the loops. Count the number of pairs that are the same.