

Identification of possible functionality of an "unknown" protein sequence

Abhishek Dasgupta (06MS07)
Sambit Bikas Pal (06MS03)

Indian Institute of Science Education and Research, Kolkata

April 19, 2008

What we've done: in a nutshell

- ▶ Picked a protein from SWISSPROT (ideally this would be an unknown protein sequence; however without access to an unknown sequence, we took a known one.)

What we've done: in a nutshell

- ▶ Picked a protein from SWISSPROT (ideally this would be an unknown protein sequence; however without access to an unknown sequence, we took a known one.)
- ▶ Did a sequence search on PFAM.

What we've done: in a nutshell

- ▶ Picked a protein from SWISSPROT (ideally this would be an unknown protein sequence; however without access to an unknown sequence, we took a known one.)
- ▶ Did a sequence search on PFAM.
- ▶ Found a match! Topoisomerase IV (PF00521)
(what are topoisomerases?)

What we've done: in a nutshell

- ▶ Picked a protein from SWISSPROT (ideally this would be an unknown protein sequence; however without access to an unknown sequence, we took a known one.)
- ▶ Did a sequence search on PFAM.
- ▶ Found a match! Topoisomerase IV (PF00521) (what are topoisomerases?)
- ▶ Did CLUSTALW alignment with the PFAM **seed**. Identified conserved residues.

What we've done: in a nutshell

- ▶ Picked a protein from SWISSPROT (ideally this would be an unknown protein sequence; however without access to an unknown sequence, we took a known one.)
- ▶ Did a sequence search on PFAM.
- ▶ Found a match! Topoisomerase IV (PF00521) (what are topoisomerases?)
- ▶ Did CLUSTALW alignment with the PFAM **seed**. Identified conserved residues.
- ▶ Did a search on Phyre server and found 1zvu as the best match.

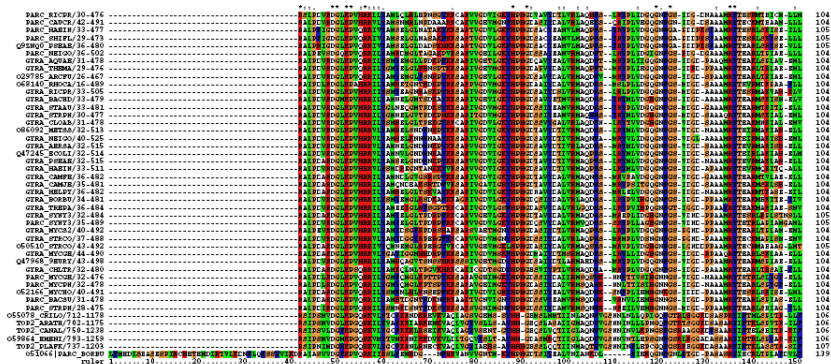
What we've done: in a nutshell

- ▶ Picked a protein from SWISSPROT (ideally this would be an unknown protein sequence; however without access to an unknown sequence, we took a known one.)
- ▶ Did a sequence search on PFAM.
- ▶ Found a match! Topoisomerase IV (PF00521) (what are topoisomerases?)
- ▶ Did CLUSTALW alignment with the PFAM **seed**. Identified conserved residues.
- ▶ Did a search on Phyre server and found 1zvu as the best match.
- ▶ Read the structure paper → important residues → are they there?

CLUSTAL X (1.83) MULTIPLE SEQUENCE ALIGNMENT

File: /home/sambit/phylo/proj/sample.ps
Page 1 of 5

Date: Sun Apr 13 21:51:25 2008




```

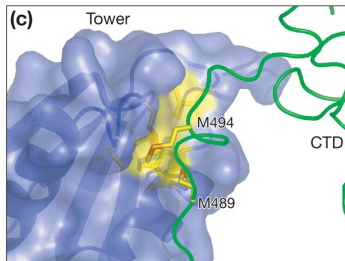
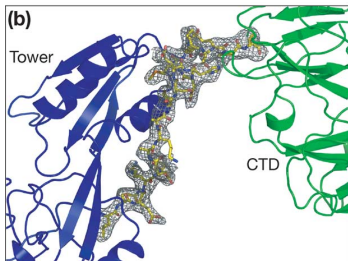
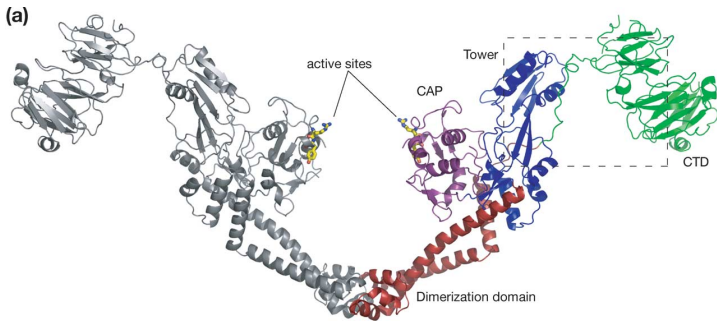
PARC_RICPR/30-476 EDIDNDTIDFSTDDSDLEPVMGASFNLLANGS82I
PARC_CAUCR/42-491 DGIDEDANDFPTDGGQDEEPVVLSEFNLLANGS82I
PARC_HAEIN/33-477 HELDQCTIDFCPNDDCTHAEPQGLARLPHILLINGTTCI
PARC_SHIFL/29-473 SELDQCTAIDFVNDGCTQEDPMLGAEPLNLLINGTTCI
Q98N00_PSEAE/36-480 SELDQCTIDFVNDGCTIDEBAVLGAEPLNLLINGTTCI
PARC_NEIOO/36-502 SEINQCTIDFVNDGCLDEDELLGLAEPLNLLINGAS2I
GTRA_AQUAE/31-478 TDIDNDTIDFCPNDDTLEPEVVLSEFNLLNGT82I
GTRA_THEMA/29-476 EDIEENTMFCNDDGCTLEPEVVLSEFNLLINGAS2I
O29785_ARCFU/26-467 ADICENTIDFVNDGATLEPEVVLSEFNLLINGS82I
O68140_RHCCA/16-489 ADIEEDTIDFCNDDGDEPTVLEAFNLLINGAG2I
GTRA_RICPR/33-505 EDIDNDTIDFVNDGSEEEPSVLEAFNLLINGAS2I
GTRA_BACHD/33-479 NDINNDTIDFCNDDGSEEPVVMSEFNLLINGAS2I
GTRA_STAAU/33-481 NDINNDTIDFCNDDGNEREPVLEAFNLLINGAS2I
GTRA_STRPH/30-477 NDINNDTIDFCNDDANEREPVLEAFNLLINGAT2I
GTRA_CLOAB/31-478 NDICENTIDFVNDGSEEPVVLSEFNLLINGAS2I
O86092_METSS/32-513 ADIDNETIDFVNDGSEEPDLIMAFNLLINGS82I
GTRA_NEIOO/40-525 ADIEETMFCNDDGSEEPVLEAFNLLINGS82I
GTRA_AERSA/32-515 ADLENETIDFVNDGCTEMIDAMTFNLLINGS82I
Q47245_BOOLI/32-514 ADLENETIDFVNDGCTEIPDMTFNLLINGS82I
GTRA_PSEAE/32-515 ADLENETIDFVNDGCTEQIDAMTFNLLINGS82I
GTRA_HAEIN/33-511 TDLDNETMFCNDDGCLMIDVLETFDALLINGS82I
GTRA_CAMPE/36-482 EDLDNDTIDFVNDSDSLEPDLVLEAFNLLINGS82I
GTRA_CAMPE/35-481 NDIDNDTIDFVNDGSEESDVLSEFNLLINGS82I
GTRA_HELBY/36-482 NDIDNDTIDFVNDGTLREDDILESLNLLINGAN2I
GTRA_BORBU/34-481 NDIDNETMFCNDDSDSLEPEIMSEFNLLINGS82I
GTRA_TREPA/36-484 EDIEKETVSEFNDDSDVEPTVLECFNLLINGS82I
GTRA_SYNT3/32-484 NDIEAETIDFCNDDGSEQEPTVLEAFNLLINGS82I
PARC_SYNT3/35-489 EGISSEAVIDFCNDDNSQREPTVLEAFNLLINGC82I
GTRA_MYCS2/40-492 NDIEDEETIDFCNDDGSEQEPTVLEAFNLLINGS82I
GTRA_STRCO/37-488 NDIEDEETIDFCNDDGSEQEPTVLEAFNLLINGS82I
O50510_STRCO/43-492 ESIDEDTIDFVNDGQEQEPVALLAFNLLINGAS2I
GTRA_MYCOE/44-490 NDIDNDTIDFVNDGSEEPVLEAFNLLINGS82I
Q47968_9EURY/43-498 DDIDNDTIDFCNDDGSEQEPTVLESEFNLLINGS82I
GTRA_CHLFR/32-480 EDLDNDTIDFVNDGDETELEPVMSEFNLLINGS82I
PARC_MYCOE/32-476 NDIDKLVSEFNDDSEEPVLEAFNLLINGAS2I
PARC_MYCPH/32-478 NDIEQLVSEFNDDSEEPSVLEAFNLLINGT82I
O52166_MYCOE/40-491 NDLDSEVSEFNDDSEEPVLEAFNLLINGA2I
PARC_BACBU/31-478 NDIDNDTIDFVNDGDDTSEEPVLEAFNLLINGET2I
PARC_STRPH/28-475 QDIEKETVSEFNDDTSEEPVLEAFNLLINGET2I
O55078_CRILO/712-1178 FPEDDDTLIFL-EDNQVEPEVIFIPMVLINGAB2I
TOP2_ARATH/702-1175 FDDDLILIN-EDGQIEPTVLEAFNLLINGAB2I
TOP2_CANAL/750-1238 NFDLDFLITVFC-DEEQTVEPEVIFIPMVLINGAB2I
O59864_EMENI/793-1259 FPHDDPILVLE-EDGSEEPVIFIPMVLINGAD2I
TOP2_DLAFL/737-1203 NEFDLDFLILIN-EBGQIEPTVLEAFNLLINGCB2I
O51066|PARC_BORBU S---KETTIESSDDCHNNEPLLGAEPVILLIQSB2I
ruler .....160.....170.....180.....15

```



From the structure paper for 1ZVT (doi:10.1016/j.jmb.2005.06.029) we see that topo IV is mainly encoded by **parC**, **parE**
parC comprises two domains:

- ▶ (28-158) helix turn helix motif similar to CAP; contains active sites needed for DNA cleavage: Arg119 and Tyr120. Arg119 and Tyr120 was found **exactly** conserved across the test sequences (taken from Pfam seed PF00521) and our unknown protein.
- ▶ (159-340) **tower** packs against CAP domain; structural support.
- ▶ also there's a compact α -helical bundle connected by long α -helices to the tower domain and the C-terminal domain.



Knowing the unknown

So what is the unknown protein?

- ▶ It is most probably a **topoisomerase IV**. It contains the Arg119 and Trp120 residues as mentioned earlier.
- ▶ The topoisomerase portion starts only from sequence 42, as is evidenced by the lack of alignments before that residue.

URLs

- ▶ www.expasy.ch/sprot/
- ▶ www.rcsb.org/
- ▶ pfam.sanger.ac.uk/
- ▶ www.sbg.bio.ic.ac.uk/phyre/
- ▶ www.ebi.ac.uk/clustalw/

Thanks!

We are especially grateful to Dr. Rana Bhadra without whose help this project would not have been realised.