

CMSC 838T Project 1

In this paper we use on-line bioinformatics tools to analyze an unknown protein sequence, attempting to determine its 3D structure and function.

Target CASP-5 sequence

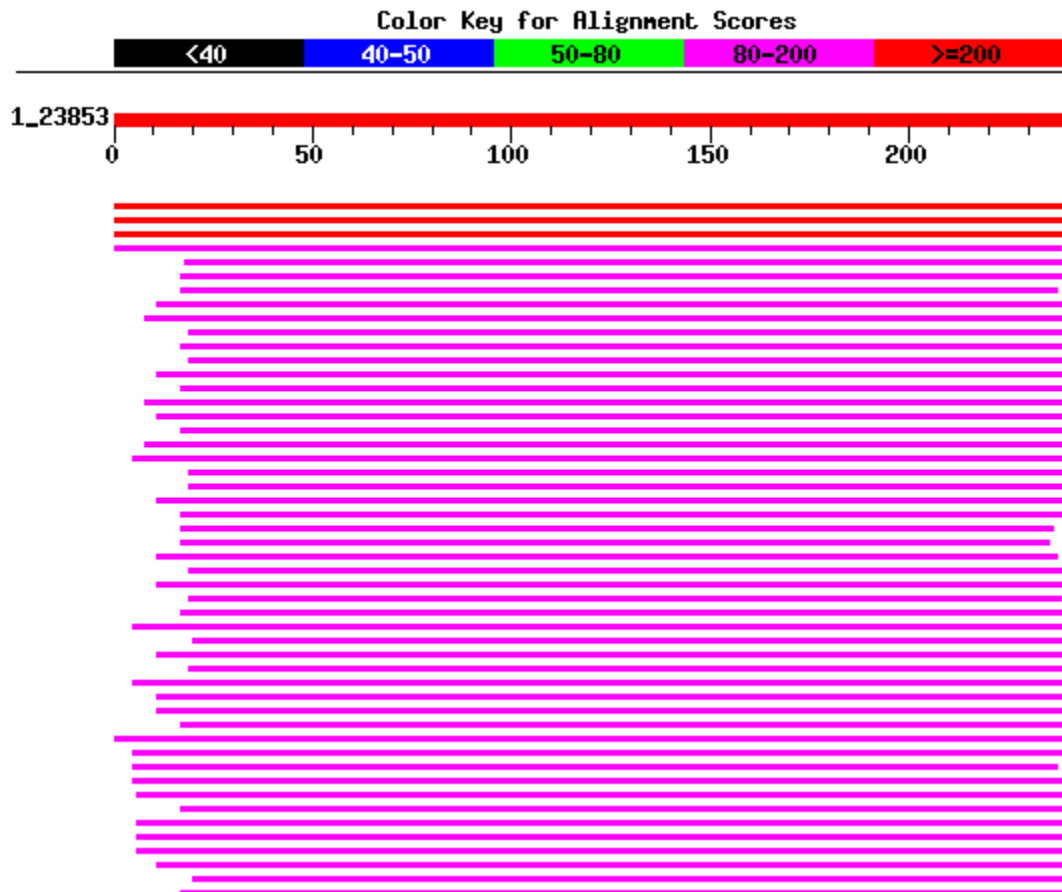
```
MGRAFEYRRAAKEKRWDKMSKVFPKLAKAITLAAKDGGSEPDTNAKLRTAILNAKAQNMPKDNI
DAAIKRASSKEGNLSEITYEGKANFGVLIIMECMTDNPTRTIANLKS YFNKTQGASIVPNGSLE
FMFNRKSVFECLKNEVENLKLSLEDLEFALIDYGLEEELEEVEDKIIIRGDYNSFKLLNEGFESL
KLPILKASLQRIATTPIELNDEQMELTEKLLDRIEDDDVVALYTNIE
```

Perform the following analyses for each protein sequence

1) Find similar protein sequences

- BLASTP returns 1 exact match, 2 almost exact matches, and many high-quality matches.

(For this project we ignore all exact matches in the sequence / structure databases, since the goal is to determine protein function based on its similarity to other proteins.)



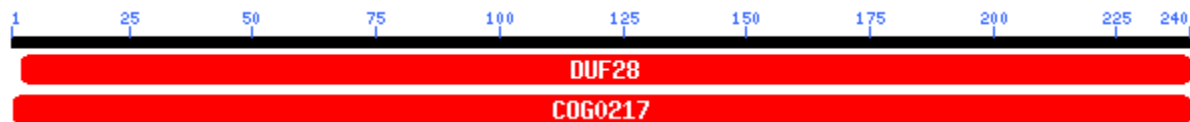
Top 10 matching sequences, their reference numbers, bit scores, and E-values

```
gi|15644791|ref|NP_206961.1| conserved hypothetical protein... 439 e-122
gi|15611219|ref|NP_222870.1| putative [Helicobacter pylori ... 434 e-121
gi|15792496|ref|NP_282319.1| hypothetical protein Cj1172c [... 226 3e-58
gi|23136969|ref|ZP_00118681.1| hypothetical protein [Cytoph... 174 1e-42
gi|7339729|gb|AAF60954.1|AF104936_2 unknown [Riemerella ana... 156 3e-37
gi|23054807|gb|ZP_00080944.1| hypothetical protein [Geobact... 139 4e-32
gi|16079834|ref|NP_390660.1| similar to spore coat protein ... 130 1e-29
gi|20807595|ref|NP_622766.1| conserved hypothetical protein... 129 5e-29
gi|28211819|ref|NP_782763.1| glucose-1-phosphate adenylyltr... 128 7e-29
gi|23502575|ref|NP_698702.1| conserved hypothetical protein... 128 9e-29
gi|17986604|ref|NP_539238.1| GLUCOSE-1-PHOSPHATE ADENYLYLTR... 127 1e-28
```

Matches were mostly to hypothetical bacterial proteins (i.e., proteins predicted from bacterial DNA) with unknown function.

2) Find protein family / conserved regions using automated tools

- RPS-BLAST found two closely matching protein families



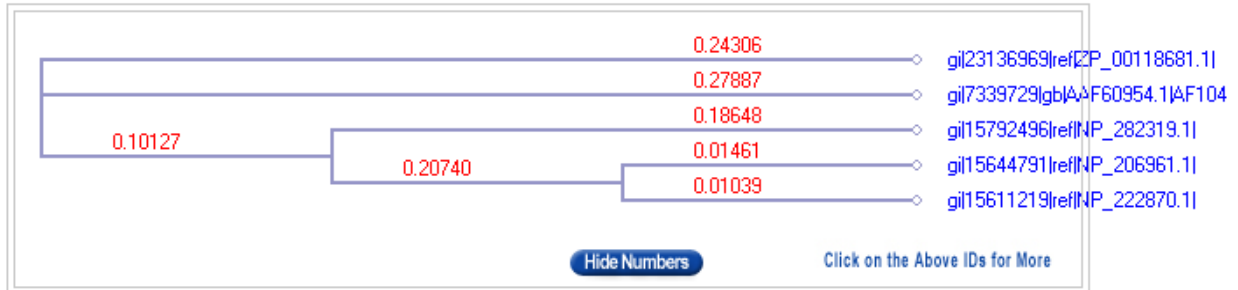
```
gnl|CDD|9344 pfam01709, DUF28, Domain of unknown function DUF28. This domai... 257 6e-70
gnl|CDD|10092 COG0217, COG0217, Uncharacterized conserved protein [Function ... 244 6e-66
```

Both protein families have unknown function. One is derived from bacterial/yeast DNA.

- HMMER found no matches to known protein families in the PIR homology domain.

3) Perform multiple sequence alignment to identify conserved regions and infer phylogenetic trees

- Alignment using CLUSTALW of top 4 matches to target yields the following phylogenetic tree and MSA, where target sequence number is 15644791 (3rd sequence in list).



```
gi|23136969|ref|ZP_00118681.1|
gi|7339729|gb|AAF60954.1|AF104
gi|15644791|ref|NP_206961.1|
gi|15611219|ref|NP_222870.1|
gi|15792496|ref|NP_282319.1|
```

```
MGRAYELRKGRRAKRFDRMAKAFTRLGKEIVMAAQGGNIETNSRLRTA
-----MAKTFSKIGKDIALAVKAGGPDPSNPALRRC
MGRAFEYRRAAKEKRWDKMSKVFPKLAKAITLAAKGGSEPDINAALRTA
MGRAFEYRRAAKEKRWDKMSKVFPKLAKAITLAAKEGGSEPDINAALRTA
MGRAFEYRRASKEARWDKMSKLFPKLAKAIQVAAKEGGTDDPMNPKLRSA
*: * * . . . * * : * . * * : : * . * * .
```

```
gi|23136969|ref|ZP_00118681.1|
gi|7339729|gb|AAF60954.1|AF104
gi|15644791|ref|NP_206961.1|
gi|15611219|ref|NP_222870.1|
gi|15792496|ref|NP_282319.1|
```

```
VNNAKSVNMPKDRIDAAIKRASGKDES DYEEVVFEGYGPYGIAILVECAT
IQNAKGANMPKDNVERAIKKASGADAENYEEITYEGYQGGVAVFFVECTT
ILNAKAQNMPKDNIDAAIKRASKE-GNLSEITYEGKANFGVLIIMECMT
ILNAKAQNMPKDNIDAAIKRASKE-GNLSEITYEGKANFGVLIIMECMT
IATAKANMMPKDNIDAAIKRASGKDSADIKNIHYEGKAAHGALVIVECMS
: . * * . * * * . : : * * * * . : : * * . * . : * * :
```

```
gi|23136969|ref|ZP_00118681.1|
gi|7339729|gb|AAF60954.1|AF104
gi|15644791|ref|NP_206961.1|
gi|15611219|ref|NP_222870.1|
gi|15792496|ref|NP_282319.1|
```

```
DNNTRTVANVRSYFTR-SGGALGKTGSLEFLFERKGVK-----IDG--T
NNSTRTVANVRAIFNK-FDGNLKGKNGELSFLFDRKGIFT-----LEKSLI
DNPTRTIANLKS YFNKTQGASIVPNGSLEFMFNKRSVF ECLKNEVEN--L
DNPTRTIANLKS YFNKTQGASIVPNGSLEFMFNKRSVF ECLKSEVEN--L
DNPTRTVANVKAI FSK-NGGEVLQNGSLGFMFTRKAVFH-----LEK--F
:* * * * * : * . : . : . * . * * * * * : : :
```

```
gi|23136969|ref|ZP_00118681.1|
gi|7339729|gb|AAF60954.1|AF104
gi|15644791|ref|NP_206961.1|
gi|15611219|ref|NP_222870.1|
gi|15792496|ref|NP_282319.1|
```

```
GLDPEELELELIDFGAEEIVKDGNEIFIYTA FVDFGTMLKELEQRNITVI
NMDWEEFEMEMIDGGAEDIDSDETEVMVTTAFEDFGSLSHKLEDELGLIEVK
KLSLEDELFALIDYGLEEELEVEDKIIIRGDYNSFKLLNEGFESLKLPII
KLSLEDELFALIDYGLEEELEVEDKIIIRGDYNSFKLLNEGFESLRLPII
AGDLEELELDLIDAGLEEELEQNEELVISGDYTA FGE LSSAIEAKGLVLK
. * * * : * * * * : . : : : * : : : : :
```

```
gi|23136969|ref|ZP_00118681.1|
gi|7339729|gb|AAF60954.1|AF104
gi|15644791|ref|NP_206961.1|
gi|15611219|ref|NP_222870.1|
gi|15792496|ref|NP_282319.1|
```

```
NAETERIPNTTTTLTTEQQEEIYKLEKFEDDDVQAVFHNMAETE----
NAELQRI PNISKSVSEEQFIANMKMLQRFEEDDDVQNVYHNMEITDELMK
KASLQRIATTPIELNDEQMELTEKLLDRIEDDDVVALYTNIE-----
KAGLQRIATTPIELNDEQMELTEKLLDRIEDDDVVALYTNIE-----
KAGLEYIPNPNVFSFEEQLSDIEKLLDKLEDDDDVQAVYTNID-----
:* : * . . . . * * * * * : * * * * * : * :
```

```
gi|23136969|ref|ZP_00118681.1|
gi|7339729|gb|AAF60954.1|AF104
gi|15644791|ref|NP_206961.1|
gi|15611219|ref|NP_222870.1|
gi|15792496|ref|NP_282319.1|
```

```
--
KL
--
--
--
```

- Alignment using T-COFFEE of top 4 matches to target yields the following MSA and consensus sequence, where target sequence number is 15644791 (1st sequence in list).

```

gi | 15644791 | ref  MGRAFEYRRAAKEKRWDKMSKVFPKLAKAITLAAKGGSEPDNTNAKLRRTAILNA
gi | 15611219 | ref  MGRAFEYRRAAKEKRWDKMSKVFPKLAKAITLAAKEGGSEPDNTNAKLRRTAILNA
gi | 15792496 | ref  MGRAFEYRRASKEARWDKMSKLFPKLAKAIQVAAKEGGTDPDMNPKLRSATATA
gi | 23136969 | ref  MGRAYELRKGRRAKRFDRMAKAFTRLGKEIVMAAQGGGNIETNSRLRTAVNNA
gi | 7339729 | gb | A -----MAKTFFSKIGKDIALAVKAGGPDPSNPALRRCIQNA

Cons                *: * * . : : * * : * * * * : : * . * * . : . *

gi | 15644791 | ref  KAQNMPKDNIDAAIKRASSKEG-NLSEITYEGKANFGVLIIMECMTDNPTRTIA
gi | 15611219 | ref  KAQNMPKDNIDAAIKRASSKEG-NLSEITYEGKANFGVLIIMECMTDNPTRTIA
gi | 15792496 | ref  KANNMPKDNIDAAIKRASGKDSADIKNIHYEGKAAHGALVIVECMSDNPTRTVA
gi | 23136969 | ref  KSVNMPKDRIDAAIKRASGKDESDYEEVVFEGYGPYGIALLVECATDNNTRTVA
gi | 7339729 | gb | A KGANMPKDNVERAIKKASGADAENYEEITYEGYQGQGVAFVVECTTNNSTRTVA

Cons                * . * * * * . : : * * : * * . : : : : * * . * . : : * * : : * * * * : *

gi | 15644791 | ref  NLKSYFNKTQGASIVPNGSLEFPMNRKSVFECLKNEVENLKLSLEDLEFALIDY
gi | 15611219 | ref  NLKSYFNKTQGASIVPNGSLEFPMNRKSVFECLKSEVENLKLSLEDLEFALIDY
gi | 15792496 | ref  NVKAIFSKN-GGEVLQNGSLGFMFTRKAVF-----HLEKFAGDLEELELDLIDA
gi | 23136969 | ref  NVRSYFTRSGG-ALGKTGSLEFLFERKGVIKIDGT-----GLDPEELELELIDF
gi | 7339729 | gb | A NVRAIFNKFDG-NLGKNGELSFLFDRKGIFTLEKSLI---NMDWEEFEMEMIDG

Cons                * : : : * . : * : : . * . * * : * * * * : : : : * * : : * * : : * *

gi | 15644791 | ref  GLEELEEVEDKIIIRGDYNSFKLLNEGFESLKLPIKASLQRIATTPIELNDEQ
gi | 15611219 | ref  GLEELEEVGDKIIIRGDYNSFKLLNEGFESLRLPIIKAGLQRIATTPIELNDEQ
gi | 15792496 | ref  GLEELEQNEEELVISGDYTAFGELSSAIEAKGLVLKAGLEYIPNNPVSFSEEQ
gi | 23136969 | ref  GAEEIVKDGNEIFIYTAQVDFGTMLKELEQRNITVINAETERIPNTTTTTLTEQ
gi | 7339729 | gb | A GAEDIDSDETEVMVTTAFEDFGSLSHKLDDELGIEVKNAELQRIPIKSKSVSEEQ

Cons                * * : : . : : : : * : : : : * : * . . . . * *

gi | 15644791 | ref  MELTEKLLDRIEDDDVVVALYTNIE-----
gi | 15611219 | ref  MELTEKLLDRIEDDDVVVALYTNIE-----
gi | 15792496 | ref  LSDIEKLLDKLEDDDDVQAVYTNID-----
gi | 23136969 | ref  QEEIYKLEKFEDDDVQAVFHNMAETE-----
gi | 7339729 | gb | A FIANMKMLQRFEEDDDVQNVYHNMEITDELTKKLL

Cons                * : * : : : * * * * : : * :

```

The multiple sequence alignment appears to be somewhat conserved across a large portion of the protein.

4) Predict secondary structure, 3D structure

- Swiss-Model finds (in addition to actual structure of protein) two proteins with known 3D structure with around 30% sequence identity to use as templates for comparative modeling.

Found 1lfpA.pdb with $P(N)=5e-26$

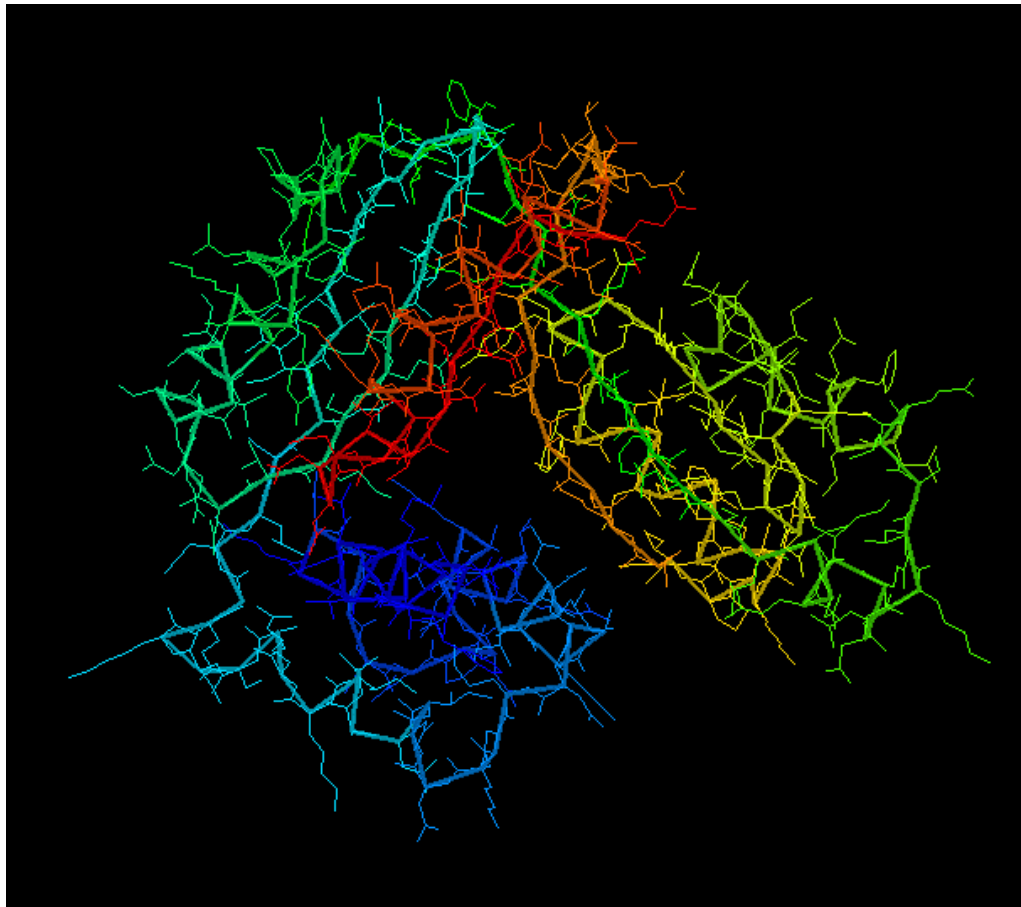
Found 1konA.pdb with $P(N)=9e-21$

1lfpA.pdb: 34.3 % identity

1konA.pdb: 32.6 % identity

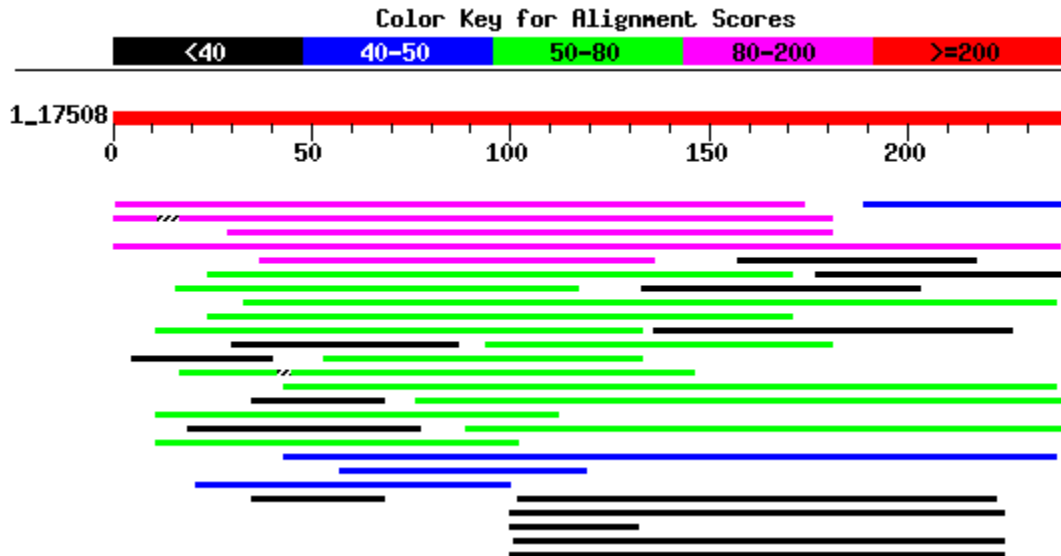
PDB entry 1lfpA is in addition described as an RNA binding protein, and a related MEDLINE link to a paper titled "Crystal structure of conserved hypothetical protein Aq1575 from Aquifex aeolicus" indicates its structure seems to be a new fold.

- 3D-PSSM finds a 100% match to a fold (c1mw7a) in its library, and good matches to two other folds (c1kona, c1lfpA) in the library (with E-values of 0.0027 and 28-30% identity). Both additional folds are bacterial proteins, one hypothetical that is labeled as an RNA binding protein.
- Overall, results seem to indicate lack of significant structural similarity with other proteins.



5) Look for source of protein in genomic DNA, cDNA

- TBLASTN finds a few potential genes expressing this protein in the GenBank DNA database.



Top 10 matching sequences, their reference numbers, bit scores, and E-values

gi 17330420 gb BH384278.1 BH384278	AG-ND-125J15.TR ND-TAM A...	153	3e-36
gi 17320126 gb BH373984.1 BH373984	AG-ND-121K13.TF ND-TAM A...	140	9e-34
gi 17338337 gb BH392196.1 BH392196	AG-ND-132H5.TR ND-TAM An...	112	8e-25
gi 20373967 gb BH771010.1 BH771010	LLMGtag732 MG1363 Random...	105	1e-21
gi 17314411 gb BH368367.1 BH368367	AG-ND-152E23.TF ND-TAM A...	104	2e-21
gi 17332712 gb BH386570.1 BH386570	AG-ND-101I16.TF ND-TAM A...	64	3e-21

- What are possible sources of target protein?

The top three matches are from the genomic DNA of an African malaria mosquito, *Anopheles gambiae*, stored in a BAC clone library. The fourth match is from random sequence tags of the bacteria, *Lactococcus cremoris*.

6) Summary

Based on these results, it appears likely the target protein is an hypothetical protein with unknown properties, either from bacteria/yeast with or possibly the African mosquito.