Bioinformatics for the terrified

This course will give you a broad overview of how bioinformatics can enable bench-based research. It is aimed at experimental researchers in the molecular life sciences who have little or no previous experience of using bioinformatics databases or tools.

An undergraduate degree in a subject related to the molecular life sciences would be an advantage.

Image courtesy of Wikimedia Commons [3].

This course was last updated in March 2018.

Learning objectives:

- Assess the role of bioinformatics in molecular life-science research
- Describe the key features of primary and secondary databases
- List strategies for describing data consistently
- Identify some of the different types of data analysis that can be applied to solving biological problems
What is bioinformatics?

Put simply, bioinformatics is the science of storing, retrieving and analysing large amounts of biological information. It is a highly interdisciplinary field involving many different types of specialists, including biologists, molecular life scientists, computer scientists and mathematicians.

The term bioinformatics was coined by Paulien Hogeweg and Ben Hesper to describe "the study of informatic processes in biotic systems" and it found early use when the first biological sequence data began to be shared. Whilst the initial analysis methods are still fundamental to many large-scale experiments in the molecular life sciences, nowadays bioinformatics is considered to be a much broader discipline, encompassing modelling and image analysis in addition to the classical methods used for comparison of linear sequences or three-dimensional structures (Figure 1).

![Figure 1](image-url) A broad overview of the different types of data that fall within the scope of bioinformatics. Traditionally, bioinformatics was used to describe the science of storing and analysing biomolecular sequence data, but the term is now used much more broadly, encompassing computational structural biology, chemical biology and systems biology (both data integration and the modelling of systems).

Distinction from medical informatics

Bioinformatics is distinct from medical informatics – the interdisciplinary study of the design, development, adoption and application of IT-based innovations in healthcare services delivery, management and planning. Somewhere in between the two disciplines lies biomedical informatics – the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific enquiry, problem solving.
and decision making, motivated by efforts to improve human health.

Recently initiated projects, such as the 100,000 Genomes Project [4], are bridging the gaps between these disciplines, but on the whole bioinformatics deals with research data and uses it for research purposes, medical informatics deals with data from individual patients for the purposes of clinical management, (diagnosis, treatment, prevention...) and biomedical informatics attempts to bridge these two extremes.

Who is bioinformatics for?

Leon, postdoc
Goal: to understand what makes a normally harmless bacterium pathogenic in the lungs of people with cystic fibrosis.
Tasks: “I’m using a combination of transcriptomics, proteomics and metabolomics to understand these pathogenic changes better.”

Barend, plant geneticist
Goal: to identify new crop strains resistant to drought, salt and fungal diseases.
Tasks: “We’re doing linkage studies to find out which genes are involved in resistance to different types of stress. We’ve got genomic and expression QTLs that we need to map on to well-characterised plants.”

Ola, clinician–scientist
Goal: to identify proteomics-based biomarkers in urine for the early detection of bladder cancer
Tasks: “I do mass spectrometry of samples from patients coming in for biopsies. I’ve found a phosphoprotein that seems to be upregulated in some patients.”

Figure 2 The (fictional) personas illustrated here, whilst not bioinformaticians in the classical sense, all use bioinformatics to enable their research. Cartoons courtesy of Jenny Cham, EMBL-EBI.

The molecular life sciences have become increasingly data driven by and reliant on data sharing through open-access databases (1 [5]). This is as true of the applied sciences as it is of fundamental research. Furthermore, it is not necessary to be a bioinformatician to make use of bioinformatics databases, methods and tools. However, as the generation of large data-sets becomes more and more central to biomedical research, it’s becoming increasingly necessary for every molecular life scientist to understand what can (and, importantly, what cannot) be achieved using bioinformatics, and to be able to work with bioinformatics experts to design, analyse and interpret their experiments (Figure 2).

The role of public databases

Bioinformatics centres of excellence

There are a small number of bioinformatics centres of excellence worldwide that have taken on the responsibility to
collect, catalogue and provide open access [6] to published biological data (Figure 3). Among these centres are:

- The EMBL-European Bioinformatics Institute [7] (EMBL-EBI)
- The US National Center for Biotechnology Information [8] (NCBI)
- The National Institute of Genetics [9] in Japan (NIG)

This work began in the early 1980s when DNA sequence data began to accumulate in the scientific literature. The EMBL Data Library (now the European Nucleotide Archive [10]) was developed to store DNA sequences published in the scientific literature. The NCBI's GenBank and NIG's DDBJ followed.

**Figure 3** The role of bioinformatics centres of excellence in making biological data available for the research community.

**Data sharing collaborations**

As the number of published sequences increased, the workflow changed: by opening discussions with publishers of the scientific literature, the organisations behind these databases convinced publishers to request that researchers submit their sequences to one of the public databases before submitting the paper. In return, authors gained an accession [11] number that could then be cited in the paper. This model has now been followed by many providers of public biological data. Table 1 provides some significant examples.

The other important aspect of these collaborations is that their participants exchange data and/or assign workload in such a way as to avoid duplication of effort whilst ensuring that data are annotated and made available in a consistent way.
Table 1 Some examples of global collaborations established to manage the public record of different biological data types.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide sequences</td>
<td>International Sequence Database Collaboration [12]</td>
</tr>
<tr>
<td>Protein sequences</td>
<td>UniProt Consortium [13]</td>
</tr>
<tr>
<td>Macromolecular structures</td>
<td>Worldwide Protein Data Bank [14]</td>
</tr>
<tr>
<td>Molecular interactions</td>
<td>The International Molecular Exchange Consortium [15]</td>
</tr>
<tr>
<td>Protein identifications</td>
<td>The ProteomeXchange Consortium [16]</td>
</tr>
<tr>
<td>Metabolomics data</td>
<td>Coordination of Standards in Metabolomics [17]</td>
</tr>
<tr>
<td>Genomic and clinical data</td>
<td>Global Alliance for Genomics and Health [18]</td>
</tr>
</tbody>
</table>

Who owns the data?

From the very beginning, the bioinformatics community has championed open data sharing, and turned it into a reality through collaborations such as the ones summarised in Table 1. This open policy has enabled the research community to make good use of data emerging from major international projects such as the Human Genome Project [2] and the Encode project [19]. However, it’s important to realise that open data sharing is not just for these big collaborations. Recent changes in funding policy have begun to reflect the now widely held view that, if public money is being spent on research (in any field), the data from that research should be made publicly available for others to make use of.

What makes a good bioinformatics database?

While recording biological data itself is useful, the way in which it is recorded makes a huge difference to the value of the database to scientists and informaticians alike. In this section we will discuss two different types of public databases and the mechanisms that they use to describe data consistently.

Primary and secondary databases

Primary databases

In bioinformatics, and indeed in other data intensive research fields, databases are often categorised as primary or secondary (Table 2). Primary databases are populated with experimentally derived data such as nucleotide sequence, protein sequence or macromolecular structure. Experimental results are submitted directly into the database by researchers, and the data are essentially archival in nature. Once given a database accession [11] number, the data in primary databases are never changed: they form part of the scientific record.

Secondary databases

By contrast, secondary databases comprise data derived from the results of analysing primary data. They are often referred to as curated databases but this is a bit of a misnomer because primary databases are also curated to ensure that the data in them is consistent and accurate.

Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary), controlled vocabularies (see later section [20]) and the scientific literature. They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science.

Secondary databases have become the molecular biologist’s reference library over the past decade or so,
providing a wealth of (often daunting) information on just about any gene or gene product that has been investigated by the research community. The potential for mining this information to make new discoveries is vast. It’s our job in this course to reduce your activation energy to make more of these resources for your research.

Table 2 Essential aspects of primary and secondary databases.

<table>
<thead>
<tr>
<th></th>
<th>Primary database</th>
<th>Secondary database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonyms</td>
<td>Archival database</td>
<td>Curated database; knowledgebase</td>
</tr>
<tr>
<td>Source of data</td>
<td>Direct submission of experimentally-derived data from researchers</td>
<td>Results of analysis, literature</td>
</tr>
<tr>
<td>Examples</td>
<td>• ENA [21], GenBank [22] and DDBJ [23] (nucleotide sequence)</td>
<td>• InterPro [27] (protein domains)</td>
</tr>
<tr>
<td></td>
<td>• ArrayExpress Archive [24] and GEO [25] (functional genomics data)</td>
<td>• UniProt Knowledgebase [28] (sequence and functional information on proteins)</td>
</tr>
<tr>
<td></td>
<td>• Protein Data Bank [26] (PDB; coordinates of three-dimensional macromolecular structures)</td>
<td>• Ensembl [29] (variation, function, regulation and more layered onto whole genome sequences)</td>
</tr>
</tbody>
</table>

A note of caution

Many data resources have both primary and secondary characteristics. For example, UniProt [30] accepts primary sequences derived from peptide sequencing experiments. However, UniProt also infers peptide sequences from genomic information, and it provides a wealth of additional information, some derived from automated annotation [31] (TrEMBL), and even more from careful manual analysis (SwissProt).

Some databases have different ‘branches’ for primary and secondary data. A good example of this is the ArrayExpress suite of data resources: the ArrayExpress [24] contains experimentally-derived functional genomics data whereas the Expression Atlas [32] uses a subset of high-quality data from the ArrayExpress to derive knowledge about gene expression [33] patterns under different conditions.

Describing data consistently

The importance of metadata

To be useful, data need to be set in context. One way of doing this is to associate them with metadata - essentially data about the data. If you’re involved in sequencing samples from the environment, perhaps to understand biodiversity in different conditions, or to investigate associations between crop yield and differences in soil flora, it would be useful to know when and where your samples were collected for instance. Standardised descriptors of collection time and geographical location can then be associated with any sequence derived from each sample. Metadata is so important that there are databases dedicated to organising it. For example, the BioSamples database [34] contains metadata on samples used to generate data stored in ENA, PRIDE and ArrayExpress. Storing metadata in this way ensures that a specific sample is referred to consistently in several data resources.

Let’s imagine that the same germplasm sample stored in a seed bank has been used for genomic sequencing, proteomics and RNAseq; these three related experiments can be related to each other by all pointing back to the same record in the BioSamples database. It’s then possible to look at patterns of gene expression [33] and protein production in this sample and compare them to others to learn about how the seed is adapted to a specific environment. Storing the metadata in just one database, rather than as part of the records in three or more separate ones, is also more cost-effective in terms of data storage - an issue that has to be taken extremely seriously in the age of big data.
Describing data and metadata consistently

It is vital that both the data and the metadata are represented in a consistent manner. To take a simple example, let’s imagine that two groups have been working on the effect of antidepressants on gene expression in primary cell cultures of neurones. One of them uses the generic names of the drugs to describe their experiments; the other uses proprietary names. Furthermore, despite isolating their cells from the same tissue using very similar methods, they have different names for their cell lines and use these in their database submission. A computer would think that these two experiments were completely unrelated; and even a human searching for one experiment would be unlikely to find the other. This is why there are agreed standards to describe data - and why databases like those at EMBL-EBI require researchers to annotate their data using these standards when submitting their data to us.

For many types of metadata there are accepted international standards applicable to many fields; for example, if we want to represent location, we can use the standard notation for longitude and latitude. However, as new areas of biology have emerged, and as new technologies have been developed to study them, the research community has had to develop and agree on new standards.

Minimum information standards

Minimum information standards [35] are sets of guidelines and formats for reporting data derived by specific high-throughput methods. Their purpose is to ensure the data generated by these methods can be easily verified, analysed and interpreted by the wider scientific community. Ultimately, they facilitate the transfer of data from journal articles (unstructured data) into databases (structured data) in a form that enables data to be mined across multiple data sets. Minimal information standards are available for a vast variety of experiment types including microarray (MIAME), RNAseq (MINSEQE), metabolomics (MSI) and proteomics (MIAPE). You can explore the data standards available for your field at FAIRsharing.org [36].

MIAME for microarray experiments

The typical example of a minimum information standard is the ‘Minimum Information about a Microarray Experiment’ (MIAME [37]), developed by the Functional Genomics Data Society (FGED [38]; originally the Microarray Data Society) (Figure 5). The community that ultimately became FGED included representatives from labs generating microarray data (academic and industrial), data repositories, microarray manufacturers and journal
publishers. Many journals and funding agencies now require authors reporting on microarray-based transcriptomics [39] experiments to comply with the MIAME standard.

Figure 5 A schematic representation of the six kinds of data captured in MIAME; adapted from the original publication (6 [5]). In this figure Experiment: the set of hybridisation experiments as a whole; Array: each array used and each element (spot, feature) on the array; Sample: samples used, extract preparation and labelling; Hybridisation: procedures and parameters; Data: images, quantification and specifications; Normalisation: types, values and specifications of controls; White boxes: represent external links

Minimum information standards typically have two parts. Firstly, there is a set of reporting requirements - typically presented as a table or a checklist. Secondly, there is a data format. Information about an experiment needs to be converted into the appropriate data format for it to be submitted to the relevant database. In the case of MIAME, the data format is provided in spreadsheet format (MAGE-TAB [40]). Some of the communities that maintain minimum information standards also provide tools to help experimental researchers to annotate their data. For example, Annotare [41] helps researchers to construct MIAME-compliant annotation [31] files based on the MAGE-TAB format.

FAIRSharing.org [42] provides a list of minimum information standards, tagged as reporting guidelines or models/formats.

**Controlled vocabularies**

Controlled vocabularies are a vital ingredient for annotating data stored in databases. In contrast to completely free text entry, controlling the vocabulary helps computers and humans to categorise information and helps reduce redundancy and errors. Whenever you select from a list of prescribed terms, for example choosing an airport when booking a flight, you are using a controlled vocabulary. Biological databases use several different types of controlled vocabularies, which will be discussed in the next sections.

**Non-hierarchical list**

The simplest type of controlled vocabularies are non-hierarchical lists of terms, such as a list of countries.
Annotating data with these lists makes it easier to filter or search for related records in a database. For example, if you use Europe PMC [43]'s advanced search to filter searches by language, you are choosing from items in a list determined by a controlled vocabulary (Figure 6).

Figure 6 Europe PMC [43], EMBL-EBI's literature resource, uses a non-hierarchical list to help you find publications in different languages.

**Taxonomy**

A taxonomy is a classification scheme. The oldest form of taxonomy, and the one that is most likely to be familiar to you if you are a life scientist, is the Linnean classification of organisms. Typically, taxonomies are hierarchical: each 'child' term (more specific term) has one 'parent' term (more generic term), and a parent term can have one or many children. This is the case in the Linnean taxonomy (Figure 7): we start off with living organism, move to
domain, then kingdom, then phylum, class, order, family, genus and finally species.

Taxonomies do not have to be hierarchies; they may also be networks, in which a child term can have one or many parents. The main advantage of using a taxonomy over a non-hierarchical list is that you can find everything annotated as a sub-category of the search term.

![Figure 7](image.png) The Linnean classification of organisms - possibly the archetypal example of a taxonomy.

**Thesaurus**

In the field of [information retrieval](https://www.ebi.ac.uk/training/online), a thesaurus is defined as a controlled and structured vocabulary in which concepts are represented by terms. Terms are organised so that relationships between concepts are made explicit.
Preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms (see the relevant international standard [46]). One example of a thesaurus that you might have come across is MeSH [47], the US National Library of Medicine’s thesaurus of Medical Subject Headings (Figure 8).

The NLM provides a series of webinars and tutorials [48] about MeSH; one that may be of particular interest is Searching Drugs or Chemicals in PubMed [49]. The primary advantage of a thesaurus compared with a non-hierarchical list is that, if the synonyms are well defined, the user can find information without using the exact terminology used by the person who created the database record.

Figure 8 A snapshot of the entry for protein tyrosine kinases in MeSH [50]; synonyms are labelled as ‘Entry term’.

Using ontologies to provide controlled vocabularies

What is an ontology?

In informatics and computer science, an ontology is a representation of the shared background knowledge for a community (7 [5]). An ontology describes the categories of objects described in a body of data, the relationships between those objects, and the relationships between those categories. In doing so, an ontology describes the objects themselves and sometimes defines what you need to know to recognise one of those objects. The labels used to describe the objects can be used to deliver a controlled vocabulary [51], but an ontology is much more than a controlled vocabulary.

The Gene Ontology

The archetypical example of an ontology in the molecular life sciences is the Gene Ontology (GO), created and maintained by the Gene Ontology Consortium [52]. GO describes the function and cellular localisation of gene products across all species (Figure 9), and you can find out more about it in our GOA and QuickGo: Quick tour [53].
GO is used to describe genes and their products in major public databases, including UniProt [30], Ensembl [54] and many model organism databases such as FlyBase [55], SGD [56] and MGI [57]. It provides a powerful means of analysing data sets, for example, it can be used to determine if genes identified as over-expressed in a functional genomics experiment have related functions or similar locations (GO enrichment analysis).

**Figure 9** Snapshot of the protein tyrosine kinase entry in GO, showing the richness of an ontology and highlighting the terms that are used to populate controlled vocabularies.

Since the development of GO in the late 1990s, many more ontologies have been developed using the same principles. These principles include open access, collaborative development and interoperability. These ontologies are collected in the OBO Foundry [58] and are also listed at FAIRsharing.org [59].

If you have a data set and want to annotate it using an ontology, it can be quite a challenge to identify the most appropriate one. The Ontology Lookup Service [60] is a good place to start, requiring no prior knowledge of what ontologies exist.
Where do I submit my data?

No matter how small-scale your research, you can contribute to the public record by submitting your data to appropriate public databases. This is typically done through dedicated submission tools. If you want to submit your data to EMBL-EBI but don’t know where to start, there is a wizard to help you decide which database to submit your data to. The vast majority of public databases also have help pages and helpdesk staff who are there to assist you.

The central EMBL-EBI submission portal [61] can be found at or by clicking on the ‘Deposit data’ button from the EMBL-EBI homepage [7]. Figure 10 shows you the submissions portal and wizard.

If you are submitting data then are a few simple steps that you can take to make the process smoother:

• Start early - begin collecting data and metadata [63] at the beginning of your experiment
• Identify the correct database
• Speak to the curators who work with that database – check what you need to submit!
• Learn about the metadata requirements and data standards used in your field. You can look these up on FAIRsharing.org [64].
• Use an ontology [65] to annotate the data, for example the Experimental Factor onotology [66].
• Remember that curators are there to help!
Tips on managing and sharing data

It's worth asking yourself whether you, your colleagues and collaborators are managing your data in such a way that you can maximise its re-use in the future. Regardless of whether you are working in an academic or a commercial environment, enabling the reuse of your data is becoming a central part of professional practice. There are several excellent papers to get you started (3-5 [5]). A wealth of information about open access [6], open science and open data is available from the FOSTER training portal [67].

When is open sharing not appropriate?

There are two major reasons that you might frequently come across for not openly sharing data:

1. To protect the individual: any data relating to identifiable individuals is sensitive and should be protected by ethical policies. You can learn much more about this in our Biomedical data: Ethical, legal and social implications [68] course. These data can still be shared in databases, but researchers wishing to access them must apply to the relevant ethics committee for permission to gain access to the data for their research.

2. To protect intellectual property or other competitive information. If data are potentially commercially applicable (let’s say, for example, that you have been modelling the docking of zika virus onto transmembrane proteins and think you have evidence for a cell-surface-based target for therapy), the data can still be made publicly available but they should be protected first through appropriate patents [69].

Bioinformatics as an experimental science

Bioinformatics isn't just about storing biological data in databases, it also concerns conducting experiments on that data. Finding a database entry you are interested in is database searching, but as soon as you want to draw a conclusion from your search - inferring homologues of a protein of interest for example - you are conducting an experiment and need to apply the same scientific methodologies in terms of controls etc. that you would to an experiment in the laboratory.

It’s also important that you make an informed choice when deciding which computational method to use, and that you understand the advantages and limitations of the method that you’re using.

Types of bioinformatics experiments

We’ve explored how bioinformatics data are stored and how they are structured and annotated. Now we will learn how you can get to the data and how might you use them to inform the scientific discovery process.

There are a large number of techniques for analysing huge amounts of biological data. In this course we will treat the core databases as a gateway to scientific literature with added, structured, data to help you perform more systematic searches than you would be able to perform using a literature database alone.
As part of that, in the following sections we consider four different types of bioinformatics experiment: searching, comparing, modelling and integrating.

Searching

Perhaps the simplest kind of bioinformatics experiment that you can perform is to search the public databases for information on a specific gene or protein. With EBI Search [70], you can search across a large number of public databases simultaneously, without needing to know in advance which database is most relevant to you (Figure 11).

Figure 11 Data types that can be searched using EBI Search. You can explore the EMBL-EBI’s data using an interactive, zoomable [71] version of this map.

Controls
Performing a simple search is not necessarily an experiment and thus doesn't need a control. However, as soon as you use the results of a search to answer a biological question it becomes an experiment. For example, if you want to find all protein sequences with the keyword 'globin', this is a search, not an experiment. But if you want to find which kinase proteins are i) in a particular reaction pathway and ii) are upregulated in a particular disease state, this is an experiment and you should add some controls to check that you are drawing correct conclusions from the results. For example, you could check if your search terms map to other unrelated pathways. You might also check to see if your search terms are correctly attached to the biological entry - can you find the same terms in related entries, for example?

To learn how to use EBI Search to study potential targets, follow the [guided example on ZAP70](https://www.ebi.ac.uk/training/online) [72].

**Guided example 1: ZAP70 as a potential target for immunosuppressive therapy**

**Background**

ZAP70 is a cytoplasmic tyrosine kinase found mainly in T cells and is important both for T-cell development and for **T-cell-receptor signalling** [73] (Figure 12). Its deficiency leads to severe combined immune disorder (SCID). ZAP70 is considered to be a relevant target for autoimmune disease, in which T-cell activation plays an important role. ZAP70 is also overexpressed in some individuals with chronic lymphocytic leukaemia, and these individuals have an especially poor prognosis. ZAP70 might therefore also be a useful therapeutic target in these patients.
Scenario

Melissa is a discovery biologist at a major international pharmaceutical company working on autoimmune disease. She wants to be able to suppress activation of T cells following T-cell-receptor binding. She wants to know what proteins are involved in T-cell signalling so that she can identify druggable targets.

Step 1 - Searching for ZAP70 using EBI-search

To use EBI Search, go to the EMBL-EBI homepage [75]. Type ZAP70 in the search box. (Figure 13).
Explore EMBL-EBI and our mission

The European Bioinformatics Institute (EMBL-EBI) shares data from life science experiments, performs basic research in computational extensive user training programme, supporting researchers in academia and industry. We are part of EMBL, Europe's flagship laboratory for the life sciences.

[76]

**Figure 13** Searching for ZAP70 on the EMBL-EBI homepage [75].

**Step 2 - Refining your search**
Figure 14 Search results for ZAP70.

The main panel of the results pane (Figure 14) is divided into three sections. Databases are clustered together into groups which can also be used to filter the information on the left (A). The top section, ‘Gene and protein summaries’ (B), has summary pages for frequently studied species. These summaries draw from numerous data resources at EMBL-EBI to bring together commonly sought information such as gene and protein sequences, expression patterns and three-dimensional structural information. The full results are presented below (C).

Click on the Gene & protein summaries for human ZAP70 (1).

### Step 3 - Exploring information on the ZAP70 protein

The gene and protein summary
Figure 15 Exploring the ZAP70 protein entry.

The tabs on the left-hand side (A) guide you to summary information about the gene, its expression, the protein, its structure and any supporting literature. Explore these tabs to answer the quiz questions.

Check your learning

Comparing

Comparing two or more things in biological data allows us to examine how closely related they might be, either in terms of function, evolution, or both.

The most frequently used type of comparison in bioinformatics is sequence comparison to work out how closely related a nucleotide or protein sequence is to others in the public databases. This is done by aligning the sequences - rearranging them to find the best match possible - and takes into consideration insertions, deletions, and substitutions that may have occurred since divergence from a theoretical common ancestor. If a match is found we might be able to infer something about the relationship between sequences. We can perform pairwise sequence alignments and multiple sequence alignments; there are numerous different tools for performing such alignments, and the right one to use will vary depending on the context.

Controls

When it comes to comparing a sequence to entries in a sequence database (sometimes called sequence similarity searching) the challenge is in assessing whether a particular alignment is significant, not in the alignment itself. In this case, an alignment is significant when the likelihood of it occurring by chance (i.e. randomly) is small. This is expressed as the expectation score (also known as an e-value) where the smaller the score, the more significant the alignment, and the more likely it is due to the existence of a shared ancestor and thus homology.
Controls to check the validity of a sequence similarity search include comparing random sequences, and assessing the score of unrelated sequences.

To learn more about sequence similarity searches watch our video on sequence similarity tools, below.

**Quiz: Sequence similarity searching quiz**

- Questions: 3
- Attempts allowed: Unlimited
- Available: Always
- Pass rate: 100%
- Backwards navigation: Allowed

**Guided example 2: does sheep ZAP70 have an active tyrosine kinase?**

**Scenario**

Melissa is working on a drug discovery programme to create new inhibitors of ZAP70 for use in autoimmune disease. Greg, a medicinal chemist in her group, has synthesised a series of protein tyrosine kinase inhibitors. She needs to devise a cellular assay [80] to see which ones inhibit the ZAP70 tyrosine kinase. She has ready access to primary T cells from the following animals:

- Mouse
- Rat
- Sheep

To decide which of her models has a ZAP70 protein most closely related to the human one, she has decided to do a quick multiple sequence alignment. Using EBI search, she located the canonical sequences (isoform 1) for the human, mouse, rat and sheep ZAP70 proteins. For each one, she downloaded the protein sequence in FASTA [81] format. She then performed a multiple sequence alignment of these proteins using the Clustal Omega [82] tool.

You can repeat her experiment if you like.

On the basis of this alignment, Melissa has decided to use sheep T cells for her assay but she wants to double check that sheep ZAP70 has an active tyrosine kinase domain before wasting time on developing an assay. For this, she’s going to use a tool called InterProScan. This tool compares your chosen sequence with all the sequences in InterPro, a data resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. If you’d like to learn a bit more about InterPro before continuing, you can read the InterPro Quick Tour [83] or follow the InterPro tutorial [84]. Go straight to the section on sequence searching [85] if you’re short of time.

**Step 1 - Examining the ZAP70 sequence with InterProScan**

Go to the InterProScan [86] webpage and paste your sequence into the search box, or copy and paste it from here.
Figure 16: Pasting in a sequence into the InterProScan search box.

**Step 2 - Interpreting the search results**

When your search is complete, your search results will look like this (Figure 17).
Figure 17: Search results for ZAP70 on InterProScan.

At the top of the results, you will see the reassuring news that sheep ZAP70 is indeed a member of the tyrosine protein kinase family, and that it has two SH2 domains and a protein tyrosine kinase domain. One thing to check is whether our sequence matches the patterns for the active site and ATP binding site, since any mutations in this area may make the protein enzymatically inactive. In this case, however, we can see that sheep ZAP70 contains the protein profile for both the ATP-binding site and the active site (towards the bottom of the figure). Melissa should be safe to use sheep cells for her assay!
Structural modelling can be used to generate hypotheses about the structure (and therefore to imply things about the biochemical function) of macromolecules.

Try our guided example on structural modelling of IRAK2 [87].

Modelling of processes is an important aspect of systems biology. A recent review by Bartocci and Lió [88] provides a good summary of the kinds of processes that are amenable to systems modelling, and some examples of tools that enable you to do this. You can download mathematical models in a number of standard formats from the BioModels Database [89]. Check out our Biomodels: Quick tour [90] to learn more about this.

Systems modelling forms part of a cycle in which experimental biology plays an equally important role: a typical systems modelling cycle involves building a model that represents what you know about the biology then testing to see whether it behaves in the same way as the biological system itself. If not, you tweak the model to come closer to the biological system and repeat the cycle until your model faithfully represents reality.

Guided example 3 - building a structural model based on sequence similarity

Homology modelling [91] is the process of estimating a theoretical structure of a protein by building a structural template from homologous [92] sequences and threading your sequence to this template and optimising it.

Scenario

Jan is interested in interleukin 1-receptor-associated kinases (IRAKs) as possible targets for cancer therapy. Several mutations in IRAK2 are associated with cancers, so Jan would like to investigate IRAK2 as a possible new target. There are quite a few small-molecule inhibitors available for related proteins and Jan wants to know whether they’d be likely to bind to IRAK2. However, there is no crystal structure available for the kinase domain of IRAK2, so Jan wants to model the structure to see whether it’s likely to resemble the 3D structures of other IRAKs.

Step 1: downloading the sequence of human IRAK2 and first approach modelling

Go to EBI Search [70] and type IRAK2 in the search box. Find human IRAK2 and download the FASTA [81] format sequence for it, or copy the sequence from here:

```
>sp|O43187|IRAK2_HUMAN Interleukin-1 receptor-associated kinase-like 2 OS=Homo sapiens GN=IRAK2 PE=1 SV=2
MACYIYQLPSWVLDDLČRNMDALSEWDWMEFASYVITDLTQLRKKSMERVQGVQVSITREL
LWWWGMRQATVQLVDDLČRLELYRAAQIILNWKPAPEIRCPIPAFPDVSKPKLPLASV
RKAĐEĐEQEEGOFVRMATFPGSSPARAHQAFLQPFEEDEAPHSLRSDLPTSSDSKDFST
SIPKQEKSSLGDSLFWSEAVQATDDFQONRKISQTFADVYRHQHKGKPFWKFKLKR
ETACSSPGSIERFFOALQICLRCCHPNVLPVGLFCAARQFHSFIYPYMANGSLQDRLQG
QGGSDELPLPWPQRVSCGLACVPGLHGEIIHNSVMSSNLQNDQNLPLSIPKQEKSSLGDSLFWSEAVQATDDFQONRKISQTFADVYRH
```

Now go to the SWISS-MODEL [93] service at the Swiss Institute of Bioinformatics. Click on the ‘Start modelling’ button and paste your sequence into the box. Click on ‘Build model’. SwissModel will automatically choose
templates based on the most closely aligned protein sequence that has a three-dimensional structure available for it, and model a structure based on that (Figure 18).

![Model Results](image)

**Figure 18** Model results.

**Controls**

There are several approaches to assessing models. One is to check that values such as bond angles, distances etc. in the model match values that have been observed in experimentally derived structures. In SWISS-MODEL the QMEAN z-score represents an estimate of how comparable the model is to experimentally derived structures of similar size. QMEAN z-scores around zero indicate good agreement between the model structure and experimental structures of similar size. Models of low quality typically have scores of -4.0 or lower. The “thumbs-up” and “thumbs-down” symbols next to the score are used to indicate whether or not the model is of good quality ([5]). Another approach is to factor in observations of the quality of the alignment and template search method - this is represented in the GMQE (Global Model Quality Estimation) score. The GMQE score reflects the expected accuracy of that alignment and is expressed as a number between 0 and 1 where higher numbers indicate higher reliability ([5]). For more information see the SWISS-MODEL documentation pages ([94]).

**Step 2: building a model based on your chosen template**
Figure 19 Choosing your own structural template.

You can also choose your own structural template. Click on ‘Search for templates’. SWISS-MODEL will look for structures in the PDB [95] that have a closely related sequence (Figure 19). Be patient; it can take some time to generate your list of templates. Once you’ve done this, you can align as many of these templates as you like to your target sequence. Here we’ve chosen structures for IRAK4 (2oib.2.A), brassinosteroid insensitive 1-associated receptor kinase 1 (3tl8.2.A) and JAK2 (4bbe.1.A). These three structures align quite closely to each other. Based on this evidence, it might be worth doing some assays to find out whether any known inhibitors of these enzymes also inhibit IRAK2.

Integrating

Data integration is a longstanding challenge for bioinformatics (Figure 20) but can be a tremendously powerful means of gathering evidence for or against a hypothesis. For example, integrating data from transcriptomics, proteomics and metabolomics experiments can help to build evidence that a particular pathway is involved in a disease, or in resistance to a drug.
As with systems modelling, data integration helps you to generate hypotheses, but must be combined with experimental approaches to test your hypothesis.

### Basic principles of integration

If you plan to integrate data, it needs to have as many similarities as possible. The same entity or concept described in different ways is not amenable to integration. Data integration therefore invariably requires some preparation.

### Tools to help integrate data

There are tools available to help - for example, PICR [96] allows you to convert a set of identifiers from one format to another. There are also mappings of different controlled vocabularies, but care needs to be taken that you don’t lose data. For example, a term in one ontology might be mapped to a term that is less granular, so you might lose specificity. At EMBL-EBI we use application ontologies, the archetypal example of which is the Experimental Factor Ontology [97], to solve this problem.

If you and your collaborators submit data to public repositories, the data will be put into a standard format and the data integration will essentially be done for you. If you work in a commercial environment, you may have your own in-house databases, or you may use private instances of the public databases. EMBL-EBI’s Embassy Cloud [98] provides EMBL-EBI’s collaborators with direct access to their datasets hosted at EMBL-EBI, and to the institute’s
powerful computing resources. This shared, high-performance workspace allows project partners in many locations to analyse their data alongside public offerings, using their own approaches. Access to the Embassy Cloud is available to collaborators working on projects with EMBL-EBI. The service has been successfully piloted with Europe PMC [99] (partners in Manchester, London and EMBL-EBI) and Tara Oceans (EMBL and global collaborators), and is now more widely available.

Where does the data come from?

It’s important to understand the origin of the data that you are integrating, and to be able to check the evidence for the involvement of each entity in the bigger picture. If, for example, you are integrating different types of omics data to understand the regulation of a pathway and its dysregulation in disease, you need to have a good understanding of the pathway in question and the disease that you’re studying, whilst remaining open and unbiased about what the data might be telling you.

Data integration requires that all the data are annotated in a consistent way. You also need to be absolutely sure that you’re comparing like with like. BioSamples [34] database will allow you to find all the experiments performed on the same sample. To learn more about BioSamples take a look at our BioSamples: Quick tour [100].

If you’re performing your own experiments, bear in mind that others may want to integrate your data with data from other sources in the future. Providing adequate metadata [63], and formatting your data in a re-usable way, should become second nature to you.

Toni Kazic’s guide for data provenance (8 [5]) is a good place to start. If you’re using other people’s data, check it as though it were your own.

Pre-canned data integration

The good news is that there are now an increasing number of resources that have done a lot of the hard work for you. We have already used one service - EBI Search [70] - that does a lot of the mapping of related entities for you. Another service that integrates a huge amount of public data relevant to discovery is Open Targets [101]. Open targets is a service that is designed to enable exploration and visualisation of drug targets associated with disease. You can learn more about Open Targets in our webinar Open Targets: Mining gene and disease associations for improved drug target identification [102].

To learn more about data integration you can join us for a face-to-face course or explore materials our past courses on Introduction to multiomics data integration [103].

Summary

- Bioinformatics is central to molecular life-science research.

- Public databases such as those available on the EMBL-EBI website provide open access [6] to a wealth of biological information, allowing you to perform in silico experiments without needing to write any code.

- The essentials of a robust bioinformatics resource are a well-thought-out database structure and adherence to community standards: it helps to have a basic awareness of these two core principles of a public data resource as this will enable you to navigate your way round the world of databases, and communicate with bioinformatics experts, more effectively.
Bioinformatics is an experimental science: it’s important to give careful consideration to the method that you use, and to build in controls exactly as you would for a wet-lab experiment.

Bioinformatics experiments can broadly be categorised as
- Search
- Compare
- Model
- Integrate

Your feedback

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

Get help and support

If you’re not sure where to find the type of data or the bioinformatics tool that you need, a good place to start is EBI Search [70], which has its own help pages [104].

If this doesn’t provide you with what you’re looking for, please do contact us at EMBL-EBI support [105]: we’re here to help.

Recommended courses

We also run an extensive programme of face-to-face training courses [106], and hold a regular webinar series [107], in addition to providing openly accessible training on demand through Train online [108].

Finally, if you can’t find what you’re looking for at EMBL-EBI, on-course [109]® provides information on a wealth of bioinformatics and other courses, both face-to-face and online.

We hope you have enjoyed this course and warmly welcome your feedback.

References


Contributors

Cath Brooksbank [1]
EMBL-EBI
Head of the EMBL-EBI Training Programme

Before joining EMBL-EBI in 2002, Cath spent a decade as an editor of scientific review journals, cutting her teeth with the Elsevier Trends Journals before launching Nature Reviews Cancer. A brief dalliance with medicine at the University of Oxford led her to seek solace in molecules in Cambridge, where she completed her PhD in biochemistry under the guidance of Robin Irvine. Cath joined EMBL-EBI to develop the outreach programme, and extended her responsibilities to include user training in 2006. Her team now coordinates a wide-ranging portfolio of training and scientific outreach activities reaching tens of thousands of individuals each year. She contributes to a number of pan-European projects including RItrain, CORBEL, BioExcel and ENLIGHT_TEN, and is co-chair of the curriculum and competencies taskforce of the International Society of Computational Biology.
Andrew Cowley [2]

EMBL-EBI
Scientific Training Officer (e-learning)

Andrew Cowley is currently working as a scientific training officer for e-learning in EMBL-EBI’s Training team. Previously he was the Bioinformatics training & support project lead for EMBL-EBI’s Web Production team which involved working on tools, training, and helping users with all kinds of questions via the EMBL-EBI help desk. He has particular expertise in sequence alignment tools and programmatic access to EMBL-EBI resources.

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 in protein structure. He worked as a Bioinformatics specialist for BBSRC Bioscience IT Services from 2005 before starting at EMBL-EBI in 2010.

Source URL: https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-0

Links
[1] https://www.ebi.ac.uk/training/online/trainers/cath
[2] https://www.ebi.ac.uk/training/online/trainers/apc
[3] https://commons.wikimedia.org/wiki/File%3AThe_Scream.jpg
[4] https://www.genomicsengland.co.uk/the-100000-genomes-project/#
[5] https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-2018/references
[6] https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-2018/references
[10] http://www.ebi.ac.uk/ena
[11] https://www.ebi.ac.uk/training/online/glossary/accession
[19] http://www.nature.com/encode/#/threads
[20] https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-2018/controlled-vocabularies
[21] https://www.ebi.ac.uk/ena
[24] https://www.ebi.ac.uk/arrayexpress/
[26] http://www.ebi.ac.uk/pdbe/node/1
[27] https://www.ebi.ac.uk/interpro/
[31] https://www.ebi.ac.uk/training/online/glossary/annotation
[32] https://www.ebi.ac.uk/gxa/home
[33] https://www.ebi.ac.uk/training/online/glossary/gene-expression
[34] https://www.ebi.ac.uk/biosamples/
[36] https://fairsharing.org/collection/MIBBI
[38] http://fged.org/
[39] https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified-transcriptomics
[96] http://www.ebi.ac.uk/Tools/picr/
[97] http://www.ebi.ac.uk/efo/
[99] https://www.ebi.ac.uk/training/online/glossary/europe-pmc
[100] https://www.ebi.ac.uk/training/online/course/biosamples-quick-tour
[101] https://www.targetvalidation.org/
[102] https://www.ebi.ac.uk/training/online/course/open-targets-mining-gene-and-disease-associations-improved-drug-target-identification
[103] https://www.ebi.ac.uk/training/events/2018/introduction-multiomics-data-integration-0
[104] https://www.ebi.ac.uk/ebisearch/documentation.ebi
[105] http://www.ebi.ac.uk/support/
[106] http://www.ebi.ac.uk/training/handson
[107] http://www.ebi.ac.uk/training/webinars
[108] https://www.ebi.ac.uk/training/online/
[109] https://www.on-course.eu/
[116] http://ontogenesis.knowledgeblog.org/66