FunVar Update

31 March, 2017

Overview

- Developing visualisation tool
  - cath-cluster-web

- Improving accessibility of FunFam alignments
  - Sequence MD5 -> UniProtKB
  - FASTA -> STOCKHOLM

cath-cluster-web

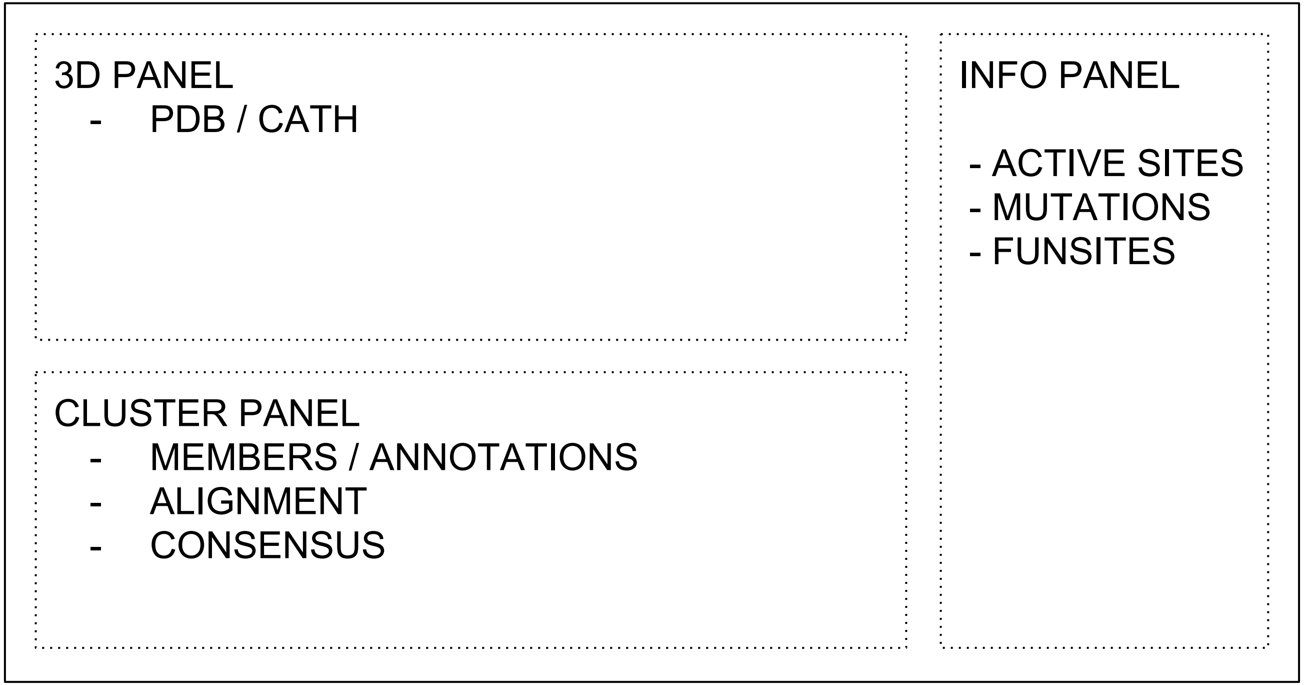
Requirements:

- 3D structural viewer
- Functional annotations
- Sequence alignments

Ideally…

- Use existing web services (CATH, PDBe, UniProtKB)
- Portable, generic, reusable

cath-cluster-web



3D PANEL

- PDB / CATH

CLUSTER PANEL

- MEMBERS / ANNOTATIONS
- ALIGNMENT
- CONSENSUS

INFO PANEL

- ACTIVE SITES
- MUTATIONS
- FUNSITES

cath-cluster-web

Issues:

- Combining existing components
  - 3D viewer, MSA viewer, Tree
- Combining data sources
  - CATH, PDBe, UniProtKB, EC, GO, ...
- Mapping between coordinate frames
  - sequence/structure
- Dynamic
  - Interaction coordinated across all components

cath-cluster-web

Choosing the right tool...

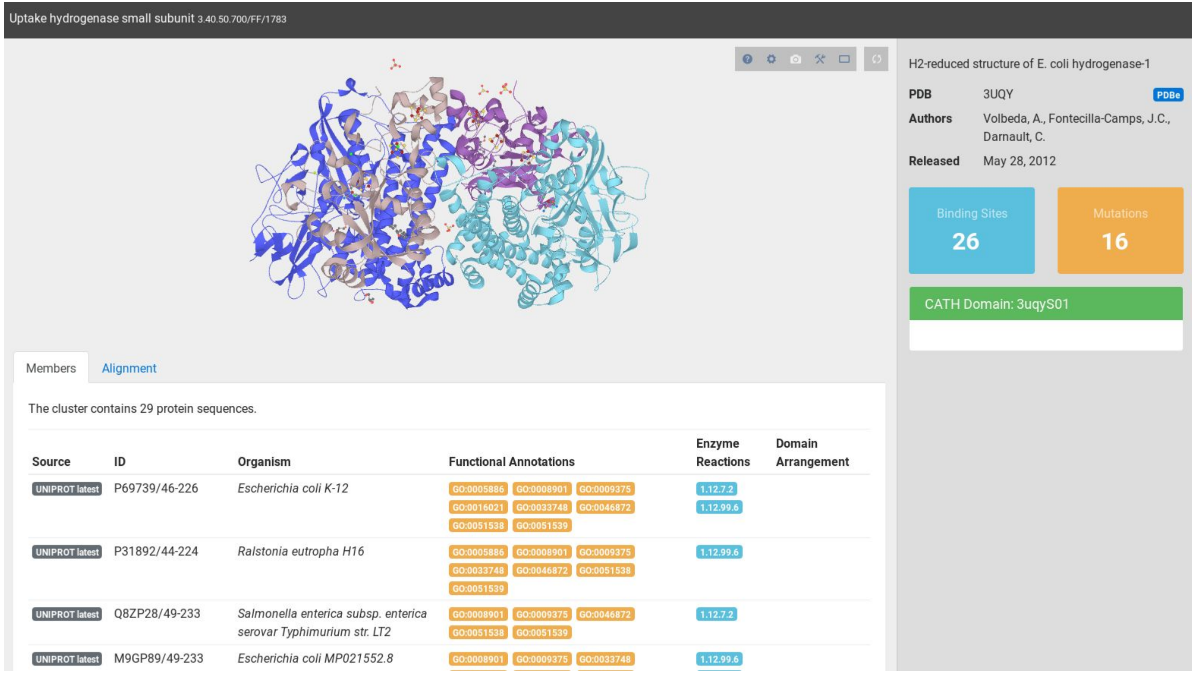
Use an existing framework to glue all views, data and events together into a single web application.

- Angular, Angular2, ReactJS, Polymer, etc, etc… ?

After discussion with PDBe dev…

- Angular2 (Google)

cath-cluster-web



Uptake hydrogenase small subunit 3.40.50.700/FF/1783

H2-reduced structure of E. coli hydrogenase-1

PDB	3UQY
Authors	Volbeda, A., Fontecilla-Camps, J.C., Darnault, C.
Released	May 28, 2012

The cluster contains 29 protein sequences.

Source	ID	Organism	Functional Annotations	Enzyme Reactions	Domain Arrangement
UNIPROT latest	P69739/46-226	*Escherichia coli K-12*	GO:0005886 GO:0008901 GO:0009375 GO:0016021 GO:0033748 GO:0046872 GO:0051538 GO:0051539	1.12.7.2 1.12.99.6	
UNIPROT latest	P31892/44-224	*Ralstonia eutropha H16*	GO:0005886 GO:0008901 GO:0009375 GO:0033748 GO:0046872 GO:0051538 GO:0051539	1.12.99.6	
UNIPROT latest	Q8ZP28/49-233	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	GO:0008901 GO:0009375 GO:0046872 GO:0051538 GO:0051539	1.12.7.2	
UNIPROT latest	M9GP89/49-233	*Escherichia coli MP021552.8*	GO:0008901 GO:0009375 GO:0033748	1.12.99.6	

cath-cluster-web

Initial data from (FASTA) alignment

- List of members (sequences)
  - Each entity based on unique sequence MD5
- Associated annotations for each member (headers)
  - CATH domains, GO terms, EC terms, UniProtKB accessions, etc
- Alignment

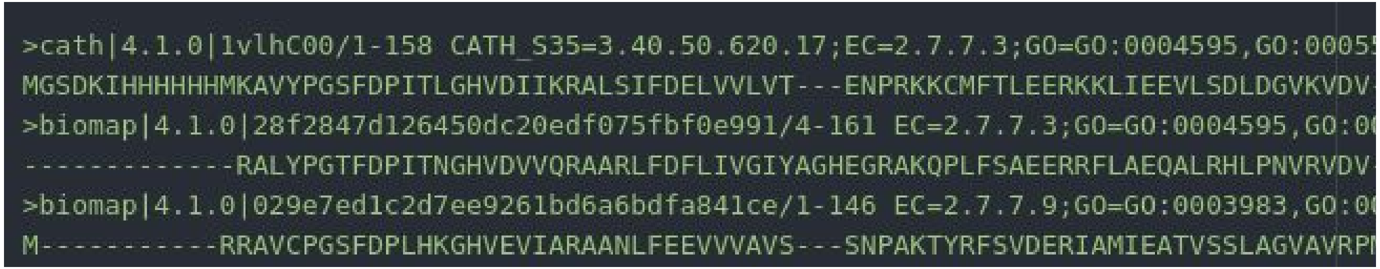
Then use web services to get...

- 3D structure, known binding sites, mutations, etc

cath-cluster-web

Meta data in FASTA is a hack...

```
> <sequence/structure_id> <annotation;...>
<ALIGNED_AA_SEQUENCE>
```



```
>cath|4.1.0|1vlhC00/1-158 CATH_S35=3.40.50.620.17;EC=2.7.7.3;GO=GO:0004595,GO:0005
MGSDKIHHHHHHMKAVYPGSFDPITLGHVDIIKRALSIFDELVVLVT---ENPRKKCMFTLEERKKLIEEVLSDLDGVKVDV
>biomap|4.1.0|28f2847d126450dc20edf075fbf0e991/4-161 EC=2.7.7.3;GO=GO:0004595,GO:0
-------------RALYPGTFDPITNGHVDVVQRAARLFDFLIVGIYAGHEGRAKQPLFSAEERRFLAEQALRHLPNVRVDV
>biomap|4.1.0|029e7ed1c2d7ee9261bd6a6bdfa841ce/1-146 EC=2.7.7.9;GO=GO:0003983,GO:0
M-----------RRAVCPGSFDPLHKGHVEVIARAANLFEEVVVAVS---SNPAKTYRFSVDERIAMIEATVSSLAGVAVRP
```

cath-cluster-web

FASTA Pros:

- Simple, easy to parse, alignments already exist

FASTA Cons:

- Forces data into unstructured headers
- No meta data (alignment id, name, date created, etc)
- No consensus information (e.g. scorecons)
- Not easy to map sequence to structure

cath-cluster-web

Also, problems using sequence MD5s?

Pros: Simple, uses existing mapping

Cons:

- Very specific to CATH-Gene3D
- Annotations are one-to-many-to-many:
  Sequence MD5
   -> one-to-many UniProtKB entries
    -> one-to-many GO/EC/organism entries

cath-cluster-web

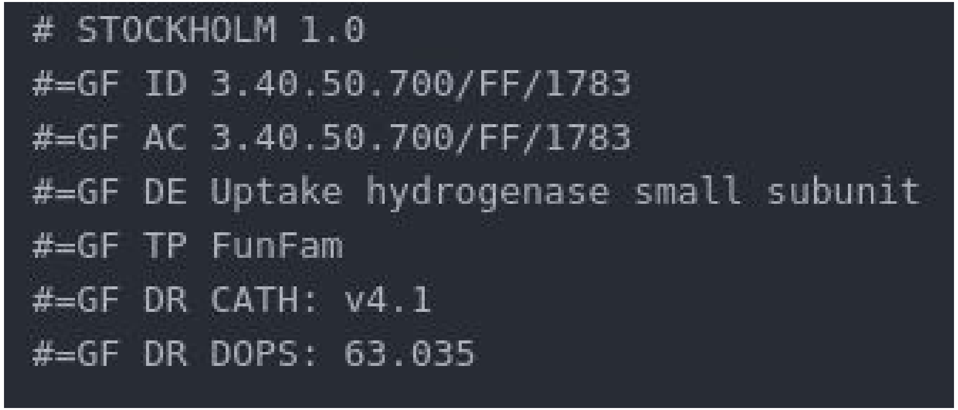
So…

1. Map all entries via UniProtKB sequences
    a. Expand sequence MD5s to UniProtKB entries
    b. Use existing filter protocols to remove redundant entries

2. Use more structured alignment format
    a. FASTA -> STOCKHOLM (as per Pfam)

STOCKHOLM

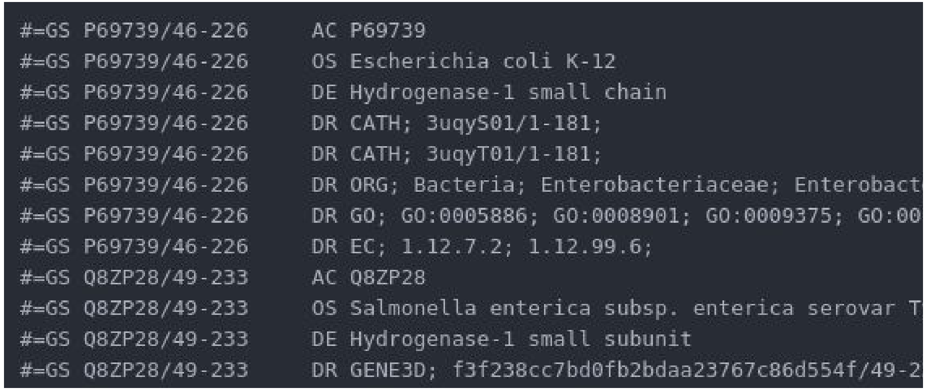
General meta data for the whole alignment...



```
# STOCKHOLM 1.0
#=GF ID 3.40.50.700/FF/1783
#=GF AC 3.40.50.700/FF/1783
#=GF DE Uptake hydrogenase small subunit
#=GF TP FunFam
#=GF DR CATH: v4.1
#=GF DR DOPS: 63.035
```

STOCKHOLM

Individual meta data for each sequence…

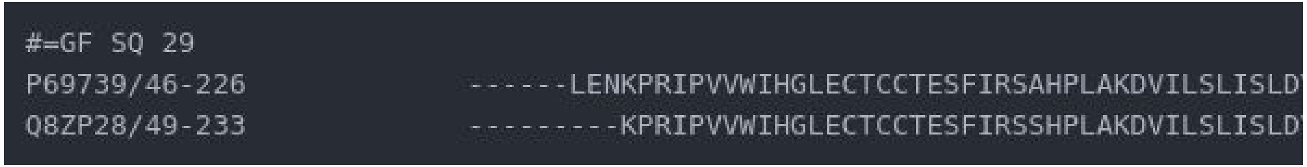


```
#=GS P69739/46-226     AC P69739
#=GS P69739/46-226     OS Escherichia coli K-12
#=GS P69739/46-226     DE Hydrogenase-1 small chain
#=GS P69739/46-226     DR CATH; 3uqyS01/1-181;
#=GS P69739/46-226     DR CATH; 3uqyT01/1-181;
#=GS P69739/46-226     DR ORG; Bacteria; Enterobacteriaceae; Enterobact
#=GS P69739/46-226     DR GO; GO:0005886; GO:0008901; GO:0009375; GO:00
#=GS P69739/46-226     DR EC; 1.12.7.2; 1.12.99.6;
#=GS Q8ZP28/49-233     AC Q8ZP28
#=GS Q8ZP28/49-233     OS Salmonella enterica subsp. enterica serovar T
#=GS Q8ZP28/49-233     DE Hydrogenase-1 small subunit
#=GS Q8ZP28/49-233     DR GENE3D; f3f238cc7bd0fb2bdaa23767c86d554f/49-2
```

CATH entry maps from UniProtKB numbering to PDB residue labels (double-checked against sequence in alignment)

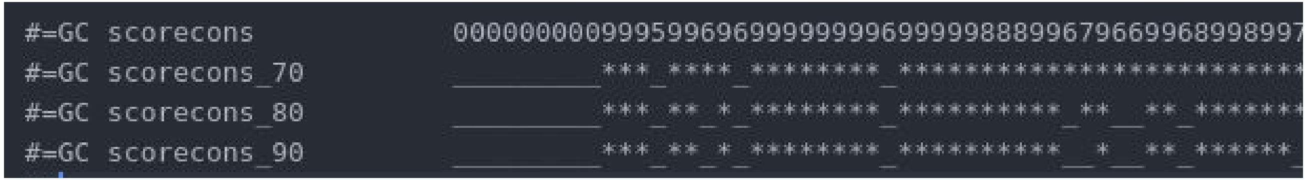
STOCKHOLM

The aligned sequences for each entry



```
#=GF SQ 29
P69739/46-226            ------LENKPRIPVVWIHGLECTCCTESFIRSAHPLAKDVILSLISLD
Q8ZP28/49-233            ---------KPRIPVVWIHGLECTCCTESFIRSSHPLAKDVILSLISLD
```

And consensus information...



```
#=GC scorecons          0000000009995996969999999969999988899679669968998997
#=GC scorecons_70       _________***_****_********_*************************
#=GC scorecons_80       _________***_**_*_********_**********_**__**_*******
#=GC scorecons_90       _________***_**_*_********_**********__*__**_******_
```

STOCKHOLM

Added Bonus:

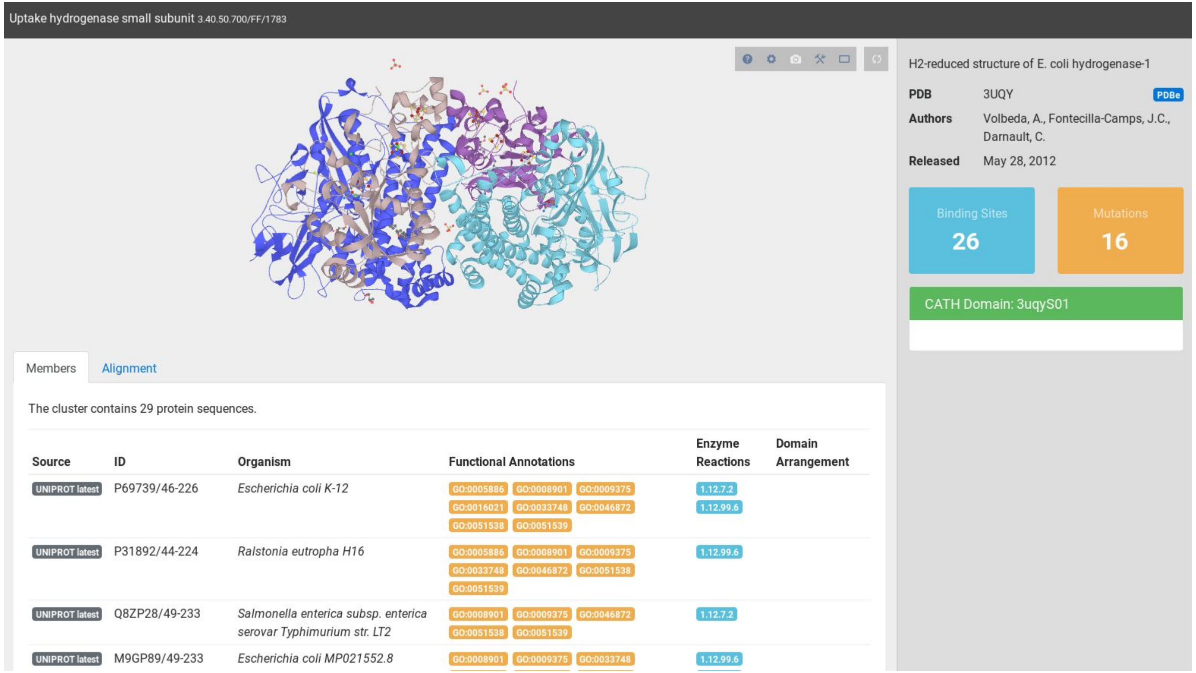
- No need for separate files (names, scorecons, DOPS, etc)
- HMMER uses this as native alignment format
  - Aligning new sequence to this alignment does what you would expect with consensus information (i.e. opens gaps)
  - No need to rerun scorecons

cath-cluster-web

Actions:

- Generate filtered STOCKHOLM alignments for all FunFams (done)
- Integrate STOCKHOLM parser into cath-cluster-web (done)

cath-cluster-web



Uptake hydrogenase small subunit 3.40.50.700/FF/1783

H2-reduced structure of E. coli hydrogenase-1

PDB	3UQY
Authors	Volbeda, A., Fontecilla-Camps, J.C., Darnault, C.
Released	May 28, 2012

The cluster contains 29 protein sequences.

Source	ID	Organism	Functional Annotations	Enzyme Reactions	Domain Arrangement
UNIPROT latest	P69739/46-226	*Escherichia coli K-12*	GO:0005886 GO:0008901 GO:0009375 GO:0016021 GO:0033748 GO:0046872 GO:0051538 GO:0051539	1.12.7.2 1.12.99.6	
UNIPROT latest	P31892/44-224	*Ralstonia eutropha H16*	GO:0005886 GO:0008901 GO:0009375 GO:0033748 GO:0046872 GO:0051538 GO:0051539	1.12.99.6	
UNIPROT latest	Q8ZP28/49-233	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	GO:0008901 GO:0009375 GO:0046872 GO:0051538 GO:0051539	1.12.7.2	
UNIPROT latest	M9GP89/49-233	*Escherichia coli MP021552.8*	GO:0008901 GO:0009375 GO:0033748	1.12.99.6	