Take a REST from manual searching

Programmatic access to Tools and Databases at EMBL-EBI

Andrew Cowley
andrew.cowley@ebi.ac.uk
support@ebi.ac.uk

EMBL-EBI

Contents

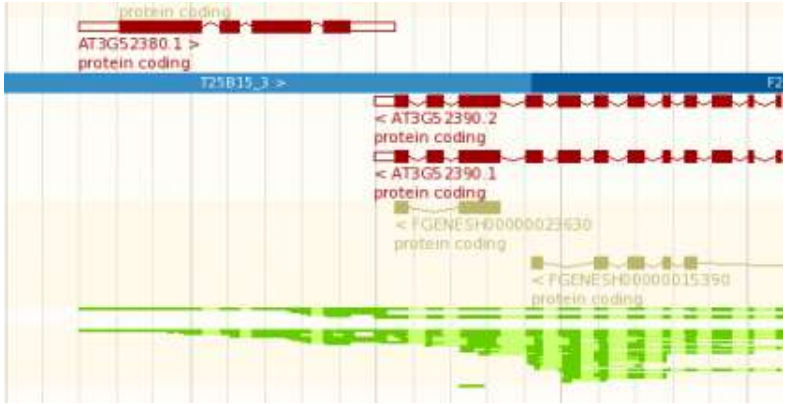
- What are we trying to do?
- Why consider programmatic access?
- How web services work
- How do you use web services (in practice)
- Clients and demo
- Tips and pitfalls
- Where to get help

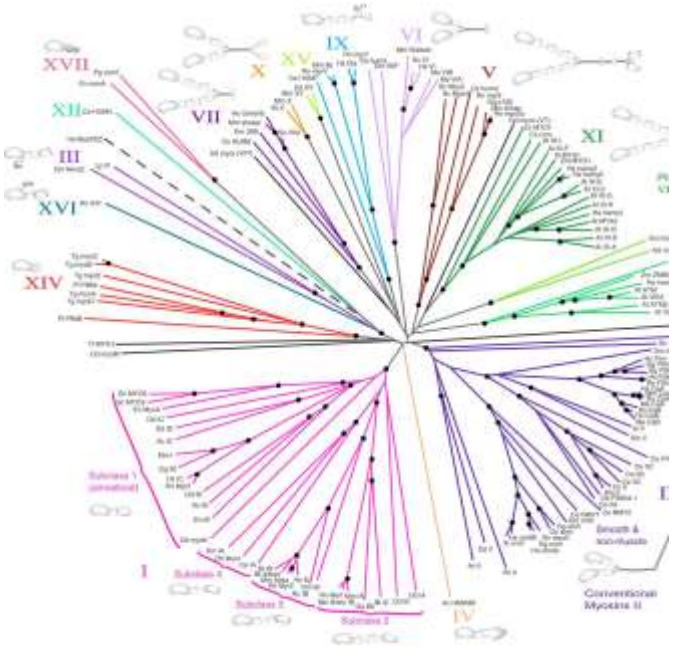
What are we trying to do?

- Bioinformatics!

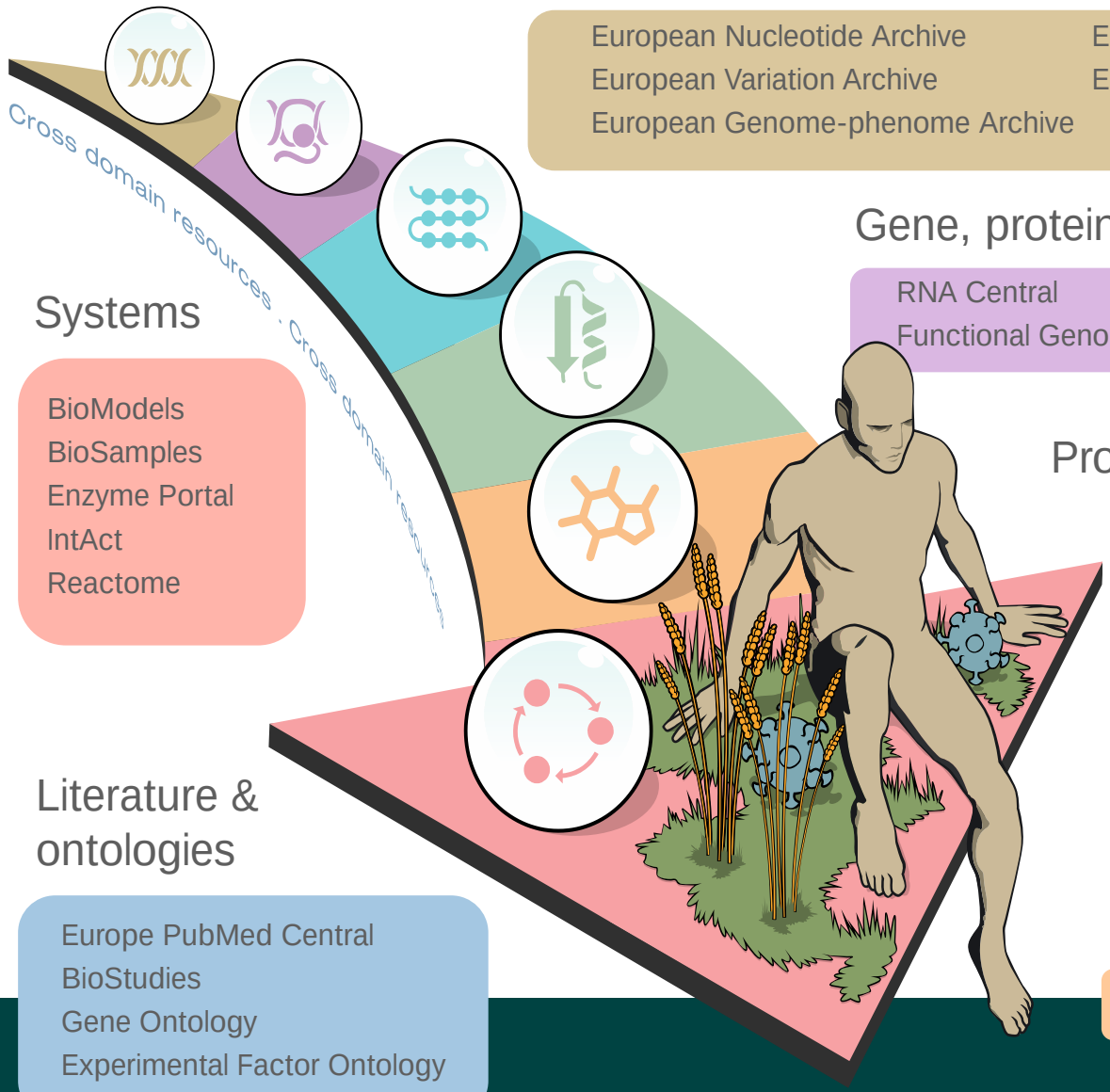
- Science of storing, retrieving and analysing large amounts of biological information

- From molecules to medicine
- Interpreting human variation
- Smarter farming etc.





Data resources at EMBL-EBI



Genes, genomes & variation

European Nucleotide Archive
European Variation Archive
European Genome-phenome Archive
Ensembl
Ensembl Genomes
GWAS Catalog
Metagenomics portal

Gene, protein & metabolite expression

RNA Central
Functional Genomics
Metabolights
PRIDE

Systems

BioModels
BioSamples
Enzyme Portal
IntAct
Reactome

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Literature & ontologies

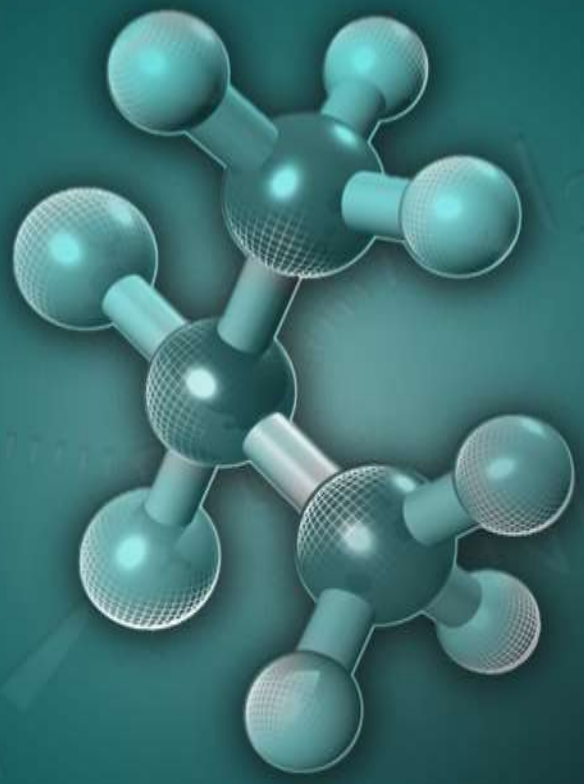
Europe PubMed Central
BioStudies
Gene Ontology
Experimental Factor Ontology

Chemical biology

ChEMBL
SureChEMBL
ChEBI

EMBL-EBI

Why consider programmatic access?



EMBL-EBI

Browser interface

- Easy interaction
- Visual input and results interfaces
- Access to the latest versions of softwares and data

Browser interface - disadvantages

- Can only run one task at a time
- Repetitive analysis is tedious
- Limited workflow capabilities

Local install

- Download tools and data locally
- Run as many tasks as you have power for
- Easy integration into workflows

Local install - disadvantages

- Expertise/privileges might be needed
- Local compute and storage requirements
- How do you keep up to date?

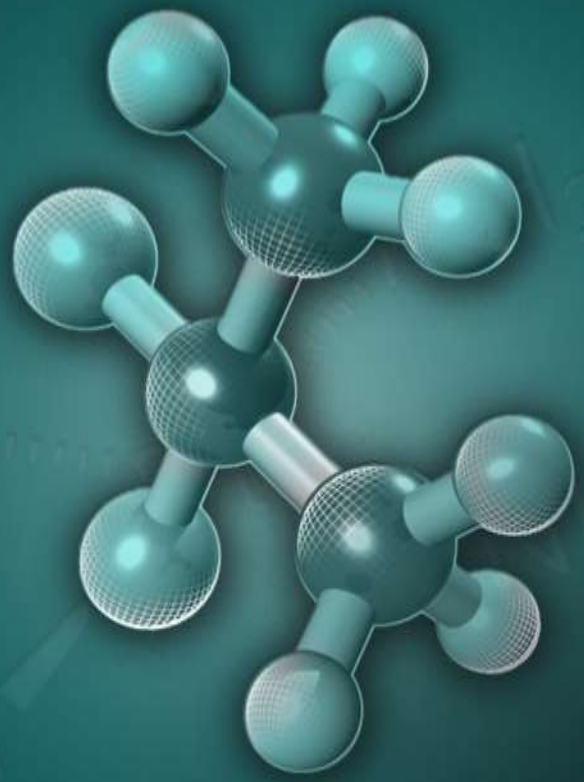
Another approach – programmatic access

- Interface with EMBL-EBI servers programmatically
- Use our compute, plus latest data
- Run many tasks simultaneously
- Easy to integrate in workflow/own website/frontend

Programmatic access - disadvantages

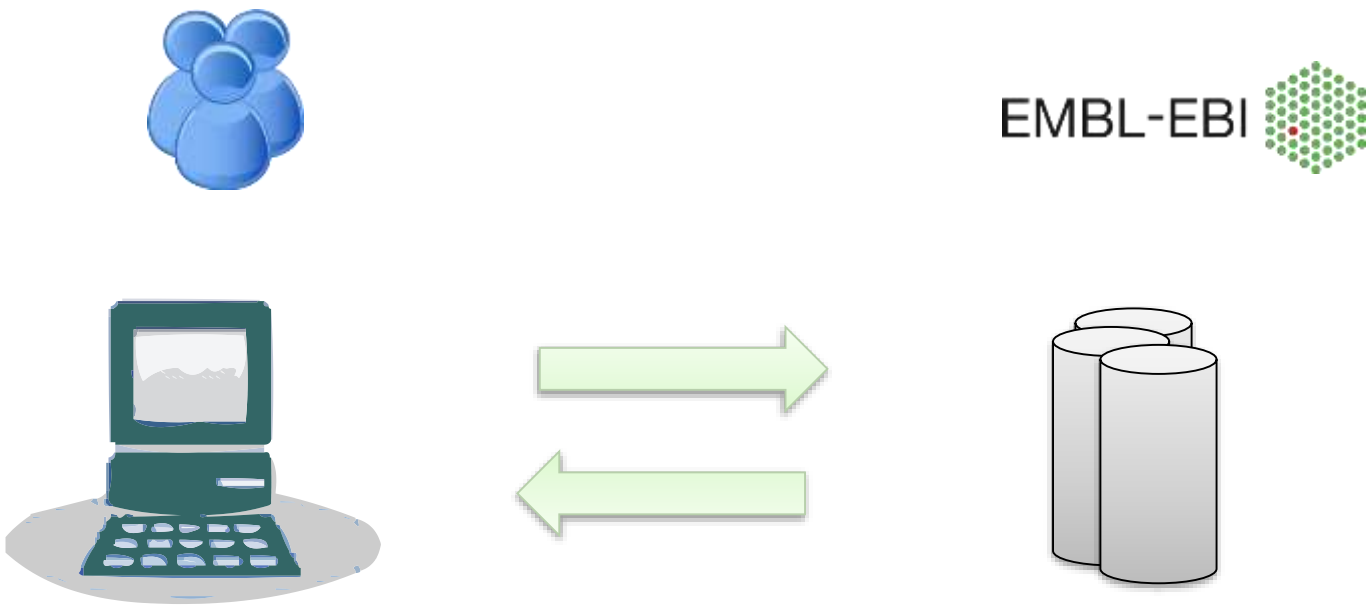
- Not unlimited access (though still more access than any other method)
- Little bit of programmatic knowledge still needed
- Still using our data

How does it work?

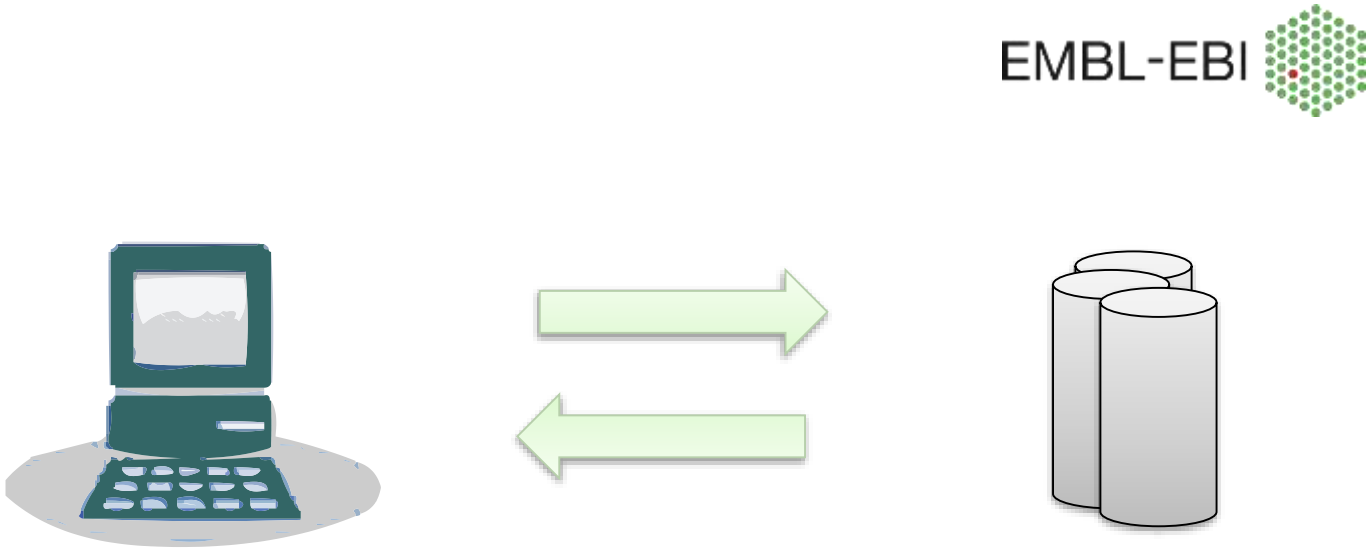


EMBL-EBI

How does it work?

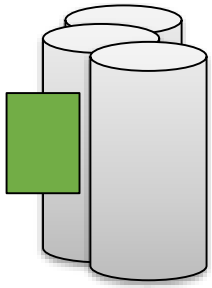


How does it work?



What are web services?

Web service = the server interface that responds to a defined request-response message system



Message systems

- Two main types of message systems

- SOAP
- REST

- Both (can) use HTTP as the protocol

SOAP

- Simple Object Access Protocol
- Wraps requests and responses in XML envelopes
- Definitions (eg parameters) described by a WSDL (Web Service Definition Language) for each service
- Historically popular, but now more commonly superseded by..

REST

- REpresentational State Transfer
- In most cases just uses URLs and HTTP verbs (GET, POST, PUT and DELETE)
- Lighter weight, quicker to process
- REST is actually a style, not a protocol, so services are described as RESTful, rather than implementing REST

REST examples

- EBI Search – search for entries in our complete collection
- https://www.ebi.ac.uk/ebisearch/ws/rest/?query=globin

https://www.ebi.ac.uk/ebisearch/ws/rest/?query=globin

This XML file does not appear to have any style information associated with it. The documen


```
- <result>
    <hitCount>151609</hitCount>
  - <domains>
    - <domain id="allebi">
        <hitCount>151609</hitCount>
      - <subdomains>
        - <domain id="genomes">
            <hitCount>59655</hitCount>
          - <subdomains>
            - <domain id="dgva">
```

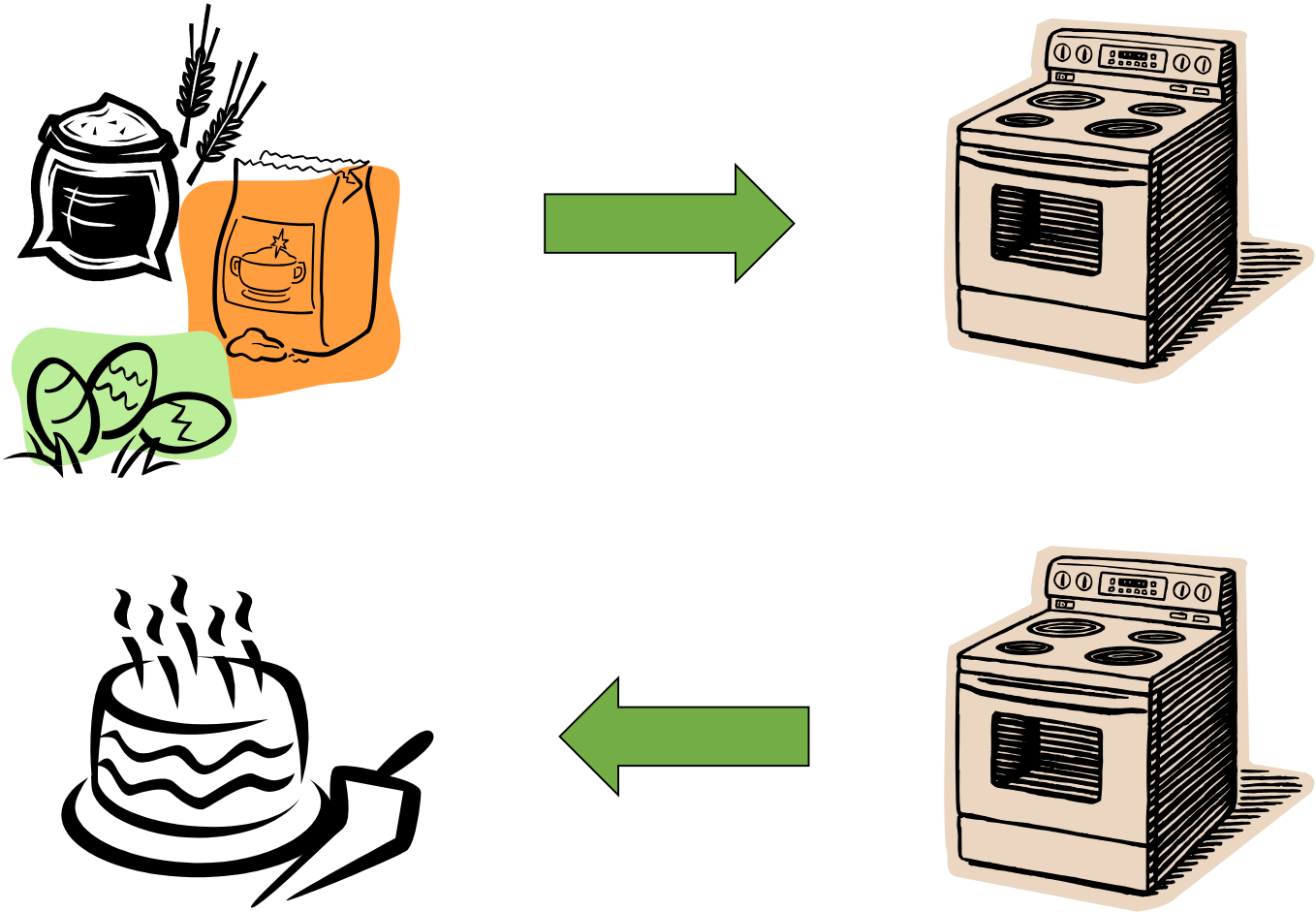
REST examples

- dbfetch – retrieve entries across our collection
- https://www.ebi.ac.uk/Tools/dbfetch/dbfetch/uniprotkb/WAP_RAT

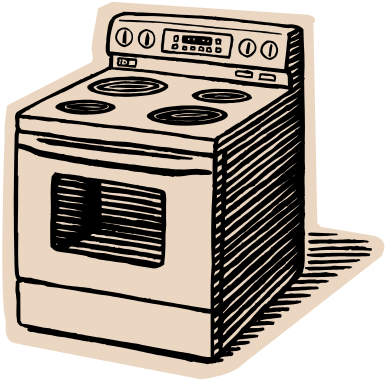
https://www.ebi.ac.uk/Tools/dbfetch/dbfetch/uniprotkb/WAP_RAT

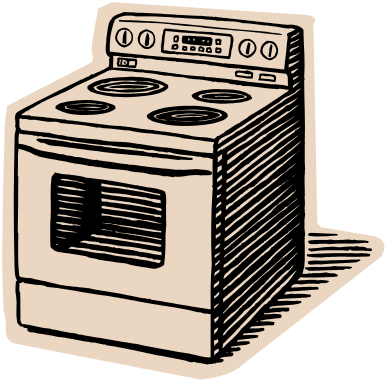
```
ID   WAP_RAT                 Reviewed;         137 AA.
AC   P01174;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1989, sequence version 2.
DT   12-APR-2017, entry version 117.
DE   RecName: Full=Whey acidic protein;
DE            Short=WAP;
DE   AltName: Full=Whey phosphoprotein;
DE   Flags: Precursor;
GN   Name=Wap;
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi;
OC   Muroidea; Muridae; Murinae; Rattus.
OX   NCBI_TaxID=10116;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   PubMed=6095207; DOI=10.1093/nar/12.22.8685;
RA   Campbell S.M., Rosen J.M., Hennighausen L.G., Strech-Jurk U.,
RA   Sippel A.E.;
```

Synchronous web services

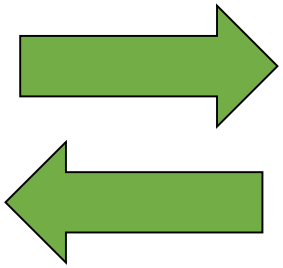


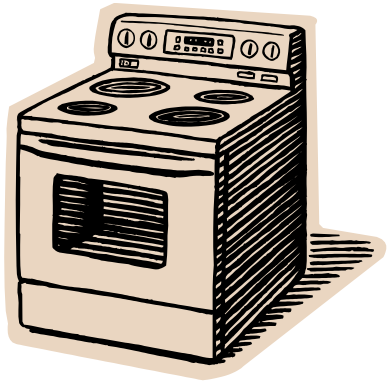
But if the task takes too long..





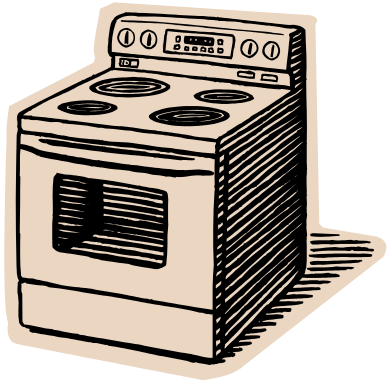
Asynchronous web services





JobID

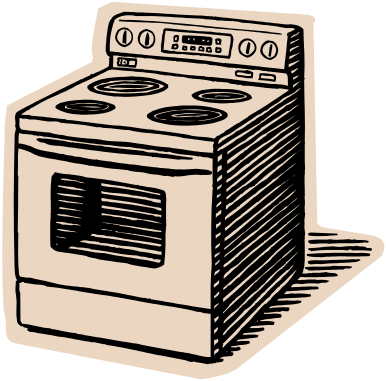
“Is JobID done?”



“Still cooking”

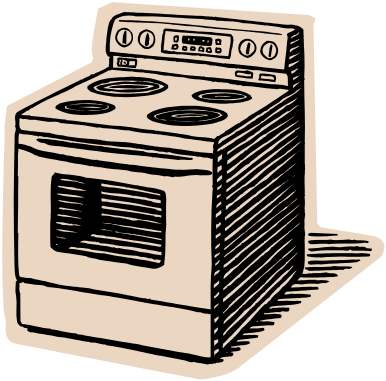
Asynchronous web services

“Is JobID done?”



“It’s ready!”

Retrieve JobID



How do I know which parameters to use?

- Look at documentation
- Use WSDL (for SOAP)
- But for REST? Might be a WADL – Web Application Description Language
- Query parameters programmatically

Querying parameters

- Many web services return details of parameters when queried

- https://www.ebi.ac.uk/Tools/services/rest/ncbiblast/parameters/

https://www.ebi.ac.uk/Tools/services/rest/ncbiblast/parameters/

This XML file does not appear to have any style information associated with it. The document tree is


```
- <parameters>
    <id>program</id>
    <id>task</id>
    <id>matrix</id>
    <id>alignments</id>
    <id>scores</id>
    <id>exp</id>
    <id>dropoff</id>
    <id>match_scores</id>
    <id>gapopen</id>
    <id>gapext</id>
    <id>filter</id>
    <id>seqrange</id>
    <id>gapalign</id>
    <id>compstats</id>
    <id>align</id>
```

Querying parameters

- https://www.ebi.ac.uk/Tools/services/rest/ncbiblast/parameterdetails/program/

https://www.ebi.ac.uk/Tools/services/rest/ncbiblast/parameterdetails/program/

This XML file does not appear to have any style information associated with it. The document tree i


```
– <parameter>
    <name>Program</name>
  – <description>
      The BLAST program to be used for the Sequence Similarity Search.
    </description>
    <type>COMMAND</type>
  – <values>
    – <value>
        <label>blastp</label>
        <value>blastp</value>
        <defaultValue>true</defaultValue>
      – <properties>
        – <property>
```

Querying parameters

- These can be built into a website that uses the Swagger framework

- Creates interactive documentation

EBI Search swagger

- https://www.ebi.ac.uk/ebisearch/swagger.ebi

All EBI search /

General information

Summary	Description	Method	Url
All EBI search	If a query parameter is specified, it will return the numbers of hits in a domain hierarchy. Otherwise, return meta-data of all domains available in EBI Search	GET	http://www.ebi.ac.uk/ebisearch/ws/rest/?query=globin

Response content type

application/xml

Parameters

SEND

Parameter name	Parameter value	Description	Data type	Parameter type
query	globin	Query string	String	Query

EBI Search swagger

- https://www.ebi.ac.uk/ebisearch/swagger.ebi

Curl

```
curl -X GET --header 'Accept: application/xml' 'http://www.ebi.ac.uk/ebisearch/ws/rest/?query=globin'
```

Request URL

```
http://www.ebi.ac.uk/ebisearch/ws/rest/?query=globin
```

Request Headers

```
{
    "X-EBI-StickySession": "true",
    "X-EBISearch-client": "ebinocle-webjs",
    "Accept": "application/xml"
}
```

Response Body (1953 ms)

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
  <result>
     <hitCount>151609</hitCount>
     <domains>
        <domain id="allebi">
           <hitCount>151609</hitCount>
           <subdomains>
              <domain id="genomes">
                 <hitCount>59655</hitCount>
```

Results

- Most web services also give a choice of results
- Different formats eg. Raw text, XML
- Images, identifiers etc.

Steps...

Meta-Information

- **List parameters**
- **Get parameter details**
  → *Name, description, values...*

Submission

- **Run** *(Email, title, values...)*
  → *Job Identifier*
- **Check status**
  → *RUNNING, FINISHED, ERROR...*

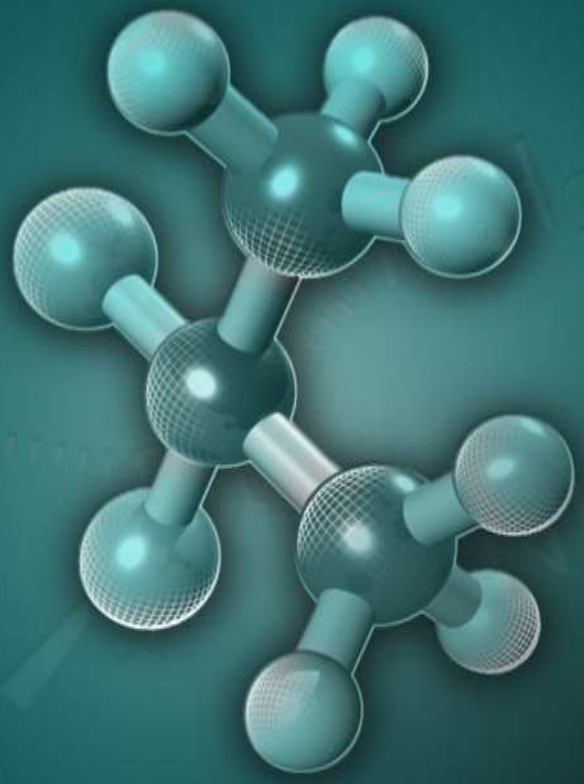
Results analysis

- **List results available**
  → *Name, description, media type...*
- **Get result**
  → *Output, text, binaries (images)...*

Input parameters

Job identifier
(e.g. iprscan-S20110708-094729-0726-35857540-pg)

How do you use them in practice?



EMBL-EBI

How do you use them in practice?

- Many ways to use them
- Generally incorporate the calls into a program or script
- Can be run from command line, or called from your own website, application or workflow
- Many EMBL-EBI services have example clients – can be used as a guide, or even by themselves

Clients

- Available for a range of programming languages
- Python, PERL, Java etc.
- Freely available to download, modify etc.

Sequence analysis tools web services

- Documentation and clients available at:

www.ebi.ac.uk/Tools/webservices

Demo

- BLAST search

NCBI BLAST+

Protein | Nucleotide | Vectors | Web services | Help & Documentation | Feedback

Tools > Sequence Similarity Searching > NCBI BLAST

Protein Similarity Search

The emphasis of this tool is to find regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of your novel sequence.

A new, more accurate, search tool combining optimal searching with iterative profile generation and over-extension error prevention is available using PSI-Search.

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected — X Clear Selection

- [x] UniProt Knowledgebase
- [ ] UniProtKB/Swiss-Prot
- [ ] UniProtKB/Swiss-Prot isoforms
- [ ] UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters
- Patents

Database selection

Sequence type

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

or upload a file: Browse... No file selected.

Sequence input

Program

STEP 3 - Set your parameters

PROGRAM

blastp

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Tools > Sequence Similarity Searching > NCBI BLAST+

Your job is currently running... please be patient

The result of your job will appear in this browser window.

Job ID: ncbiblast-I20170509-105525-0507-55904456-pg

Please note the following

- You may press Shift+Refresh or Reload on your browser at any time to check if results are ready.
- You may bookmark this page to view your results later if you wish.
- Results are stored for 7 days.

Results

Tools > Sequence Similarity Searching > NCBI BLAST

Results for job ncbiblast-I20150429-112458-0927-89262202-oy

Summary Table | Tool Output | Visual Output | Functional Predictions | Result Summary | Submission Details

Selection:
Select All | Invert | Clear

Apply to selection:

Annotations:
Show | Hide

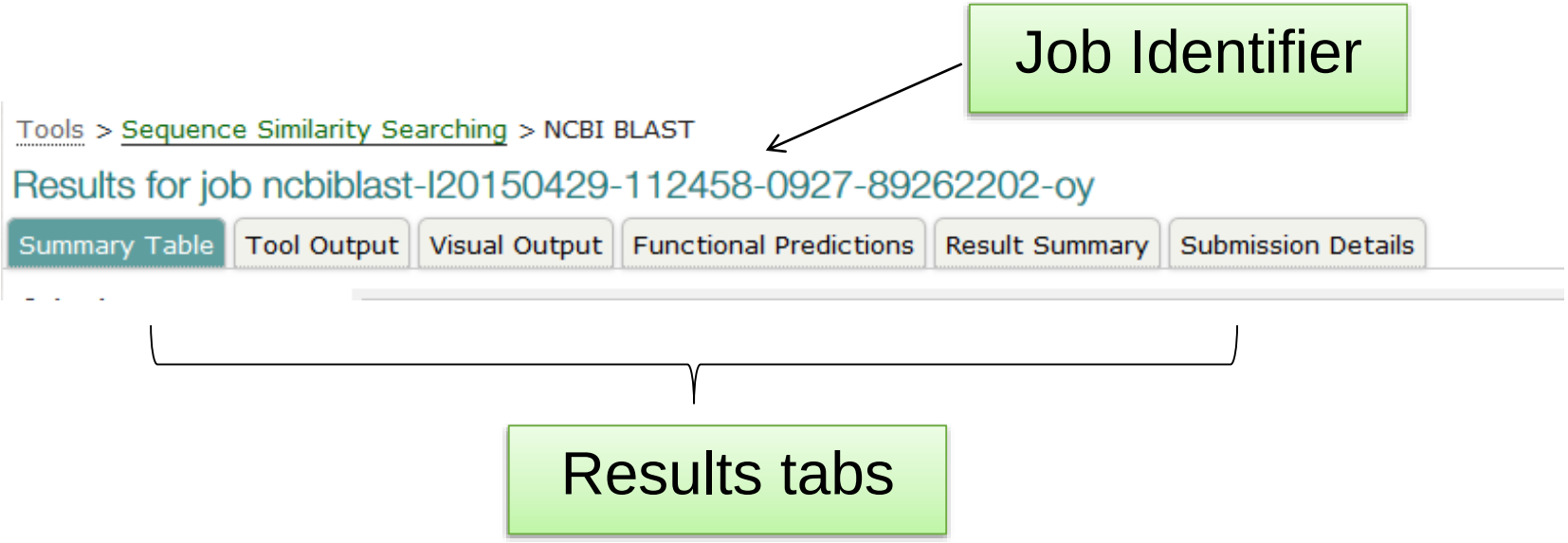
Alignments:
Show | Hide

Entries:
Download in fasta format

Tools:
Launch Clustal Omega

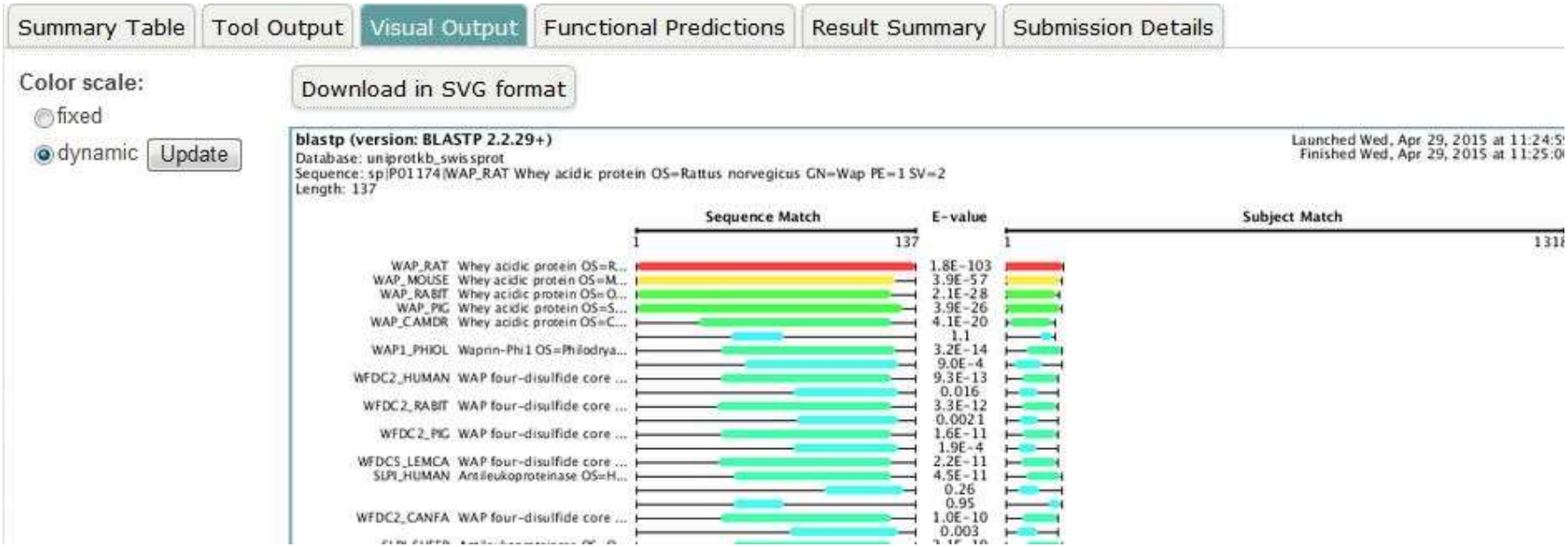
Align.	DB:ID	Source	Length	Score	Identities %	Positives %	E()
1	SP:WAP_RAT	Whey acidic protein OS=Rattus norvegicus GN=Wap PE=1 SV=2 *Cross-references and related information in:* ▶ Small molecules ▶ Nucleotide sequences ▶ Samples & ontologies ▶ Protein families ▶ Literature ▶ Protein expression data ▶ Protein sequences	137	763	100.0	100.0	1.8E-103
2	SP:WAP_MOUSE	Whey acidic protein OS=Mus musculus GN=Wap PE=1 SV=3 *Cross-references and related information in:* ▶ Small molecules ▶ Nucleotide sequences ▶ Samples & ontologies ▶ Protein families ▶ Literature ▶ Protein expression data ▶ Protein sequences	134	457	68.2	75.2	3.9E-57
3	SP:WAP_RABIT	Whey acidic protein OS=Oryctolagus cuniculus GN=WAP PE=2 SV=1 *Cross-references and related information in:* ▶ Small molecules ▶ Nucleotide sequences ▶ Genomes ▶ Samples & ontologies ▶ Protein families ▶ Literature ▶ Protein expression data ▶ Protein sequences	127	265	43.7	60.3	2.1E-28
4	SP:WAP_PIG	Whey acidic protein OS=Sus scrofa GN=WAP PE=1 SV=1	132	250	42.1	54.1	3.9E-26

Results


Job Identifier
Tools > Sequence Similarity Searching > NCBI BLAST
Results for job ncbiblast-I20150429-112458-0927-89262202-oy
Summary Table
Tool Output
Visual Output
Functional Predictions
Result Summary
Submission Details
Results tabs

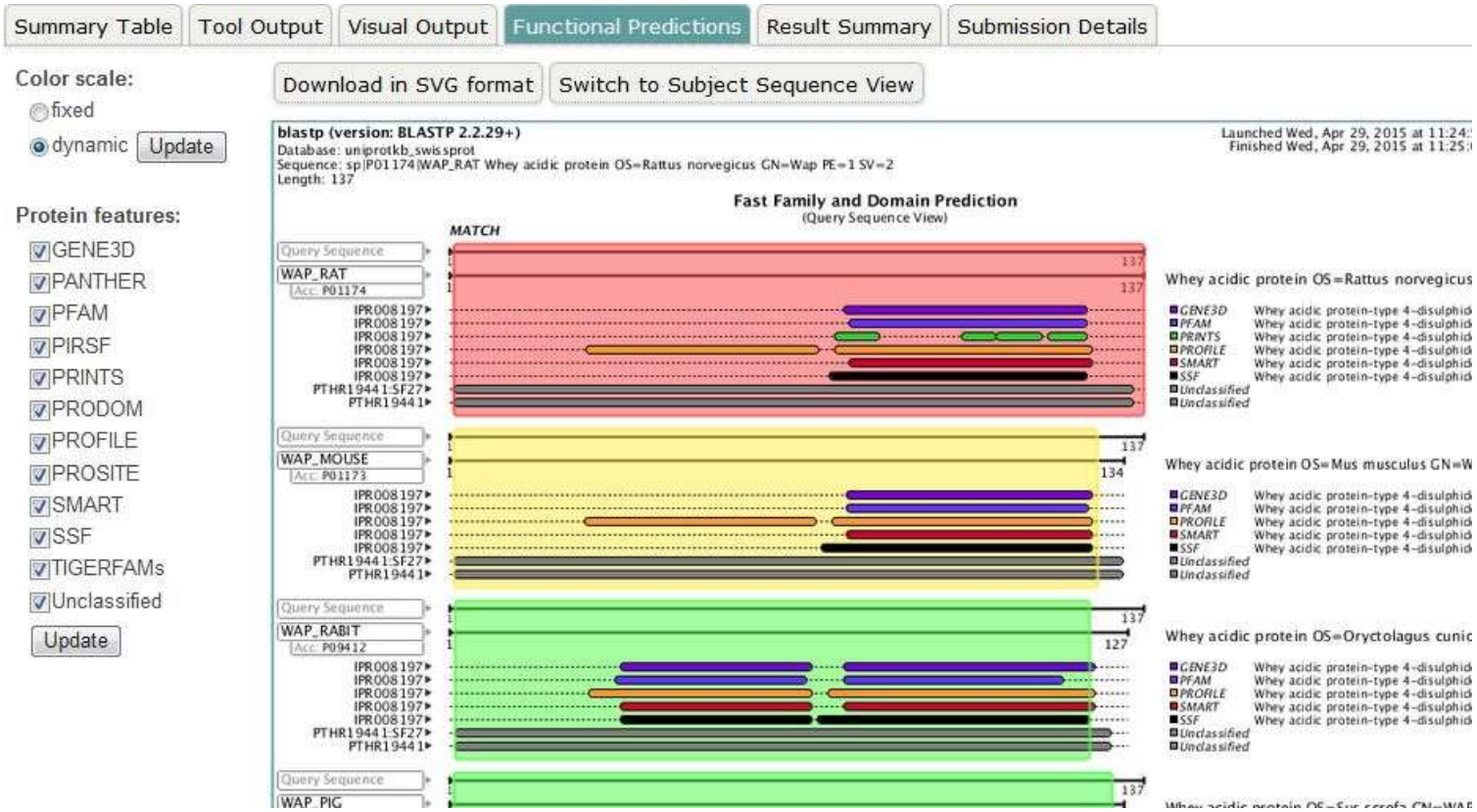
Results – Visual Output

- Shows where the alignment is occurring in the sequences



Results – Functional Predictions

- Domain and families predictions from InterPro



Now with web services..

- Download client from www.ebi.ac.uk/Tools/webservices

- Run without arguments to check usage/help

- Carry out search
  - Email address
  - Database (uniprotkb_swissprot)
  - Sequence type/stype (protein)
  - Program (blastp)
  - Input sequence

```
ebi-cli-002.ebi.ac.uk> ./ncbiblast_lwp.pl --email andrew.cowley@ebi.ac.uk --data
base uniprotkb_swissprot --stype protein --program blastp seq1.fsa
```

```
JobId: ncbiblast-R20170509-110531-0701-64935599-pg
RUNNING
FINISHED
```

```
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.out.txt
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.sequence.txt
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.ids.txt
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.xml.xml
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.visual-svg.svg
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.complete-visua
l-svg.svg
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.visual-png.png
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.complete-visua
l-png.png
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.visual-jpg.jpg
Creating result file: ncbiblast-R20170509-110531-0701-64935599-pg.complete-visua
l-jpg.jpg
```

Web Services workflow

```
$ncbiblast_lwp.pl --email email@example.org --program
blastp --database uniprotkb_human --stype protein
P01174.fasta
```

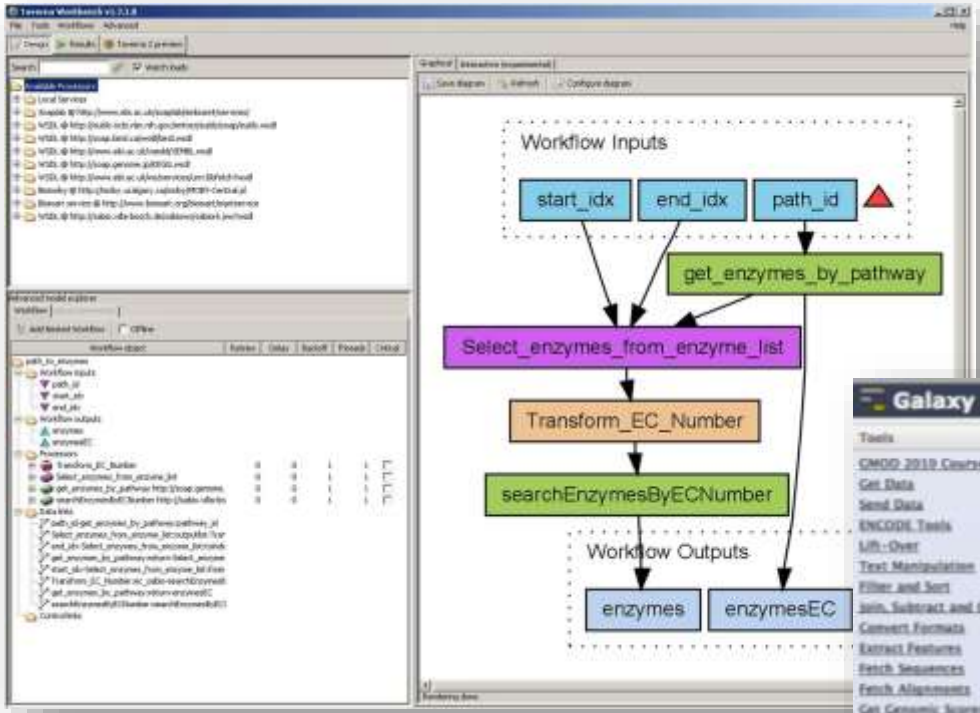
```
$wsdbfetch_soaplite.pl fetchData @<jobid>.ids.txt
fasta > P01174search2016_05_15.fasta
```

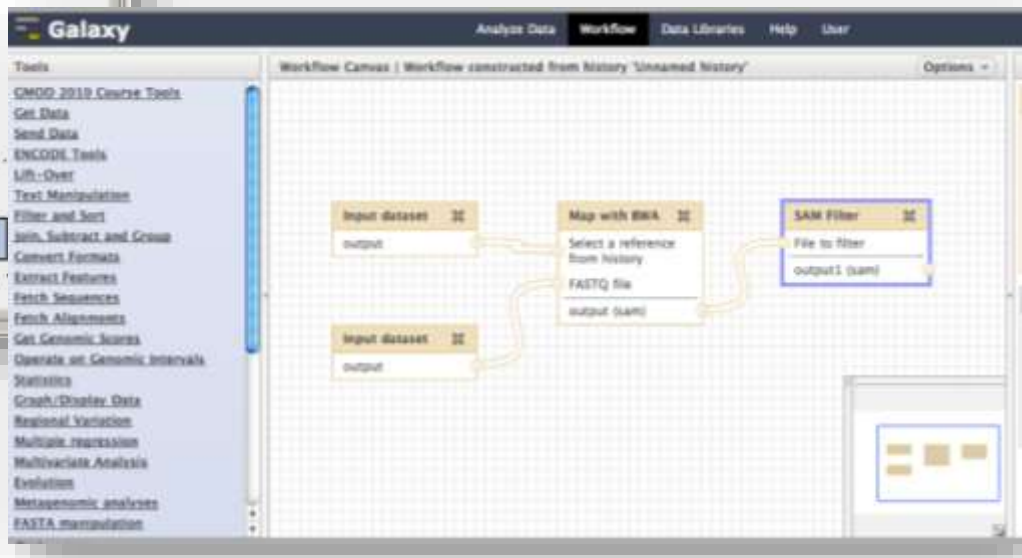
```
$kalign_soaplite.pl --email email@example.org
P01174search2016_05_15.fasta
```

Web Services workflow

```
$ncbiblast_lwp.pl --email email@example.org --program
blastp --database uniprotkb_human --stype protein
P01174.fasta | wsdbfetch_soaplite.pl fetchData @-
fasta | kalign_soaplite.pl --email email@example.org
-
```

Workflows

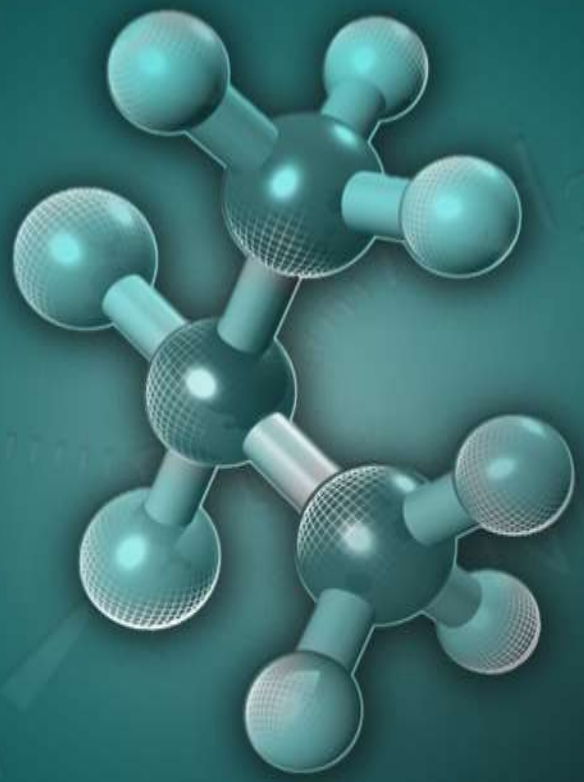




Taverna

Galaxy

Tips and pitfalls



EMBL-EBI

Parallelising and batch running

- Don’t go mad! Check terms of use
- Sequence analysis tools generally have limit of 30 simultaneous jobs
- Going beyond this can affect your job speed and if disruptive you may be blocked
- If using a third party tool, check if it’s already submitting multiples – eg BLAST2GO or runIPRscan

Parallelising and batch running

- Use script to automate new job submissions, iterating input files
  - When status FINISHED, submit new job.

- Use short cuts in clients to make things easier
  - Eg --multifasta flag - allows input to be a file containing multiple fasta format sequences, client will work through sequentially
  - Split input into 30 files, launch 30 --multifasta jobs

Bring back only the results you need

- By default all outputs are sent back, including graphics
- You can reduce space/transfer by just returning output of interest
  - Eg ID list, BLAST report etc.
  - --polljob --outformat ids --jobid <jobID>
- Check what result types are available
  - --resultTypes --jobid <jobID>

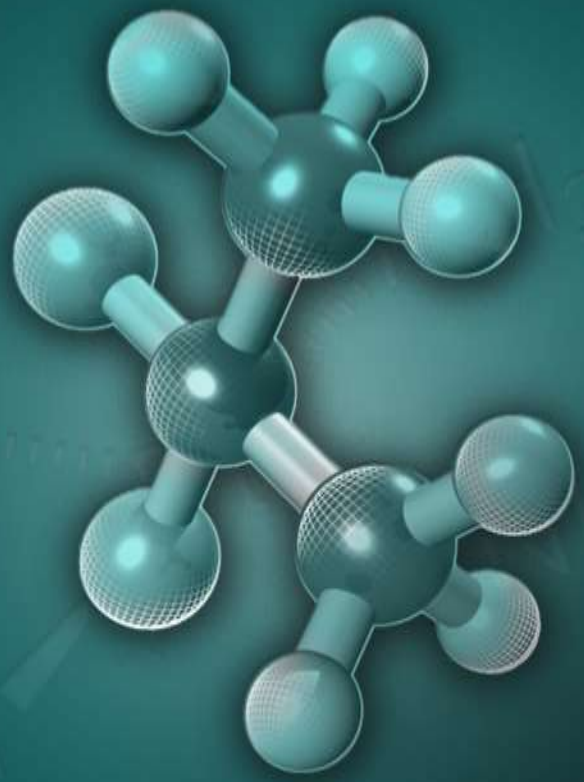
Error messages – three levels

- Client
  - May return error on invalid/missing parameters etc.
- Web service/server
  - May return error if problem with connectivity/bad http
  - Failed validation
- Tool
  - May return error if problem running/completing the job

Common errors

- NOT_FOUND
  - Check jobID
  - Check submission date – results only stored 7 days
- ERROR/FAILURE
  - Check parameters/input – validation might have failed
- FINISHED, but results include .error.txt file
  - Check error message
  - Check input – correct format, too large?
  - Check tool usage – right tool for the task?

Where to get help

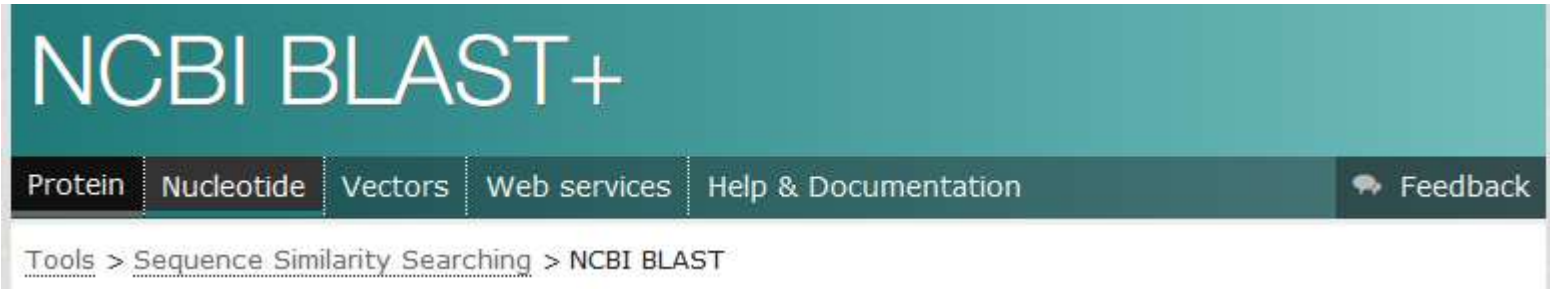


EMBL-EBI

Getting help

- Documentation available via Help & Documentation link
- Helpdesk available via Feedback button or

www.ebi.ac.uk/support/



- Current Protocols in Bioinformatics: **Unit 3.12** *Using EMBL-EBI Services via Web Interface and Programmatically via Web Services*

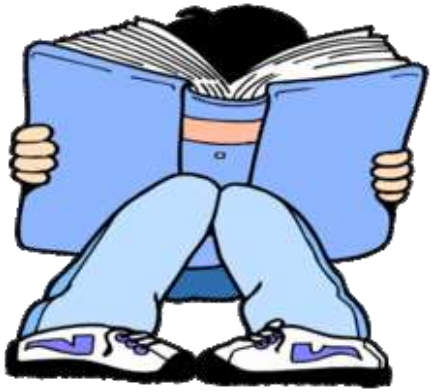
Citing use

Programmatic access to bioinformatics tools from EMBL-EBI update: 2017

Chojnacki S, Cowley A, Lee J, Foix A, Lopez R.

Nucleic Acids Res Web Server issue (2017)

DOI: 10.1093/nar/gkx273



Thank you!

Support: www.ebi.ac.uk/support/

DOI: 10.1093/nar/gkx273

DOI: 10.1002/0471250953.bi0312s48

DOI: 10.1093/nar/gkt376

Andrew Cowley

andrew.cowley@ebi.ac.uk
support@ebi.ac.uk

EMBL-EBI