

Lab 5: Protein Structure Prediction

Zackery Field

March 6, 2014

Part 0¹: Examine [sequence record](#) for PBP1A_MYCTU

The header of the sequence file reveals that the record is in the manually curated SwissProt section: "P71707 (PBP1A_MYCTU) Reviewed, UniProtKB/**Swiss-Prot**".

0.1 TMHMM comparison

TMHMM and Uniprot both predict a helical TM region at residues 139-159.²

[UniProt](#): Transmembrane - Helical 139-159

[TMHMM](#): Transmembrane - Helical 137-159

0.2 Pfam comparison

The Pfam-A matches discovered using the [Sequence Search](#) tool shows that there is agreement between the Pfam and Uniprot records. Both records place a [transglycosylase](#) at region 180-360 (aa). Both records also show a [transpeptidase](#) at region 453-734 (aa).

Part 1: [Phyre-results](#)

Part 2: Run BLAST vs PDB at NCBI

Top Hit: [3DWK_A](#)

The Pfam analysis of both the top hit and the target are in agreement. Both Pfam analyses describe approximately the same MDA. That is to say both a [transglycosylase](#) and a [transpeptidase](#) are identified, but 3DWK_A does not have an N terminal region extending from the transglycosylase, whereas p71707 does.

The transglycosylase region of p71707 is from residues 180-360. The transglycosylase region of 3DWK_A is from residues 15-191.

¹Nice use of zero-indexing.

²It is possible that the results from TMHMM/Pfam and Uniprot are related, because the Uniprot record could have been populated with data from both Pfam and TMHMM.

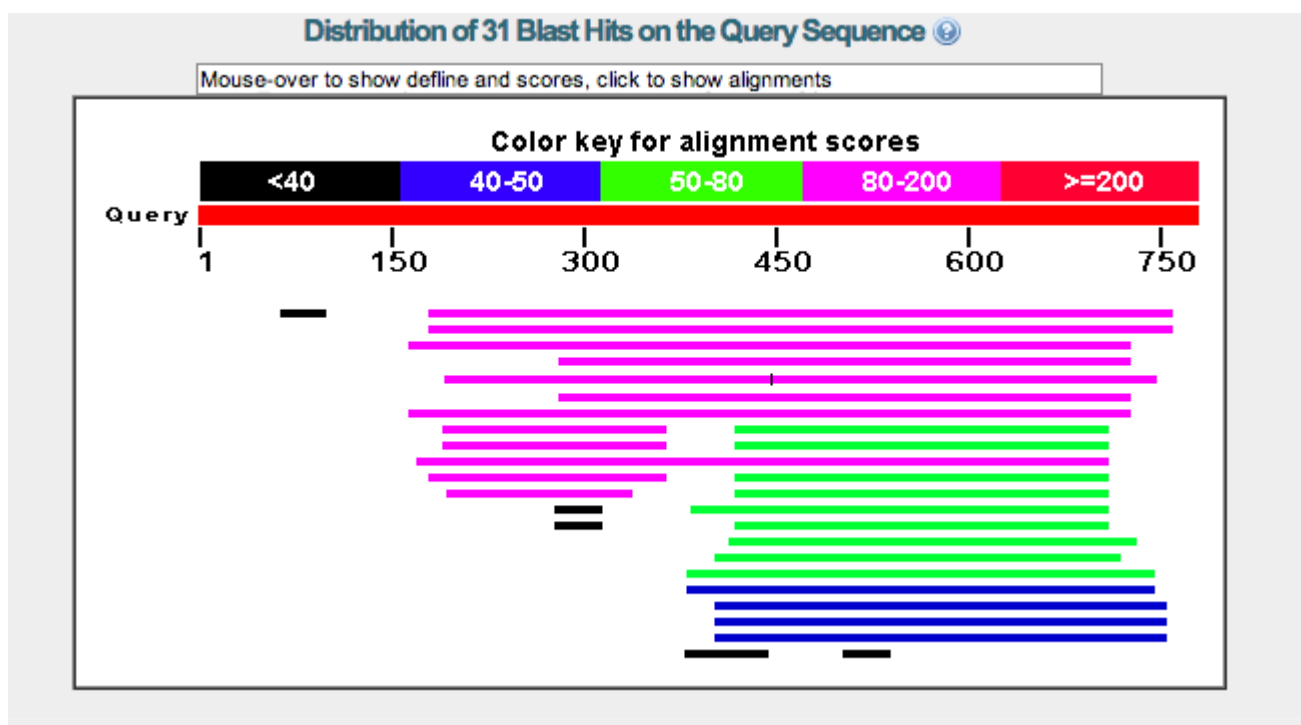


Figure 1: A screenshot of the alignment overlap display from a blast query of p71707 run against the pdb database.

The BLAST pairwise alignment for those regions:

Query	182	EIAKIVPEGNRVDVNL	SQVPMHVRQAVIAA	EDRNFYSNPGFSFTG	FARAVKNNLFGG-D	240
		E+ K +	VNL VP ++ AV+A	ED FY + +	A+ NL GG	
Sbjct	16	ELVKTL	DNGQRHEHVN	LKDVPKSMKDAVL	ATEDNRFYEHGALD	YKRLFGAIGKNTGGFG 75
Query	241	LQGGSTITQYVKNAL	VGSAQHGW	SGLMRKAKELVIAT	KMSGIEWSKDDVL	QAYLNIIYFG 300
		+G ST+TQQ VK+A +	+QH G RKA+E ++ ++	E+SKDD+ Q YLN	IY+	
Sbjct	76	SEGASTLTQQVVKDAFL--	SQHKSIG--RKAQEAYLS	YRLEQEYSKDDIFQVY	LNKIYYS	131
Query	301	RGAYGISAASKAYFDK	PVEQLTVAEGALLA	LIRRPSTLDP	PAVDPEGAHARWN	VLDGMV 360
		G GI AA+K YF+K ++ L	+AE A LA L + P+ +	P+ A R N VL M		
Sbjct	132	DGVTGIKAAAKYYFNKDL	KDLNLAEAYLAGLPQ	VPNNYNIYDHPKAAED	RKNTVLYLMH	191

A BLAST alignment was done on these two regions in particular and the results were:
 Max Score 93.6 Total Score: 93.6 Query cover: 97% E-value: 1e-28 Ident: 33%.

Since the sequence identity is on the upper edge of the twilight zone of protein sequence

alignments (20-35%PID) ³, with high sequence coverage it is possible that these two regions do not represent the same transglycosylase domain. However, as the twilight paper states, more than 95% of sequences above 30% PID were homologous. Therefore, I would agree with the Pfam classification that both of these regions contain a domain from the transglycosylase family.

The transpeptidase region of p71707 is from residues 453-743. The transpeptidase region of 3DWK_A is from residues 294-562.

The BLAST pairwise alignment for those regions:

```

Query 453  VVSIDPHNGAVRAYYG 469
           +D  G + A  G
Sbjct 294  ATILDSKTGGLVAISG 310

Query 470  G-DNANGFDFAQAG--LQTGSSFKVFA----LVAALEQGIGLGYQVDSSPLTVDGIKITN 522
           G D + + QA      TGSS K F      + ++      Q D S  VDG  N
Sbjct 311  GRDFKDVVNRNQATDPHPTGSSSLKPFLAYGPAIENMKWATNHAIQ-DESSYQVDGSTFRN 369

Query 523  VEGEGCGTCNIAEALKMSLNTSYRML--LKLNGGPQAVADAHQAGIASSFPQVAHTLS 580
           + +  GT +I +AL+ S N      +      +K N G  A      A + G      L+
Sbjct 370  YDTKSHGTVSIYDALRQSFNIPALKAWQSVKQVQAGNDAPKKFAAKLG-----LN 418

Query 581  EDGKGGPPNNGIVLGQYQTRV--IDMASAYATLAASGIYHPPHFVQKVVSANGQVLFDDAS 638
           +G  GP      VLG  +      +ASA+A +A  G Y+  H +QKVV+ +G+ +
Sbjct 419  YEGDIGPSE---VLGGSASEFSPTQLASAFAAIANGGTYNNAHSIQKVVTTRDGETIEYDH 475

Query 639  TADNTGDQRIPKAVADNVTAAMEPIAGYSRGNLAGGRDSAAKTGTTQFGDTT----- 691
           T+      +A+ +      +P  G + GH ++ G +  AKTGT  +G  T
Sbjct 476  TSHKAMSDYTAYMLAEMLKGTFKPY-GSAYGHGVS-GVNMGAKTGTGTGTYGAETYSQYNLP 533

Query 692  --ANKDAWMVGYPSTLSTAVWVGTVK----GDEPLVTASGAAIYGSGPLSDIWKATMDGA 745
           A KD W+ G+TP  + +VW+G  K      G+  V  S      P  +++  M
Sbjct 534  DNAAKDVWINGFTPQYTMSVWVGFSKVKQYGENSFVGHSSQQE-----YPQFLYENVMSK 587

```

A BLAST alignment was done on these two regions in particular and the results were: Max Score 69.3 Total Score: 69.3 Query cover: 79% E-value: 3e-18 Ident: 29%.

Since the sequence identity is well within the twilight zone of protein sequence alignments (20-35%PID) it is possible that these two regions do not represent the same transpeptidase domain. Since this 29% identity is only over 80% of the sequence, the PID is actually lower. As the twilight paper states, more than 90% of sequences below 25% PID were not homologous. I would not agree with the Pfam classification that both of these regions contain

³[Twilight zone of protein sequence alignments](#)

a transpeptidase domain. The E-value for the transpeptidase prediction for p71707 ($9.1e-50$) is 18 orders of magnitude lower than the E-value for the transpeptidase prediction for 3DWK_A ($1.5e-32$). Therefore, the transpeptidase classification for p71707 should remain, and the transpeptidase classification for 3DWK_A is more likely to be incorrect.

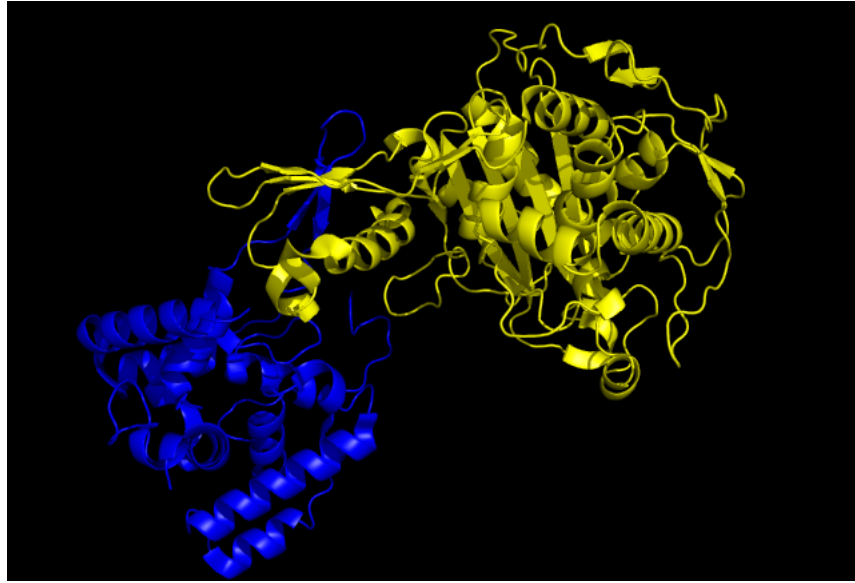


Figure 2: An image of the 3d structure of 3DWK_A with the Pfam identified domains highlighted. Blue: transglycosylase; Yellow:transpeptidase