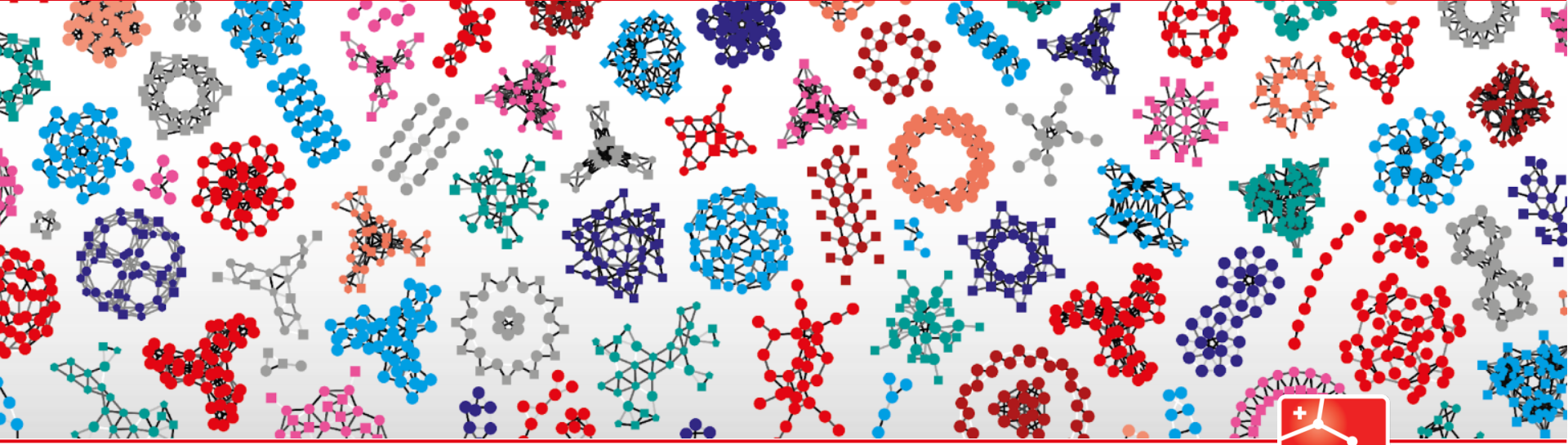


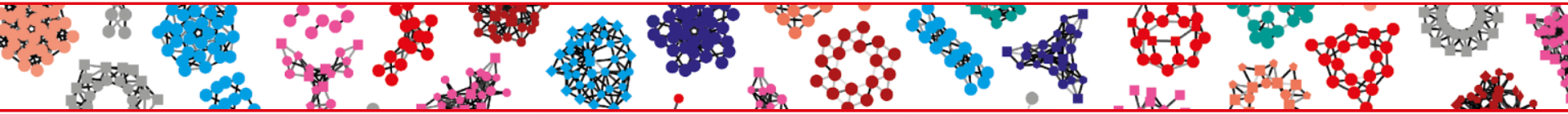
Swiss Institute of Bioinformatics

HAMAP

Jerven Bolleman, Swiss-Prot group

www.sib.swiss

Overview



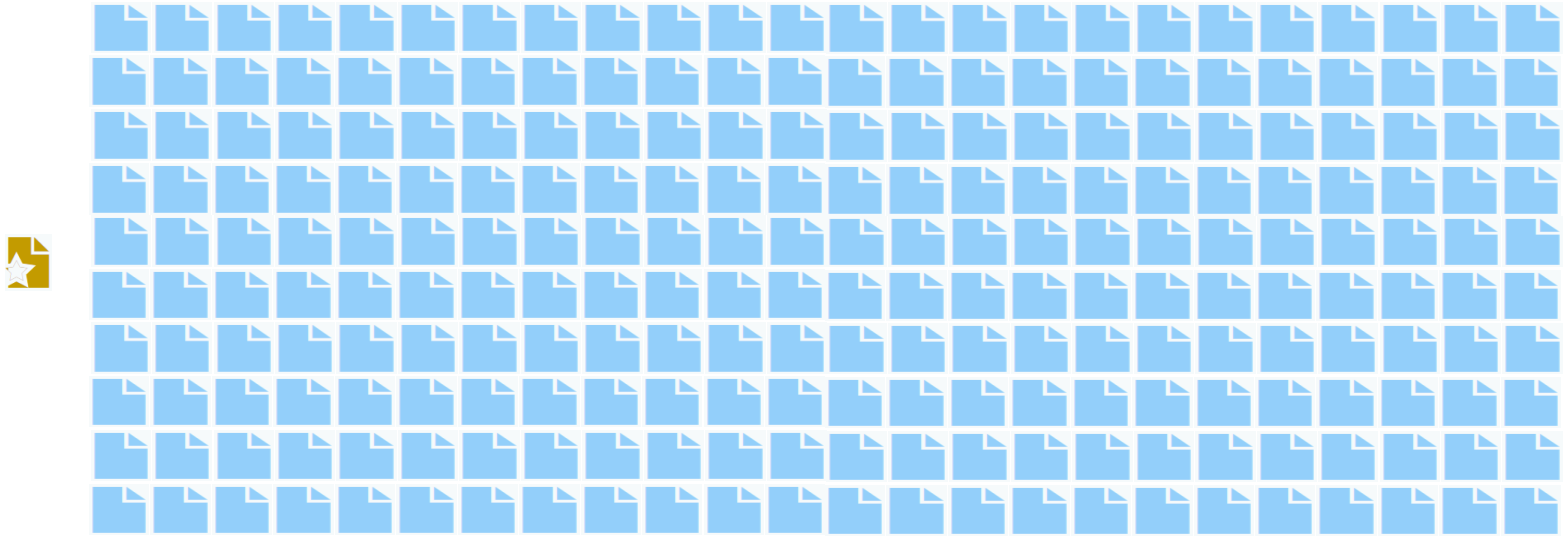
01 Biology

02 Data model

03 SPARQL

04 To Elixir

1 annotated protein vs. 220+ non annotated



Band aid ⇝ Family based annotation propagation

hamap.expasy.org

theseed.org

TIGRFAMS

J. Craig Venter INSTITUTE

www.jcvi.org/tigrfams

Similar concepts, different implementations

Family ⇝ similar proteins

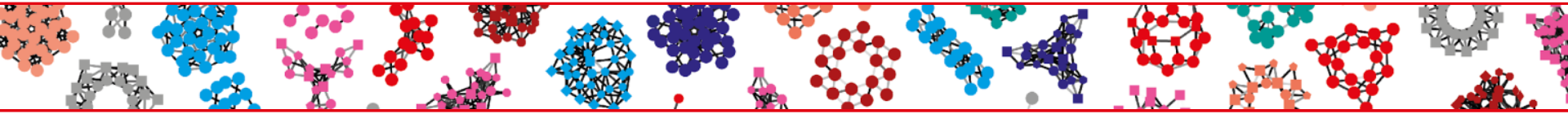
```
ADE_YEAST       ---------MVSVEFLQELPKCEHHLHLEGTLEPDLLFPLAKRNDIILPE----GFPKSV
ADE_CANGA       ---------MVPESFLLELPKCEHHLHLEGTLEPDLLFPLAKRNNIQLPD----HFPQTP
ADE_KLULA       --MAKFECTDEVTNFLTELPKCEHHLHLEGTLEPELLFQLVERNGVQLPG----TFPKTV
ADE_CANAL       --MAQYECSEHMENFLRELPKCEHHVHLEGTLEPSLLFKLAKRNNITLPE----TFPKTV
ADE_ASPFU       MC------QSPLHDFLHGLPKCEHHVHLEGCVTPELIFQLAEKNNIQLPNPATHPAYASV
ADE_ASPOR       MC------KSDLHDFLHGLPKCEHHVHLEGCLAPDLIFELAKRNNVSLPN---EPAYESI
ADE_EMENI       MCPPNTPYQSQWHAFLHSLPKCEHHVHLEGCLEPPLIFSMARKNNVSLPSPSSNPAYTSV
ADE_SCHPO       MS------NLPIYNFIRKLPKCEHHVHLEGCLSPDLVFRLAKKNGITLPS--DDAAYTTP
ADE_GIBZE       MC------KSRVHSFLQALPKVEQHLHIEGTLEPELLFTLAEKNGIELPN---DPVYESA
ADE_CAUCR       -MTDASFAPSASAEFVRGLPKAELHMHIEGSLEPELMFELAQRNGITLPFA-------SV
ADE_CAUCN       -MTDASFAPSASAEFVRGLPKAELHMHIEGSLEPELMFELAQRNGITLPFA-------SV
ADE_SPHAL       -MPDGFASHEERAAFIAGLPKAELHLHIEGSLEPELLFEFARRNRVAIPFA-------SI
ADE_RHORT       --------MAVDPAFLHALPKVELHLHIEGSLEPEMMVALAERNGLRLPYA-------SV
ADE_STRCO       --------MKRPYDALMPLPKAELHLHIEGTLEPELAFALAARNGVSLPYA-------DE
ADE_BURPP       -MTTTTVTPTPLAEKTALAPKAELHIHIEGSLEPELIFALAERNGVKLAYD-------SI
ADE_BURXL       -MTTTTVTSTPLAEKTVLAPKAELHIHIEGSLEPELIFALAERNGVKLAYD-------SI
ADE_CUPTR       -----MTIDAALAEQIRRTPKAELHVHIEGTLEPELIFRLAQRNQVALPYP-------SV
ADE_CUPNH       -----MTIDAALAEQIRRTPKAELHVHIEGTLEPELIFRLAQRNQVALPYP-------SV
ADE_CUPPJ       -----MTIDAALADKIRRTPKAELHVHIEGTLEPELIFRLAQRNHVNLPYP-------SV
ADE_RALME       -----MTIDAALADKIRRTPKAELHVHIEGTLEPERIFRLAQRNNVKLAYP-------DV
ADE_RALPJ       -----MPISSALAERIATSPKAELHIHIEGSLEPELMFALAERNGVKLPYA-------SV
ADE_RALSO       -----MPISPALAERIATSPKAELHIHIEGSLEPELMFALAERNGVKLPYA-------SV
ADE_GEOLS       -MNLTNIPRQALPELLCRMPKAELHIHIEGSLEPELIFALAERNRLQLAYP-------TI
ADE_GEOUR       -MNFDCIPREDLHGILCHMPKAELHIHIEGSLEPELIFELATRNRIQLPYP-------TI
```

Build signature of family

e.g. generalized profile (PF tools)

Hidden Markov Model (hmmer3)

Overview



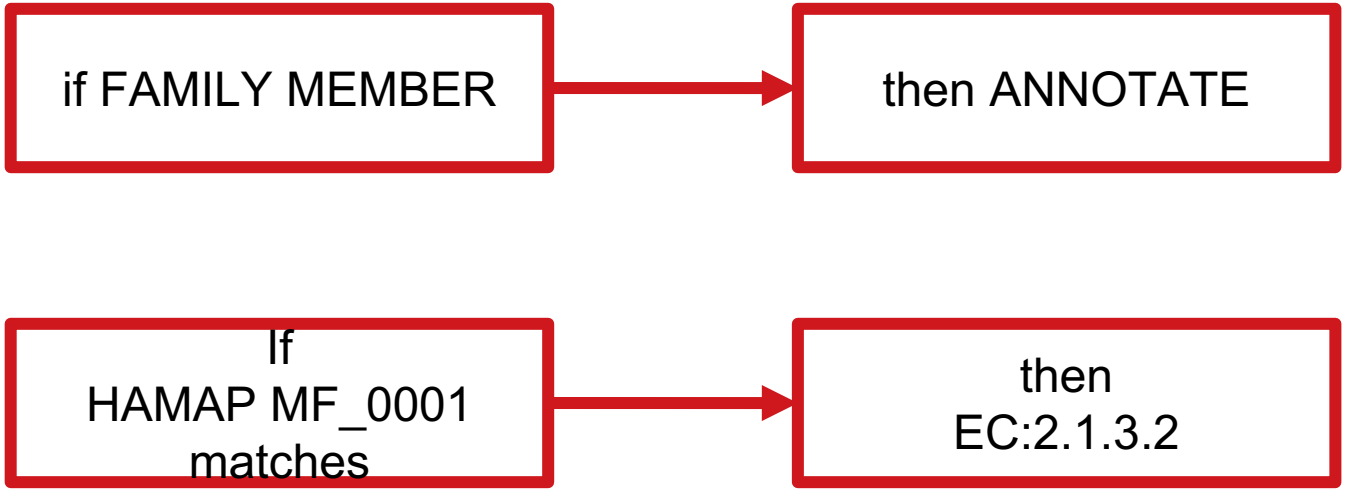
01 Biology

02 Data model

03 SPARQL

04 To Elixir

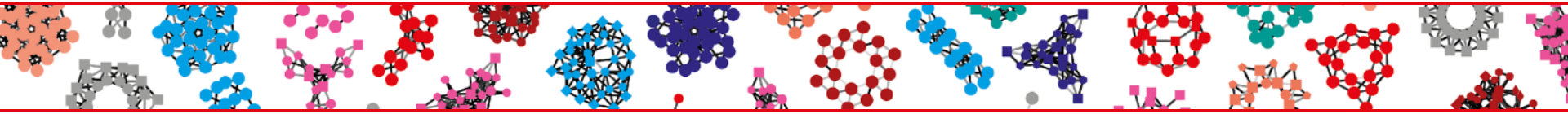
Family ⇝ similar proteins ➜ same function



Problem: non standard rule application

- You can’t get a copy of the HAMAP rule application pipeline
    - It’s married to our infrastructure
- Equivalent projects have similar issues

Overview



01 Biology

02 Data model

03 SPARQL

04 To Elixir

HAMAP as SPARQL

- **Target protein identification**
  - **Main condition: Protein family signature match**
    (requires pfscan/InterProScan results
    converted to simple RDF format)
  - **Additional conditions**
    (e.g. taxonomy, sequence length)
  - **WHERE**

HAMAP as SPARQL

- **Annotation propagation**
  - **«Simple» : GO, EC, Keywords**
  - **SELECT**
  - **«Complex» : all UniProt annotation types, incl. evidences!**
  - **CONSTRUCT**

MF_00001 “simple” rule

```
SELECT
        DISTINCT ?protein ?term
} WHERE {
        VALUES ?term {‘GO:00006221’ ‘EC:2.1.3.2’}
        VALUES ?superTaxon {taxon:2 taxon:2157}
        ?protein rdf:type up:Protein ;
                rdfs:seeAlso hamap:MF_00001 ;
                up:organism/rdfs:subClassOf ?superTaxon .
} ORDER BY ?protein ?term
```

Z9JXW0	EC:2.1.3.2
Z9JXW0	GO:00006221
Z9JII6	EC:2.1.3.2

MF_00001 “complex” and complete rule

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX uniprot:<http://purl.uniprot.org/uniprot/>
PREFIX sequence:<http://purl.uniprot.org/sequences/>
PREFIX seq_region_match:<http://example.org/sequence_region_match/>
PREFIX unirule:<http://purl.uniprot.org/unirules/>
PREFIX taxon:<http://purl.uniprot.org/taxonomy/>
PREFIX GO:<http://purl.obolibrary.org/obo/GO_>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX hs:<http://example.org/hamap_sparql/>
PREFIX up:<http://purl.uniprot.org/core/>
PREFIX seq_region_motif:<http://example.org/sequence_region_motif/>
PREFIX faldo:<http://biohackathon.org/resource/faldo#>
PREFIX keyword:<http://purl.uniprot.org/keywords/>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX predictor:<http://example.org/predictor/>
PREFIX proteome:<http://purl.uniprot.org/proteomes/>
PREFIX hf:<http://example.org/function#>
PREFIX hamap:<http://purl.uniprot.org/hamap/>
PREFIX eco:<http://purl.obolibrary.org/obo/ECO_>
PREFIX annotation:<http://purl.uniprot.org/annotation/>
PREFIX isoform:<http://purl.uniprot.org/isoforms/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
CONSTRUCT {
  _:188043 up:source unirule:MF_00001 .
  _:188044 up:source unirule:MF_00001 .
  _:188045 up:source unirule:MF_00001 .
  _:188046 up:source unirule:MF_00001 .
  _:188047 up:source unirule:MF_00001 .
  _:188048 up:source unirule:MF_00001 .
  _:188049 up:source unirule:MF_00001 .
  _:188050 up:source unirule:MF_00001 .
  _:188051 up:source unirule:MF_00001 .
  _:188052 up:source unirule:MF_00001 .
  _:188053 up:source unirule:MF_00001 .
 ?this up:alternativeName ?name1 ;
   up:annotation ?annotation3 ,
     ?annotation4 ,
     ?annotation5 ;
   up:classifiedWith GO:0004070 ,
     GO:0006221 ,
     keyword:665 ,
     keyword:808 ;
   up:recommendedName ?name0 .
 ?name0 up:ecName '2.1.3.2' ;
   up:fullName 'Aspartate carbamoyltransferase' ;
   rdf:type up:Structured_Name .
 ?name1 up:fullName 'Aspartate transcarbamylase' ;
   up:shortName 'ATCase' ;
   rdf:type up:Structured_Name .
 ?gene2 skos:prefLabel 'pyrB' .
 ?annotation3 a up:Catalytic_Activity_Annotation ;
   rdfs:comment 'Carbamoyl phosphate + L-aspartate = phosphate + N-carbamoyl-L-aspartate.' .
 ?annotation4 a up:Pathway_Annotation ;
   rdfs:comment 'Pyrimidine metabolism; UMP biosynthesis via de novo pathway; (S)-dihydroorotate from bicarbonate: step 2/3.' .
 ?annotation5 a up:Similarity_Annotation ;
   rdfs:comment 'Belongs to the ATCase/OTCase family.' .
 _:188055 a rdf:Statement ;
   up:attribution _:188044 ;
   rdf:subject ?this ;
   rdf:predicate up:alternativeName ;
   rdf:object ?name1 .
 _:188056 a rdf:Statement ;
   up:attribution _:188045 ;
   rdf:subject ?gene2 ;
   rdf:predicate up:encodedBy ;
   rdf:object ?gene2 .
 _:188057 a rdf:Statement ;
   up:attribution _:188046 ;
   rdf:subject ?this ;
   rdf:predicate up:annotation ;
   rdf:object ?annotation3 .
 _:188058 a rdf:Statement ;
   up:attribution _:188047 ;
   rdf:subject ?this ;
   rdf:predicate up:annotation ;
   rdf:object ?annotation4 .
 _:188059 a rdf:Statement ;
   up:attribution _:188048 ;
   rdf:subject ?this ;
   rdf:predicate up:annotation ;
   rdf:object ?annotation5 .
 _:188060 a rdf:Statement ;
   up:attribution _:188049 ;
   rdf:subject ?this ;
   rdf:predicate up:classifiedWith ;
   rdf:object keyword:665 .
 _:188061 a rdf:Statement ;
   up:attribution _:188050 ;
   rdf:subject ?this ;
   rdf:predicate up:classifiedWith ;
   rdf:object keyword:808 .
 _:188062 a rdf:Statement ;
   up:attribution _:188051 ;
   rdf:subject ?this ;
   rdf:predicate up:classifiedWith ;
   rdf:object GO:0004070 .
 _:188063 a rdf:Statement ;
   up:attribution _:188052 ;
   rdf:subject ?this ;
   rdf:predicate up:classifiedWith ;
   rdf:object GO:0006221 .
 _:188064 a rdf:Statement ;
   up:attribution _:188053 ;
   rdf:subject ?case6 ;
   rdf:predicate up:annotation ;
   rdf:object ?annotation7 .
}
WHERE {
  ?this up:reviewed "false"^^xsd:boolean .
  #baseURI: http://purl.uniprot.org/unirule/MF_00001
  #Rule MF_00001 Created by:hamap on:2001-06-01 Modified by:ecastro on:2014-09-26
  VALUES ?supertaxon8 {taxon:2 taxon:2157}
  ?this up:organism ?thisOrganism ;
```

MF_00001 on your system

```
wget “http://mirror.easyname.ch/apache/
   jena/binaries/apache-jena-3.13.1.tar.gz”
tar -xzvf apache-jena-3.13.1.tar.gz
```

```
wget “ftp://ftp.expasy.org/databases/hamap/sparql/hamap.simple”
```

```
./interproscan.sh -dp -appl hamap “$YOUR_SEQ”
```

```
xsltproc to_rdf.xslt “$IP_OUT” > “$INPUT_FOR_HAMAP”
```

```
./bin/sparql  --data “$INPUT_FOR_HAMAP” /
              --query <(head -n 1 hamap.simple)
```

All rules on your system

Scan all your sequences:

```
./interproscan.sh -dp -appl hamap “$YOUR_SEQ”
```

```
xsltproc to_rdf.xslt “$IP_OUT” > “$INPUT_FOR_HAMAP”
```

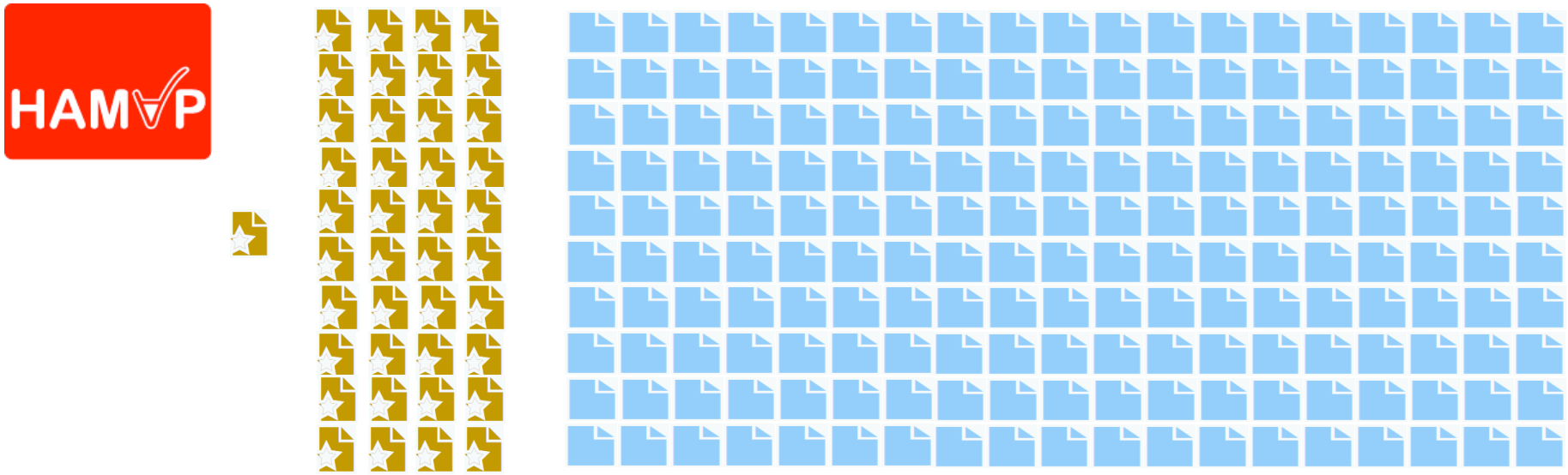
Download all rules from the FTP site and loop:

```
for rule in rules; do
  ./bin/sparql  --data “$INPUT_FOR_HAMAP” /
                --query ${rule}
done
```

Where/How to run

- **On your SPARQL endpoint**
- **In your RDBMS (via R2RML)**
- **On your cluster (via command line Sparql)**
- **SPARQL databases in the cloud**
    - **AWS Neptune**
    - **ORACLE PGX/Semnet**
    - **DB2**
    - **Pivotal/StarDog**
- **Documentation (HowTo, Examples)**

Band aid ⇝ annotates 14% to 64%



- **25 % *Escherichia coli* (most studied)**
- **64 % *Buchnera aphidicola* (small genome, core functions)**
  - **Mainly enzymes**

Acknowledgment

- **Biocuration**
  - **Andrea Auchincloss**
  - **Elisabeth Coudert**
  - **Chantal Hulo**
  - **Guillaume Keller**
  - **Patrick Masson**
  - **Ivo Pedruzzi**
  - **Catherine Rivoire**
- **Software Development**
  - **Delphine Baratin**
  - **Beature Cuche**
  - **Edouard de Castro**
- **Coordination**
  - **Alan Bridge**
  - **Nicole Redaschi**