EMBL-EBI, programmatically: take a REST from manual searches

David Armstrong [1], Melissa Burke [2], Andrew Cowley [3], Anna Foix [4], Joon Lee [5], Maria Levchenko [6], Matthew Laird [7], Michal Nowotka [8], Klemens Pichler [9]

- DNA & RNA
- Literature
- Proteins
- Structures
- Cross-domain

- Beginner
- Intermediate

- 2 hours

From BLAST searches to literature services, gene annotation and information on protein structure and function, large amounts of biological data can be accessed online with the click of a button.

When it comes to repeated tasks with large volumes of data (e.g. BLASTing a large number of sequences), the best way to get quick and automated access is to do it programmatically.

Between 9th May and 18th July 2017 we held a series of live webinars on how to access EMBL-EBI resources and tools programmatically. The webinars were recorded and combined to form a complete course.

The series includes:

- An introductory webinar that will introduce the concept of programmatic access, how to do it and best practice guidelines.
- Six webinars featuring Europe PMC [10], Ensembl [11], UniProt [12], PDBe [13], ChEMBL [14], and EMBL-EBI Sequence analysis tools [15] that demonstrate how to perform key biological
tasks programmatically. These webinars focus on REST-based APIs [16] but information and links to other methods will be provided.

The webinars include links to documentation, example scripts and live examples.

You do not need to watch all of the webinars; you can dip in and out and watch only the webinars you are interested in.

To watch the webinars simply choose the webinar of interest from the menu on the left of your screen.

Learning objectives:

- Gain an overview of EMBL-EBI webservices
- Describe processes that can be performed using EMBL-EBI RESTful APIs
- Learn where to find help, documentation and get support
- Access example scripts

**Introduction to EMBL-EBI resources**

At EMBL-EBI we have a huge variety of tools, resources and databases for sharing, accessing and analysing biological data. They cover many different topics from literature and ontologies, to systems, genes, genomes, proteins, metabolites and more (Figure 1). Importantly, these services are interlinked which helps you explore biological systems more easily.

![EMBL-EBI's resources and databases](image)

**Figure 1** EMBL-EBI's resources and databases
You can explore all of the EMBL-EBI's tools, resources and databases on our [services page](http://www.ebi.ac.uk/training/online) [17].

In [Train online](http://www.ebi.ac.uk/training/online) [18] you will find quick tours, tutorials and webinars that introduce each of their resources and show you how to use them.

The tools and resources can be accessed online or programmatically, the latter of which is very useful when accessing large volumes of data and building analytic workflows.

EMBL-EBI web services provides a [list of tools and resources that can be accessed programmatically](http://www.ebi.ac.uk/training/online) [19] with links to the relevant help pages and documentation.

In this webinar series we introduce the basic principles of programmatic access and take a more detailed look at how to access key resources programmatically.

### Introduction to programmatic access

Biology has entered the digital age and it is now possible to access and analyse many types of data online. Using EMBL-EBI tools and resources you can share your data, perform complex queries, and analyse data in many different ways. You can [BLAST](http://www.ebi.ac.uk). Nucleotide sequences, search the literature, look up variant data, learn about proteins, explore macromolecular molecules, and much, much more.

One way to make these kinds of bioinformatics tasks quicker and easier is to do them programmatically using [web services](http://www.ebi.ac.uk) [21].

This webinar took place on 9th May 2017 and introduces the concept of web services and how you can use them to access the tools and data available at EMBL-EBI.

- What are we trying to do? (0:41)
- Why consider programmatic access? (1:29)
- How web services work (4:52)
- How do you use web services (in practice) (13:56)
- Clients and demo (17:35)
- Tips and pitfalls (23:36)
- Where to get help (31:56)
You can explore all of the EMBL-EBI's tools, resources and databases on our [services page](http://www.ebi.ac.uk/training/online) [17]. On the services page you will also find a [list of tools and resources that can be accessed programmatically](http://www.ebi.ac.uk/training/online) [19] with links to the relevant help pages and documentation.

If you need help with any of the tools and resources you can [contact us](http://www.ebi.ac.uk/training/online) [22] or get in touch with resource helpdesk.

You can download the slides from this webinar [here](http://www.ebi.ac.uk/training/online) [23].

---

**Europe PMC, programmatically**

Europe PMC is an open repository, providing access to worldwide biomedical literature, including articles, books, patents, and clinical guidelines. More than 32 million abstracts (27 million from [PubMed](http://www.ncbi.nlm.nih.gov/pubmed) [24]) and 4 million full-text articles can be freely accessed through Europe PMC in a single search. The mission of Europe PMC is to build [open access](http://www.ebi.ac.uk/training/online) [25], full text scientific literature resources and deliver world-class literature services. You can learn more about Europe PMC in our [Europe PMC: Quick tour](http://www.ebi.ac.uk/training/online) [26].

This webinar took place on 23rd May 2017. It introduces the Europe PMC Articles RESTful [API](http://www.ebi.ac.uk/training/online) [16] and show you how to access publications and related information, such as citations, data links, or text-mined terms, programmatically. It includes:

- Tools and data overview (1:22)
- RESTful web service (5:08)
- Using RESTful web service: examples (8:07)
- RESTful web service output (14:37)
- Literature search: tips and suggestions (19:45)
- Related resources (24:09)
Europe PMC provides programmatic methods for accessing:

- **Publications and related information** [27].
- **Grant information from Europe PMC funders using the GRIST database** [28].
- **Publication data using a Simple Object Access Protocol (SOAP)-based service** [29].
- **Open Archives Initiative (OAI) Protocol for retrieving metadata of all items in Europe PMC** [30].

For documentation, help and support visit the [Europe PMC help pages](http://www.ebi.ac.uk/training/online) [31] or download the [developer friendly web service guide](http://www.ebi.ac.uk/training/online) [32]. For web service related question you can get in touch via the [Google group](http://www.ebi.ac.uk/training/online) [33] or contact the [helpdesk](http://www.ebi.ac.uk/training/online) [34].

You can [download the slides](http://www.ebi.ac.uk/training/online) [35] from this webinar here.

### Ensembl, programmatically

[Ensembl](http://www.ebi.ac.uk/training/online) [36] provides a genome browser that acts as a single point of access for annotated genomes. Genes, genetic variants and **regulatory features** [37] are all annotated onto genomes, with comparative genomics analyses carried out on genes and genomes. The [Ensembl REST API](http://www.ebi.ac.uk/training/online) [38] provides a number of endpoints for flexible programmatic access to all these data. You can learn more about Ensembl in our [Ensembl: Quick tour](http://www.ebi.ac.uk/training/online) [39].

This webinar took place on 6th June 2017 and was hosted by Matthew Laird. The webinar includes:

- What is Ensembl? (0:28)
- Methods to access Ensembl data (2:00)
- The Ensembl REST server documentation. (2:19)
- How to query the REST server. (3:32)
- Ensembl REST API endpoints (4:47)
  - What does a HTTP response look like? (7:52)
  - Content types (9:37)
Chaining requests and decoding the response (10:10)
- Live example - fetch a gene based on symbol (11:36)
- HTTP methods - POST requests (13:03)
- Where to go for help and documentation (16:47)

You can find out more about Ensembl REST API endpoints [38] and other Ensembl APIs [16] on the Ensembl help pages [40]. You can also find a user guide and sample clients on GitHub [41].

For help and support with Ensembl get in touch with the Ensembl help desk [42].

You can download the slides [43] and the scripts used for the gene symbol [44] and loop [45] examples using these links.

**UniProt, programmatically**

*UniProt* [12] aims to provide the scientific community with the highest quality, most comprehensive and most thoroughly annotated protein resource. Our mission includes curating detailed information on protein function and sequences including isoforms and disease variants. The *UniProt knowledgebase* [46] contains millions of sequences, a subset of which have been curated by experts who critically review experimental and predicted data for each protein. The remainder are automatically annotated based on rule systems that rely on the expert curated knowledge. You can learn more about UniProt in *UniProt: Exploring protein sequence and functional information* [47] or our introductory webinar [48].

This webinar took place on 20th June 2017 and will introduce the UniProt REST [49] API [16]. In this webinar we will:

- Introduce the UniProt resources (1:12)
- Search the UniProt Knowledgebase (3:40)
- Convert between different database identifiers (11:43)
- Search with proteomics [50] peptides (16:55)
- Show you where to get help (21:47)

You can download the slides [51] and the scripts in PDF [52] or in executable Jupyter Notebook [53].
format.

You can find more information and documentation on [how to access UniProt programmatically](#) and the [Proteins API](#) on our website.

For help and support visit the [UniProt help pages](#) or contact the [help desk](#).

**PDBe, programmatically**

The [Protein Data Bank in Europe (PDBe)](#) is a member of the Worldwide Protein Databank (wwPDB), the organisation tasked with the collection, organisation and dissemination of data about biological macromolecular structures. PDBe aims to ‘bring structure to biology’ and provides a comprehensive range of tools and services to make the extensive archive of structural and functional information in the [PDB](#) more accessible to the wider biological community. You can learn more about PDBe in our webinar [PDBe: Searching for biological macromolecular structures](#).

This webinar took place on 4th July 2017 and introduced the PDBe [REST](#) [API](#) service and show you how to access and analyse macromolecular structural data programmatically.

This webinar includes sections that focus on:

- Introduction to the data at PDBe (1:25)
- PDBe's RESTful API (3:33)
- Finding additional annotation programmatically (6:30)
- Advanced searching programmatically (11:29)
- A simple example script (16:19)

You can download the [slides](#), the [simple script](#) and an [extra example script](#) by using these links.

PDBe’s REST API is a programmatic way to obtain information from the PDB and [EMDB](#). You can access details about sample, experiment, models, compounds, cross-references, publications, quality, assemblies and more.

Have a look at the following links for help and documentation:

- [PDBe API documentation](#)
- [Frequently asked questions (FAQ)](#)
- [Get started - worked out examples](#)
ChEMBL, programmatically

ChEMBL is a database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data). The data is abstracted and curated from the primary scientific literature, and covers a significant fraction of the SAR and discovery of modern drugs. We attempt to normalise the bioactivities into a uniform set of end-points and units where possible, and also to tag the links between a molecular target and a published assay with a set of varying confidence levels. Additional data on clinical progress of compounds is being integrated into ChEMBL at the current time.

This webinar took place on 12th July 2017. It provides:

- An overview of the ChEMBL API and its use, including how to execute API calls from the browser (0:23);
- Where to find documentation (2:54);
- How to user filtering and pagination (6:23);
- Available output formats (17:13);
- Scripting examples in Python, Bash and R (27:38).

We also give examples of how the API can be used to create reusable web components and integrated into tools such as KNIME and Slack. The webinar assumes a degree of familiarity with the data in ChEMBL. If you are not famililar with ChEMBL we recommend watching our ChEMBL walkthrough: webinar [68] first.

You can download the slides (including links to example scripts) here [69].

Have a look at the following links for help and documentation:

- ChEMBL Frequently Asked Questions [70]
- ChEMBL web services and documentation [71]
- ChEMBL web services live documentation explorer [72]
- Source code via GitHub [73]

Or get in touch with the ChEMBL helpdesk [chembl-help[at]ebi.ac.uk].

Sequence analysis tools, programmatically
Providing access to core analytical tools and protein and nucleotide sequence databases is central to the EMBL-EBI's mission. Tools such as InterProScan [74], BLAST [20]+, PSI-SEARCH, Clustal [75] Omega and HMMER3 are but a few of more than 60 sequence analysis tools accessible using SOAP [76] and RESTful API [16] Web Services [21]. There are client applications written in a variety of languages including Python, Perl [77] and Java for many of these tools which can be integrated directly into your bioinformatics pipelines.

This webinar took place on 18th July 2017. It introduces these client applications and demonstrates how they can be used to form simple but effective workflows. It focuses on clients that demonstrate the RESTful API to tools such as blast+ and Clustal Omega, and data retrieval tools such as dbfetch.

This webinar is for anyone familiar with running bioinformatics tools from the command-line (Linux terminal, Windows CMD or PowerShell).

The webinar includes:

- An introduction to EMBL-EBI tools RESTful clients (0:50)
- Where to get the clients (2:52)
- How to run the clients (5:43)
- How to run a simple pipeline manually using a shell script and CWL (14:05)
- Where to get help and support (20:31)

You can download the slides [78] and example scripts [79] using these links.

You can explore all of the EMBL-EBI's tools, resources and databases on our services page [17].

On the services page will also find a list of tools and resources that can be accessed programmatically [19] with links to the relevant help pages and documentation.

Get in touch [80] with us if you need help with any of the tools and resources.

**Summary**

Biology has entered the digital age. The ability to search, access and analyse data programmatically is one way that we can make biological research and bioinformatics more efficient.
We hope that this webinar series has given you a taste of some of the resources and tools available at the EMBL-EBI and how to access them programmatically. Of course, we were not able to fit all of our resources or all of our programmatic methods into this series.

If you want to know more you can:

- Browse all of the available tools and resources and find out how to access them programmatically on our webservices page.
- Explore other tutorials and webinars in Train online that show you how to use many of our resources.

If in doubt, you can always get in touch and we will do our best to help.

Your feedback

Please tell us what you thought about this Quick tour. Your feedback is invaluable and helps us to improve our courses and enhance your learning experience.

Contributors

This webinar series was developed as a collaboration between multiple EMBL-EBI resources and the EMBL-EBI Training Team.

The working group for this project included a wide variety of people from many different backgrounds (in alphabetical order):

**ChEMBL**: Anna Gaulton, Michal Nowotka

**Ensembl**: Matthew Laird, Ben Moore, Emily Perry, Helen Sparrow

**Europe PMC**: Jyothi Katuri, Maria Levchenko

**PDBe**: David Armstrong, John Berrisford, Matt Conroy

**Protein Families**: Rob Finn

**Training team**: Melissa Burke, Sarah Morgan

**UniProt**: Leyla Jael Garcia Castro, Klemens Pichler, Sangya Pundir

**Web services**: Andrew Cowley, Anna Foix, Joon Lee, Rodrigo Lopez

Special thanks must go to those who volunteered to present the webinars (in order of appearance):
Andrew Cowley [3]

EMBL-EBI
Bioinformatics Trainer - Web production

Andrew studied Biochemistry at Cambridge before moving to York University for post-graduate studies in Bioinformatics and then Structural Bioinformatics, in particular working on the identification of distinctive features of protein structure. He then worked as a Bioinformatics specialist for BBSRC Bioscience IT Services before taking up his current position in the External Services team at the EBI where he provides training, support and documentation.

Maria Levchenko [6]

EMBL-EBI
Europe PMC Community Manager - McEntyre team: Literature services

Matthew Laird [7]

EMBL-EBI, programmatically: take a REST from manual searches
Published on EMBL-EBI Train online (http://www.ebi.ac.uk/training/online)
Ensembl Core Developer - Zerbino team: Genome Analysis

Klemens Pichler [9]
EMBL-EBI
Scientific Database Curator - Protein Function Content

David Armstrong [1]
EMBL-EBI
Scientific Database Curator - Kleywegt team: Protein Data Bank in Europe Scientific Database Curator - Velankar team: PDBe content and integration

Michal Nowotka [8]
EMBL-EBI
Technical Lead - Leach team: ChEMBL

Anna Foix [4]
EMBL-EBI
Software Engineer - Web production

Joon Lee [5]
EMBL-EBI
Full Stack Software Engineer - Web production

Source URL: http://www.ebi.ac.uk/training/online/course/embl-ebi-programmatically-take-rest-manual-searches

Links
[1] http://www.ebi.ac.uk/training/online/trainers/davida_1
EMBL-EBI, programmatically: take a REST from manual searches
Published on EMBL-EBI Train online (http://www.ebi.ac.uk/training/online)

y_py3.ipynb
[54] http://www.uniprot.org/help/programmatic_access
[55] https://www.ebi.ac.uk/proteins/api/doc/
[57] http://www.uniprot.org/contact
[58] http://www.ebi.ac.uk/pdbe/
[59] http://www.ebi.ac.uk/training/online/glossary/pdb
[60] http://www.ebi.ac.uk/training/online/course/pdbe-searching-biological-macromolecular-structure
[61] http://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/PDBe_webinar_presentation.pptx
[62] http://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/PDBe_really_simple_search_example.py.txt
[63] http://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/PDBe_get_data_from_api.py.txt
[64] http://www.ebi.ac.uk/training/online/glossary/emdb
[65] http://PDBe API documentation
[66] http://Frequently asked questions (FAQ)
[67] https://github.com/PDBeurope/PDBe_Programming
[68] http://www.ebi.ac.uk/training/online/course/chembl-walkthrough-webinar
[69] http://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/ChEMBL%20programmatically.pptx
[70] https://www.ebi.ac.uk/chembl/faq
[71] https://www.ebi.ac.uk/chembl/ws
[72] https://www.ebi.ac.uk/chembl/api/data/docs
[73] https://github.com/chembl/chembl_webservices_2
[74] http://www.ebi.ac.uk/training/online/glossary/interproscan
[75] http://www.ebi.ac.uk/training/online/glossary/clustal
[76] http://www.ebi.ac.uk/training/online/glossary/soap
[77] http://www.ebi.ac.uk/training/online/glossary/perl
[78] http://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/Sequence%20analysis%20tools%20programmatically.pptx
[79] http://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/Sequencetools_example_script.zip
[80] http://www.ebi.ac.uk/support