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.

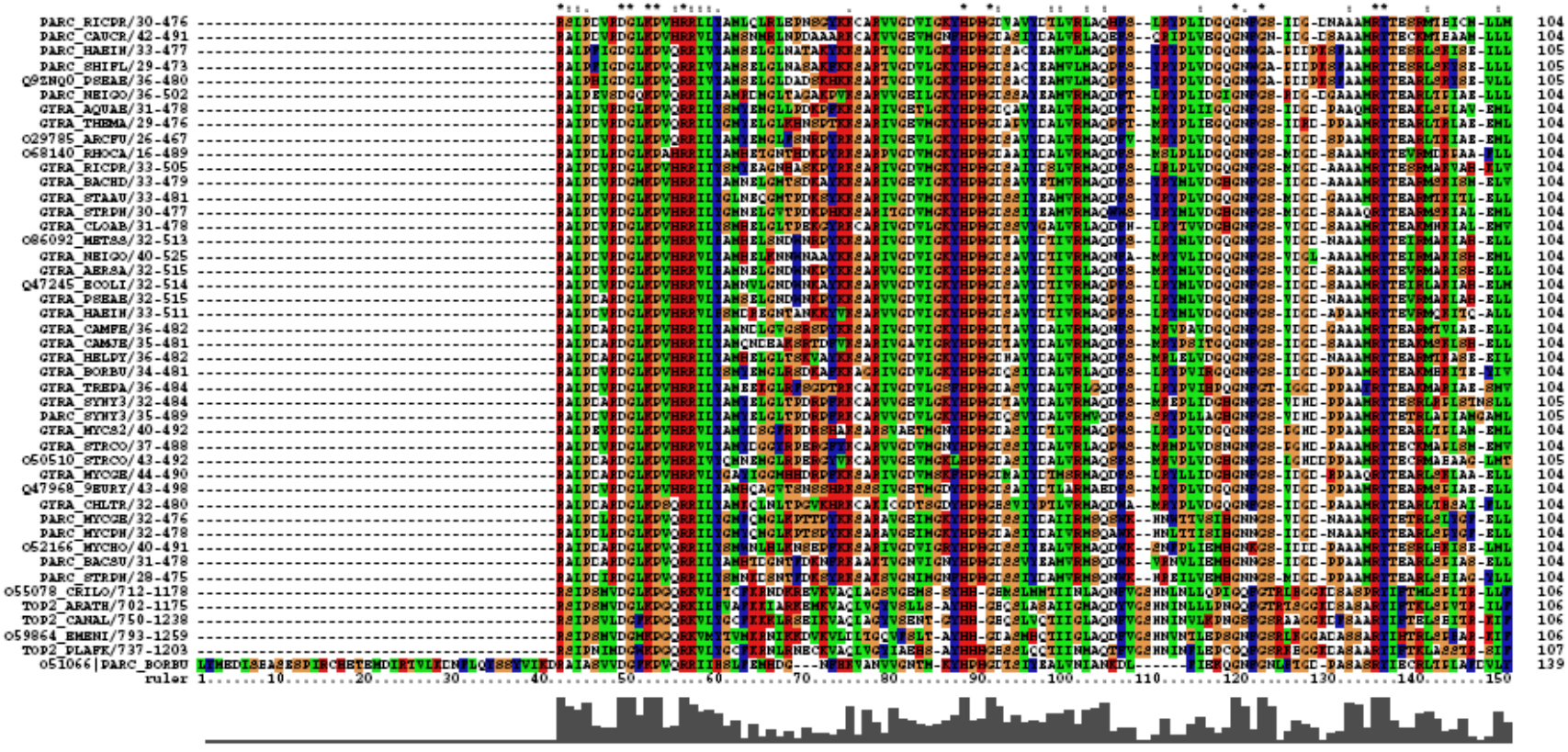
What we've done: in a nutshell

- 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 $\longrightarrow$ important residues $\longrightarrow$ are they there?

CLUSTAL X (1.83) MULTIPLE SEQUENCE ALIGNMENT

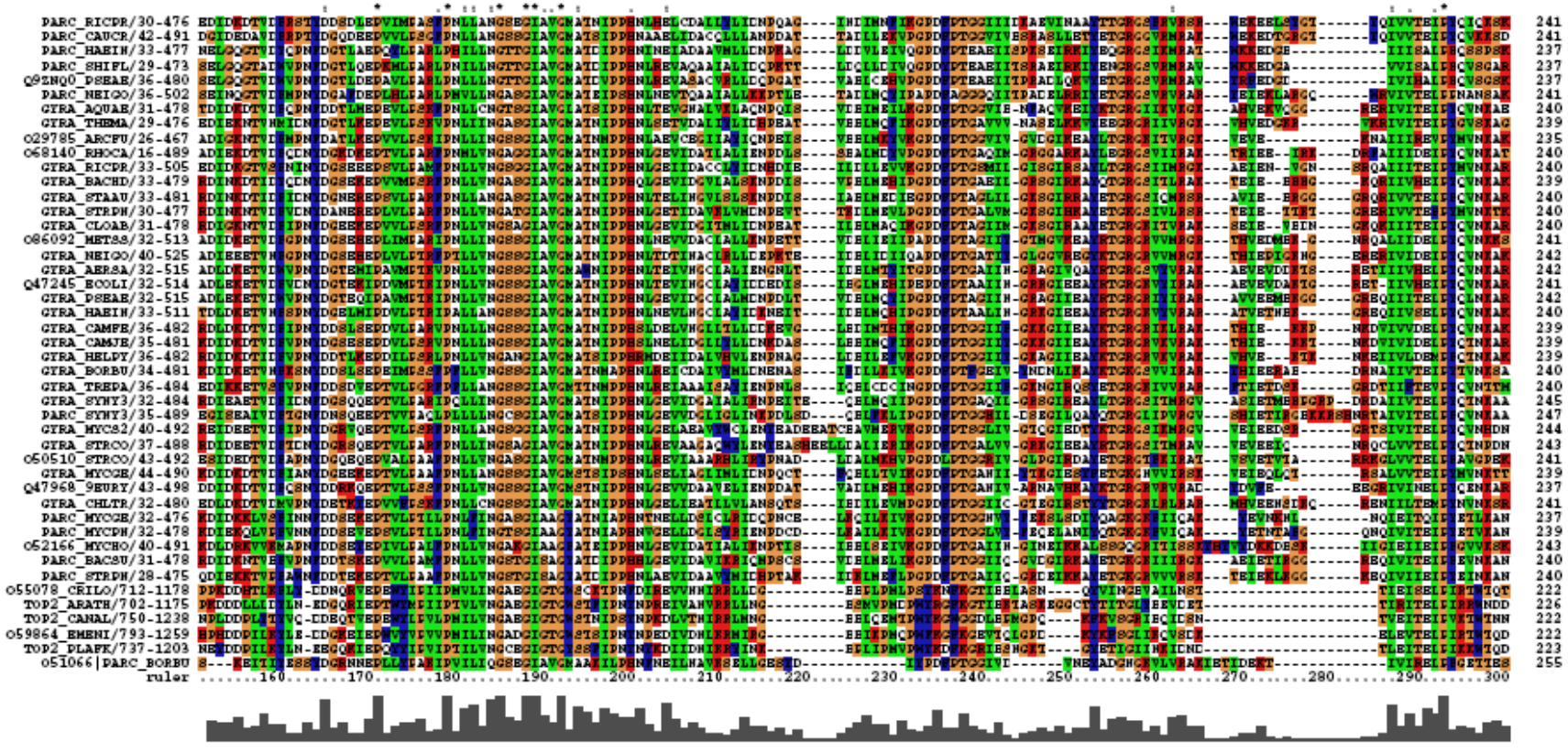
File: /home/sambit/phylo/proj/sample.ps

Date: Sun Apr 13 21:51:25 2008



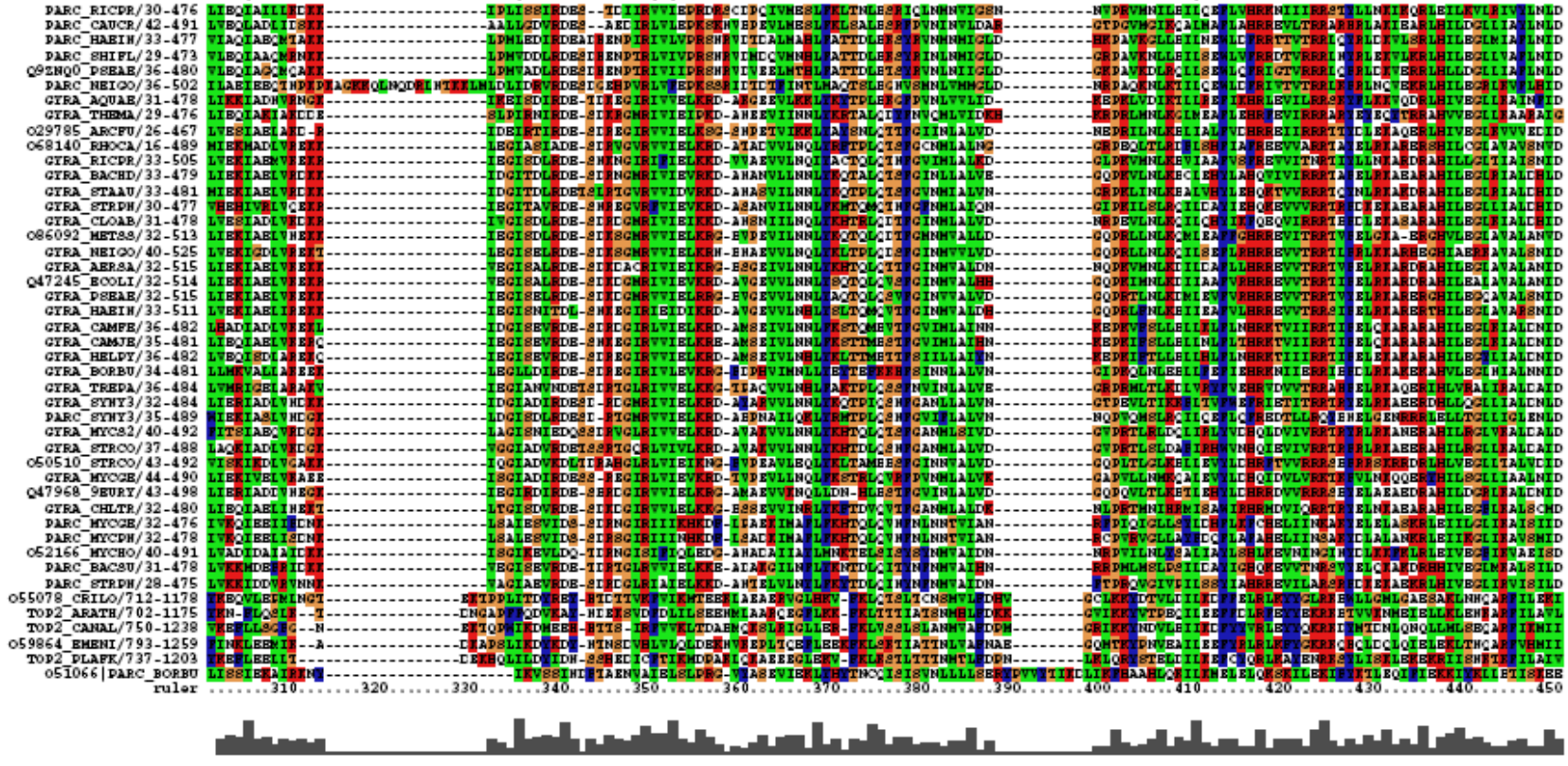
CLUSTAL X (1.83) MULTIPLE SEQUENCE ALIGNMENT

File: /home/sambit/phylo/proj/sample.ps Date: Sun Apr 13 21:51:25 2008



CLUSTAL X (1.83) MULTIPLE SEQUENCE ALIGNMENT

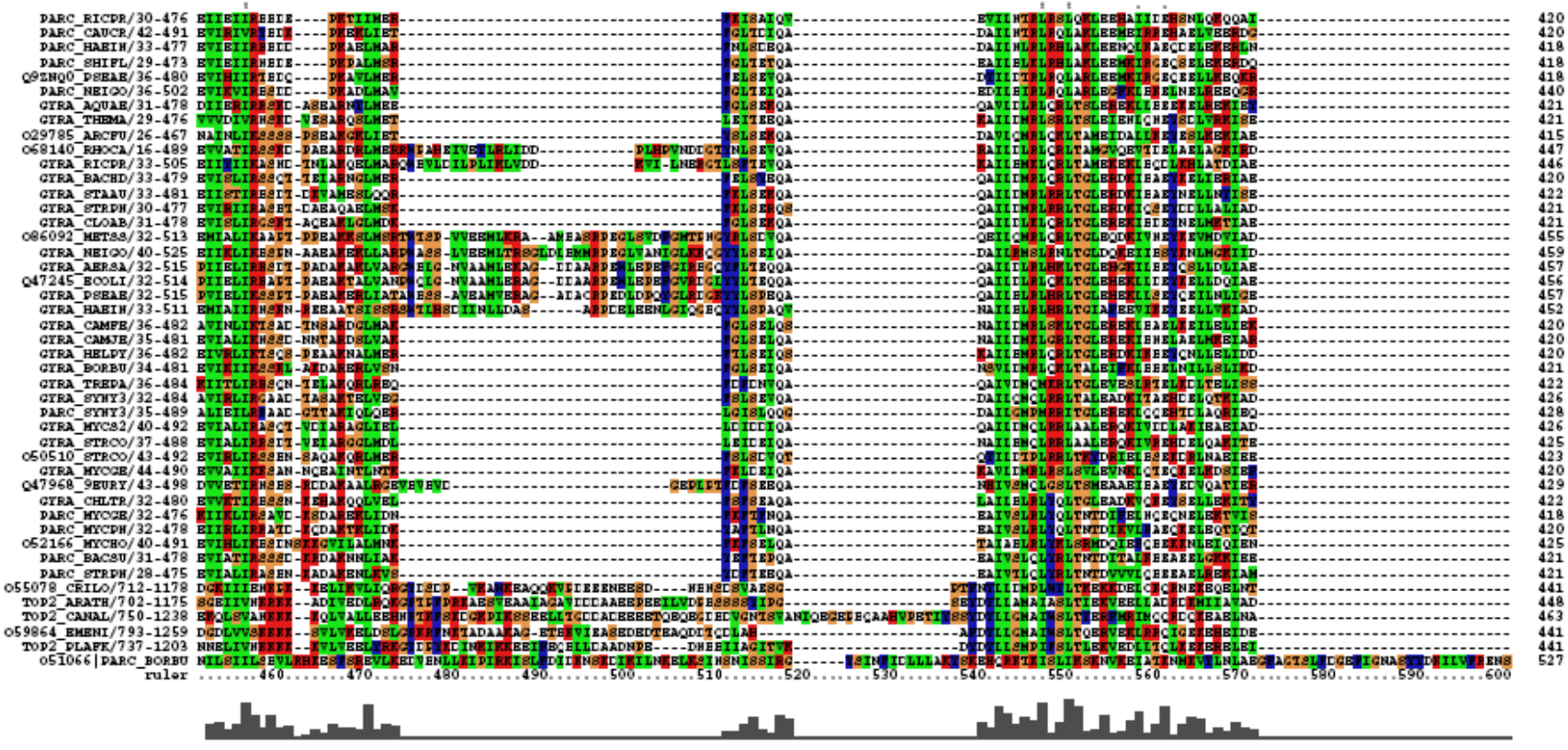
File: /home/sambit/phylo/proj/sample.ps Date: Sun Apr 13 21:51:25 2008



CLUSTAL X (1.83) MULTIPLE SEQUENCE ALIGNMENT

File: /home/sambit/phylo/proj/sample.ps

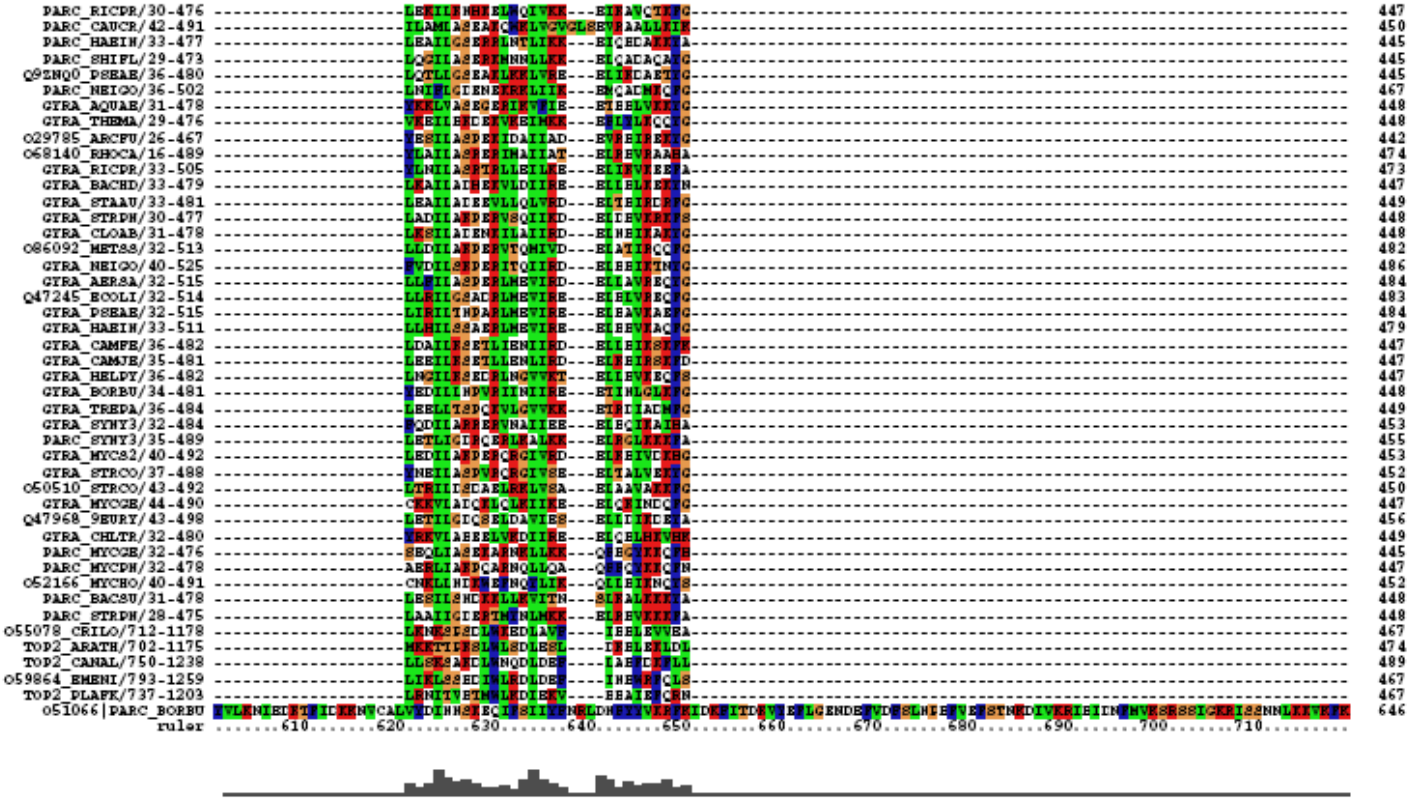
Date: Sun Apr 13 21:51:25 2008

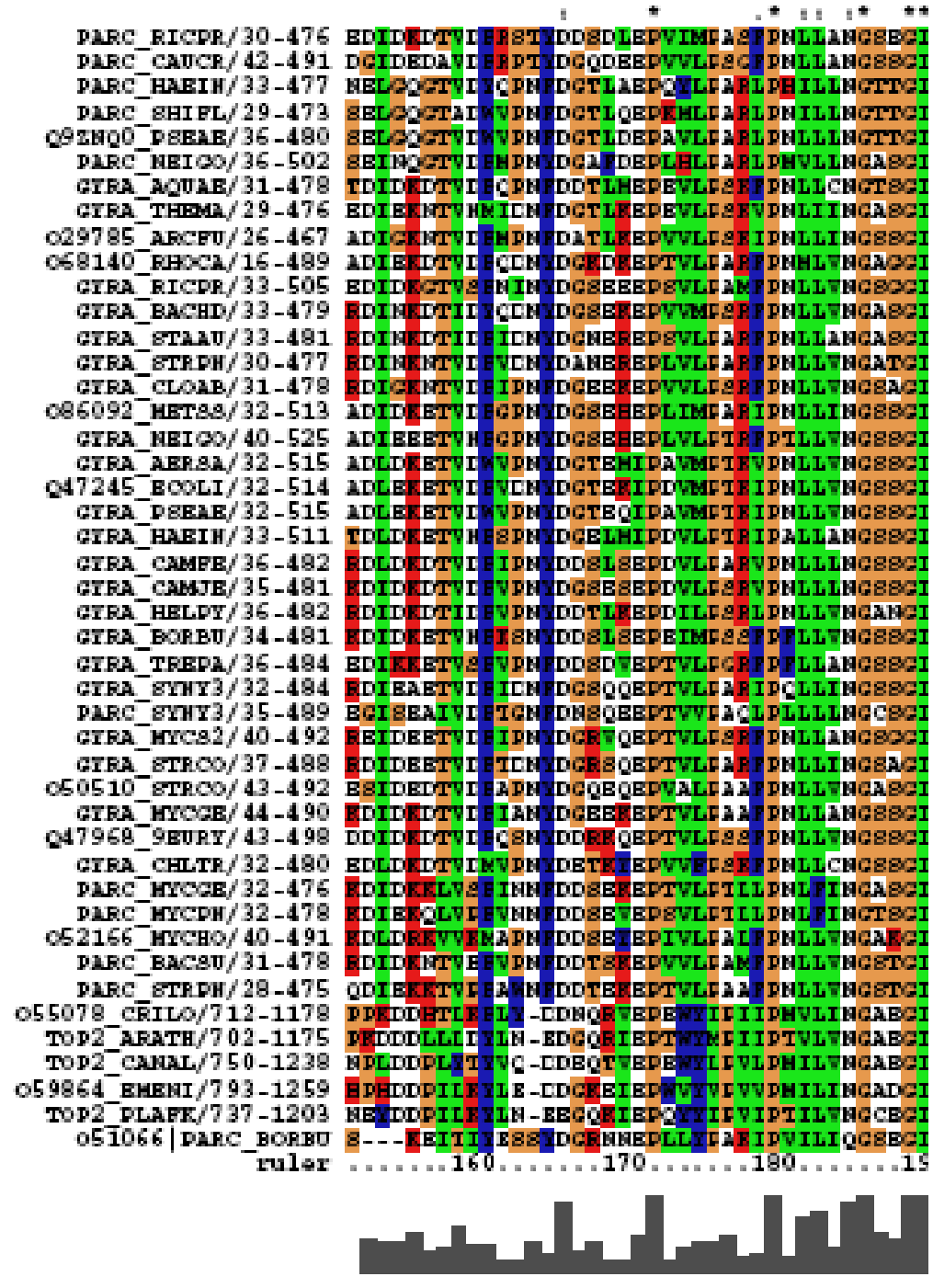


CLUSTAL X (1.83) MULTIPLE SEQUENCE ALIGNMENT

File: /home/sambit/phylo/proj/sample.ps Date: Sun Apr 13 21:51:25 2008

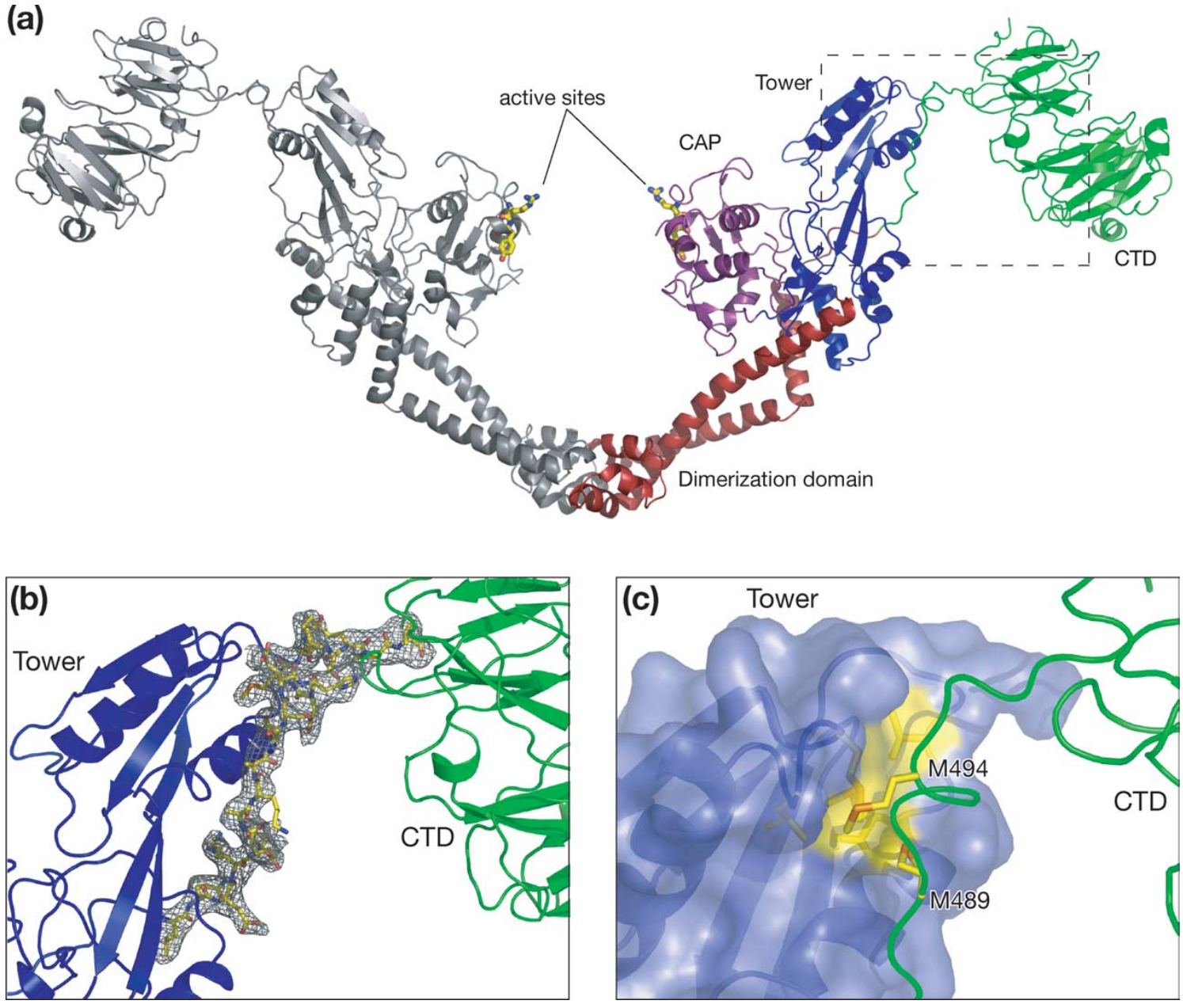
Page 5 of 5




PARC_RICPR/30-476
PARC_CAUCR/42-491
PARC_HAEIN/33-477
PARC_SHIFL/29-473
Q9ZNQ0_PSEAE/36-480
PARC_NEIGO/36-502
GYRA_AQUAE/31-478
GYRA_THEMA/29-476
O29785_ARCFU/26-467
O68140_RHOCA/16-489
GYRA_RICPR/33-505
GYRA_BACHD/33-479
GYRA_STAAU/33-481
GYRA_STRPN/30-477
GYRA_CLOAB/31-478
O86092_METSA/32-513
GYRA_NEIGO/40-525
GYRA_AERSA/32-515
Q47245_ECOLI/32-514
GYRA_PSEAE/32-515
GYRA_HAEIN/33-511
GYRA_CAMFE/36-482
GYRA_CAMJE/35-481
GYRA_HELPY/36-482
GYRA_BORBU/34-481
GYRA_TREPA/36-484
GYRA_SYNY3/32-484
PARC_SYNY3/35-489
GYRA_MYCA2/40-492
GYRA_STRCO/37-488
O50510_STRCO/43-492
GYRA_MYCGE/44-490
Q47968_9EURY/43-498
GYRA_CHLTR/32-480
PARC_MYCGE/32-476
PARC_MYCPN/32-478
O52166_MYCHO/40-491
PARC_BACSU/31-478
PARC_STRPN/28-475
O55078_CRILO/712-1178
TOP2_ARATH/702-1175
TOP2_CANAL/750-1238
O59864_EMENI/793-1259
TOP2_PLAFK/737-1203
O51066|PARC_BORBU
ruler
160
170
180

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.


(a)
active sites
CAP
Tower
CTD
Dimerization domain
(b)
Tower
CTD
(c)
Tower
M494
CTD
M489

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.