On-line resources for biologists

Jean-Christophe Nebel, Kingston University.

What we will learn

Human genes are around 27,000 base pairs long, which is a headache if you have to count ATCG etc. That is why bioinformatics is important to digitalize the process and make practical the process of looking for mutations, homologies etc. In this chapter we aim to support the biologist wishing to take advantage of the bioinformatics revolution by introducing them to several essential free on-line resources. After introducing this revolution, the types and natures of biological data it involves are reviewed. Then, the main on-line resources are presented identifying the most useful databases and analysis tools. Next, due to their critical role in the process of rational drug design, resources especially dedicated to prediction and analysis of 3D protein structures are discussed. Finally, some domains of application are briefly presented.

The bioinformatics revolution

The publication of the first 'Atlas of Protein Sequence and Structure' (Dayhoff et al., 1965) which comprised the sequences of 65 proteins, arguably funded the field of bioinformatics, even if the term itself was only coined in 1970 by Paulien Hogeweg and Ben Hesper. This atlas gave researchers the opportunity to compare sequences and establish evolutionary relationships between proteins and predict their function by simply using a computer. Instead of only performing *in vivo* and *in vitro* experiments, biologists were given the possibility to conduct *in silico* studies. By identifying corresponding characters between two sequences, the evolutionary process which has taken place could be reconstructed. Since the average size of a gene comprises several tens of thousands of characters and they evolve through mutations where characters are replaced, added and removed, the alignment of two sequences is not a trivial matter. The 'Needleman–Wunsch' algorithm (Needleman & Wunsch, 1970) has provided an effective automatic method to produce an exact solution to the global alignment of two sequences and it is still at the core of the latest search engines, which allow finding the best alignment between a given sequence and relevant entries in a large database containing millions of sequences.

With international efforts such as HUGO (the Human Genome Project) which sequenced the 3 billion DNA characters of the human genome (2001), thousands of complete genomes are now available and this number is increasing at an exponential pace. Their analysis has required not only the applications of conventional data mining and pattern recognition approaches, but also the development of completely novel techniques to handle the specificity and sheer size of genomics data. With the expectation that deciphering the human genome will result in dramatic improvement of health, the international community has required from bioinformatics to produce fast, efficient and robust computational techniques tailored to genome analysis. As a consequence, nowadays bioinformatics organisations deliver mature and powerful tools which serve millions of scientists. Not only can biologists access free databases containing most genomics and proteomics data produced by the international community, but also services allowing their analysis are also available largely free of charge. The genomics revolution is becoming a reality: evolutionary relationships can be inferred, common ancestors can be reconstructed, functions of genes and proteins can be predicted, disease mechanisms can be deciphered and potential drugs can be designed. Eventually, personalised medicine will be delivered.

The flow and transformation of genetic information

Biology data have a variety of formats which not only may depend on standards used for their representation, but also on biological entities and their state. Guided by the central dogma of molecular biology which describes the flow of genetics information from DNA to protein, important data types are illustrated in Figure 1.

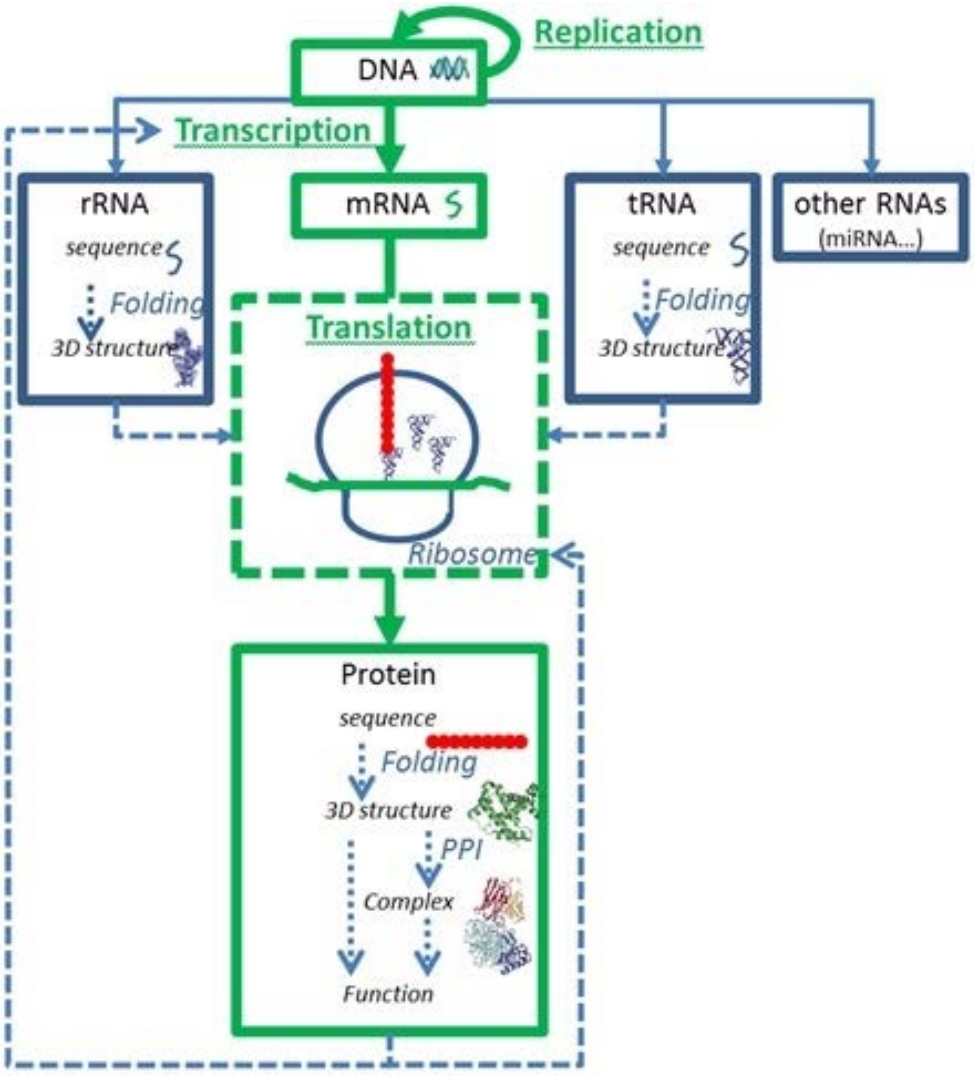


Figure 1: The flow and transformation of genetic information. Each box represents a different type of molecule and shows their possible forms, i.e. linear sequence, 3D structure or complexed. The central dogma of molecular biology is represented in green. Other transcription products and interactions/transformations are shown in blue.

All living organisms store their genetic material using nucleotides which form the DNA double helix. Note that the genetic material of viruses is stored as either DNA or RNA; viruses are generally not considered as living organisms since they do not have the endogenous molecular machinery allowing replication, for that they need a host cell. DNA has the ability to replicate itself as part of cell division and the growth of organisms. The central dogma of molecular biology (represented in green in

Figure 1) shows how that information is transformed in an organism. Subsets of DNA, called genes, are transcribed into messenger RNA (mRNA), which, in turn, is translated into sequences of amino acids called proteins. They are molecules playing a central role in biological processes. Whereas this describes the main flow of information, other less common processes exist especially in viruses: they include reverse transcription, where DNA is generated from RNA, and RNA replication. While the central dogma models the flow of information, it does not consider the different forms that genetic information can adopt and is not concerned by non-protein coding DNA. Figure 1 includes that information: DNA transcription may lead to the production of ribosomal RNA (rRNA) – they are component of the ribosome, a complex molecular machine involved in protein synthesis -, transfer RNA (tRNA) - they translate codons (triplets of nucleotides) into amino acids - and other RNA forms such as microRNA. In order to be active both rRNA and tRNA need to adopt a 3D shape. Similarly, proteins need to fold into a 3D structure to become functional. However, often they also require interacting with other proteins to form protein complexes. Finally, it is worth mentioning that the function of some proteins is to regulate the transcription process; they are called transcription factors.

As a consequence of the variety of genetic materials and the different forms they can adopt, molecular data have very different formats. DNA (including genes) and RNA sequences are represented by strings of characters representing four different types of nucleotides (A, C, G and T for DNA, and the corresponding A, C, G and U for RNA). Protein sequences are encoded by linear chains of amino acids: the 20 types of amino acids may be stored either as strings of characters (a subset of the Roman alphabet: A, C, D, E, F, G, H, I, K, L, M, N, P Q, R, S, T, V, W and Y) or sequences of 3-letter codes which are abbreviations of their full names, e.g. Alanine and asparagine can be represented, respectively, by either by ‘A’, resp. ‘N’, or ‘ALA’, resp. ‘ASN’. Structural data include for each atom of each nucleotide or amino acid its 3D coordinates (x, y, and z) in angstroms. Since protein conformations typically rely on sub-structures, called secondary structures, protein sequences can also be annotated with the secondary structure to which each amino acid belongs. The most common ones are alpha helices, beta sheets and coils, which are usually represented by the characters ‘H’, ‘E’ (for extended) and ‘C’ (or ‘-‘). Figure 2 shows various data encodings of calmodulin, a calcium-binding protein involved in processes including inflammation, smooth muscle contraction, memory and the immune response. In addition, more dynamic data, often associated to a specific type of tissues, can be acquired: for example, gene/protein expression or protein interactions can be measured.

```
>ENA|M27319|M27319.1 Human calmodulin mRNA, complete cds.
CAGCATCGGAGGTACCCCCGCCGTCGCAGCCCCCGCGCTGGTGCAGCCACCCTCGCTCCCTCTGCTCTTCCTCCCTTCACTCGCACCATGGCTG
ATCAGCTGACCGAAGAACAGATTGCTGAATTCAAGGAAGCCTTCTCCCTATTTGATAAAGATGGCGATGGCACCATCACAACAAAGGAACTTGG
AACTGTCATGAGGTCACTGGGTCAGAACCCAACAGAAGCTGAATTGCAGGATATGATCAATGAAGTGGATGCTGATGGTAATGGCACCATTGAC
TTCCCCGAATTTTTGACTATGATGGCTAGAAAAATGAAAGATACAGATAGTGAAGAAGAAATCCGTGAGGCATTCCGAGTCTTTGACAAGGATG
GCAATGGTTATATCAGTGCAGCAGAACTACGTCACGTCATGACAAACTTAGGAGAAAAACTAACAGATGAAGAAGTAGATGAAATGATCAGAGA
AGCAGATATTGATGGAGACGGACAAGTCAACTATGAAGAATTCGTACAGATGATGACTGCAAAATGAAGACCTACTTTCAACTCCTTTTTCCCC
CCTCTAGAAGAATCAAATTGAATCTTTTACTTACCTCTTGCAAAAAAAAGAAAAAAGAAAAAAGTTCATTTATTCATTCTGTTTCTATATAGCA
AAACTGAATGTCAAAAGTACCTTCTGTCCACACACACAAAATCTGCATGTATTGGTTGGTGGTCCTGTCCCCTAAAGATCAAGCTACACATCAG
TTTTACAATATAAATACTTGTACTACCTTAATGATAAGGACTCCTTA

>sp|P0DP23|CALM1_HUMAN Calmodulin-1 OS=Homo sapiens OX=9606 GN=CALM1 PE=1 SV=1
MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFD
KDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK

>1CLL:A|PDBID|CHAIN|SEQUENCE
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDK
DGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK

ALA ASP GLN LEU THR GLU GLU GLN ILE ALA GLU PHE LYS GLU ALA PHE SER LEU PHE ASP LYS ASP GLY
ASP GLY THR ILE THR THR LYS GLU LEU GLY THR VAL MET ARG SER LEU GLY GLN ASN PRO THR GLU ALA
GLU LEU GLN ASP MET ILE ASN GLU VAL ASP ALA ASP GLY ASN GLY THR ILE ASP PHE PRO GLU PHE LEU
THR MET MET ALA ARG LYS MET LYS ASP THR ASP SER GLU GLU GLU ILE ARG GLU ALA PHE ARG VAL PHE
ASP LYS ASP GLY ASN GLY TYR ILE SER ALA ALA GLU LEU ARG HIS VAL MET THR ASN LEU GLY GLU LYS
LEU THR ASP GLU GLU VAL ASP GLU MET ILE ARG GLU ALA ASP ILE ASP GLY ASP GLY GLN VAL ASN TYR
GLU GLU PHE VAL GLN MET MET THR ALA LYS

ATOM      1  N   LEU A   4      -6.873  21.082  25.312  1.00 49.53           N
ATOM      2  CA  LEU A   4      -6.696  22.003  26.447  1.00 48.82           C
ATOM      3  C   LEU A   4      -6.318  23.391  25.929  1.00 46.50           C
ATOM      4  O   LEU A   4      -5.313  23.981  26.352  1.00 45.72           O
ATOM      5  N   THR A   5      -7.147  23.871  25.013  1.00 46.77           N
ATOM      6  CA  THR A   5      -6.891  25.193  24.428  1.00 46.84           C
ATOM      7  C   THR A   5      -6.801  26.228  25.543  1.00 45.36           C
ATOM      8  O   THR A   5      -5.829  26.999  25.561  1.00 47.41           O
ATOM      9  CB  THR A   5      -7.923  25.626  23.323  1.00 46.33           C
ATOM     10  OG1 THR A   5      -9.238  25.386  23.908  1.00 48.28           O
ATOM     11  CG2 THR A   5      -7.704  24.943  21.974  1.00 44.31           C
ATOM     12  N   GLU A   6      -7.781  26.217  26.419  1.00 46.42           N
ATOM     13  CA  GLU A   6      -7.858  27.154  27.557  1.00 45.65           C
ATOM     14  C   GLU A   6      -6.502  27.343  28.224  1.00 44.22           C
ATOM     15  O   GLU A   6      -6.040  28.483  28.423  1.00 42.97           O
ATOM     16  CB  GLU A   6      -8.881  26.726  28.592  1.00 50.13           C
ATOM     17  CG  GLU A   6      -9.358  27.699  29.657  1.00 55.79           C
ATOM     18  CD  GLU A   6     -10.322  27.149  30.668  1.00 58.96           C
ATOM     19  OE1 GLU A   6     -10.775  26.034  30.334  1.00 62.89           O
ATOM     20  OE2 GLU A   6     -10.648  27.701  31.706  1.00 60.58           O
```

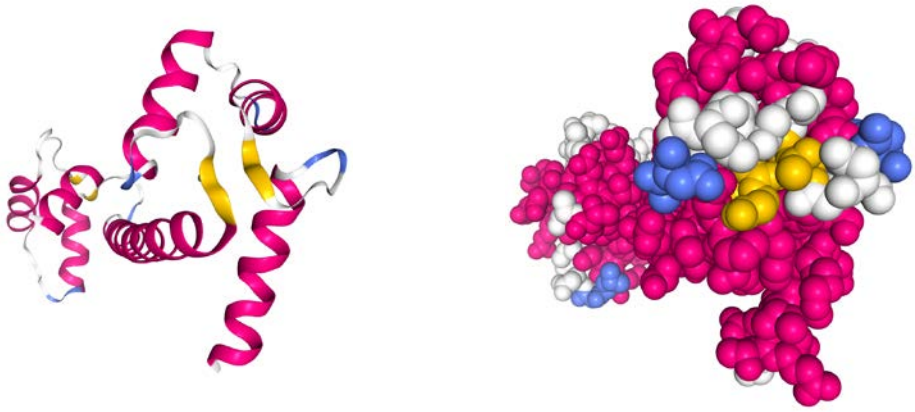


Figure 2: Some data encodings associated to homo sapiens Calmodulin-1 (CaM-1): a) coding region of the CaM-1 gene (fasta format), b) reference sequence of the CaM-1 protein, c) sequence of a 3D model (1CLL) of the CaM-1 protein, d) sequence of 1CLL using 3-letter code format, e) Atom 3D coordinates (in bold) of amino acids 4-6 of 1CLL (PDB format) and f) cartoon (left) and spacefill (right) representations of 1CLL, where alpha helices and beta sheets are coloured in pink and yellow, respectively.

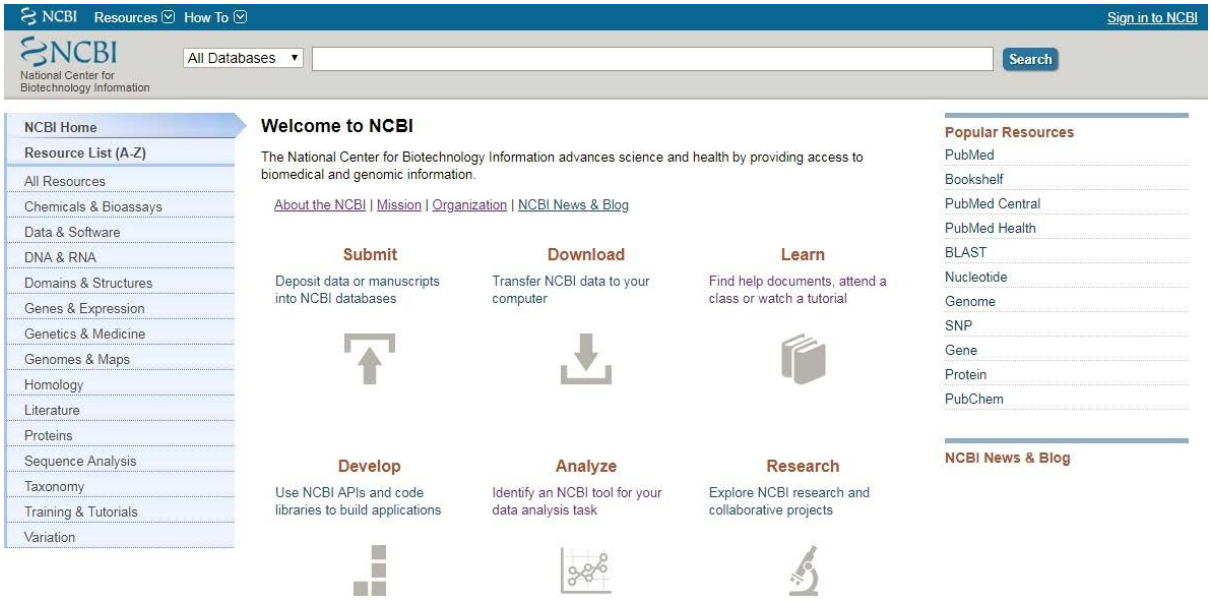
The main on-line resources

If one looks for “bioinformatics resources” using one’s favourite search engine, it returns tens of thousands of results giving access to websites which provide links to generally free databases and tools. Among them, the bioinformatics web portals offered by the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EMBL-EBI) may be considered as the references for the scientific community:

- NCBI was founded in 1988 as part of the US National Institutes of Health (NIH). As a national resource for molecular biology information, NCBI has been charged with “*creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally*”. Its impact can be illustrated by its daily provision of over 35 Terabytes of data to scientists from all over the world.

- EBI is part of the European Molecular Biology Laboratory (EMBL) an intergovernmental organisation created in 1974 “to promote molecular biology across Europe, and to create a centre of excellence for Europe's leading young molecular biologists”. Currently funded by 24 European governments and, science institutes and company from over 60 countries, one of its missions (delivered by EBI) is to support the scientists and engineers world-wide “through open data, connecting the global scientific community by providing data services and fostering collaborative Research”. Addressing the exponential growths of biological data, it stores over 120 Petabytes of data and every weekday its websites fulfil over 27 million requests. It serves over 3 million individual scientists every month.

Both resources not only offer user-friendly websites allowing access to freely available data and services, but also a variety of educational resources including manuals, webinars, training material and online courses. In addition, well documented access is offered for high throughput usage. Note that this will not be discussed further since it is out of the scope of this chapter.



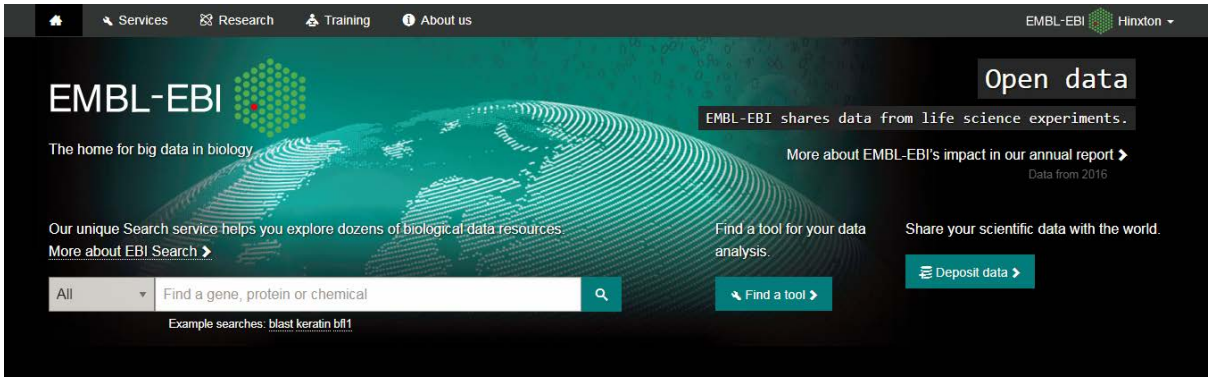


Figure 3: Welcome pages of the bioinformatics web portals offered by a) NCBI (top) and b) EBI (bottom).

Using the web interface, data can be accessed using a simple search box (Figure 3), where users type the name of their molecule of interest (gene or protein). The associated search engine will return links to all relevant existing resources, including, when appropriate and available, gene and protein sequences, expression data, 3D structures, interaction data, disease associations and related literature. While on the same page (Figure 3a, menu on the left) NCBI provides lists of its resources organised by data and tool types, an extra click on 'Find a tool' is necessary on the EBI page to reach their 'Services' page (Figure 4). There, the most popular tools and databases are on display. Those lists can be expended by clicking on 'See all tools' or 'See all data resources' at the bottom of the page. In addition, on the right there is menu to select both tools and databases relevant for a given data type. Note that the search box at the centre of the page allows searching for specific tools and databases.

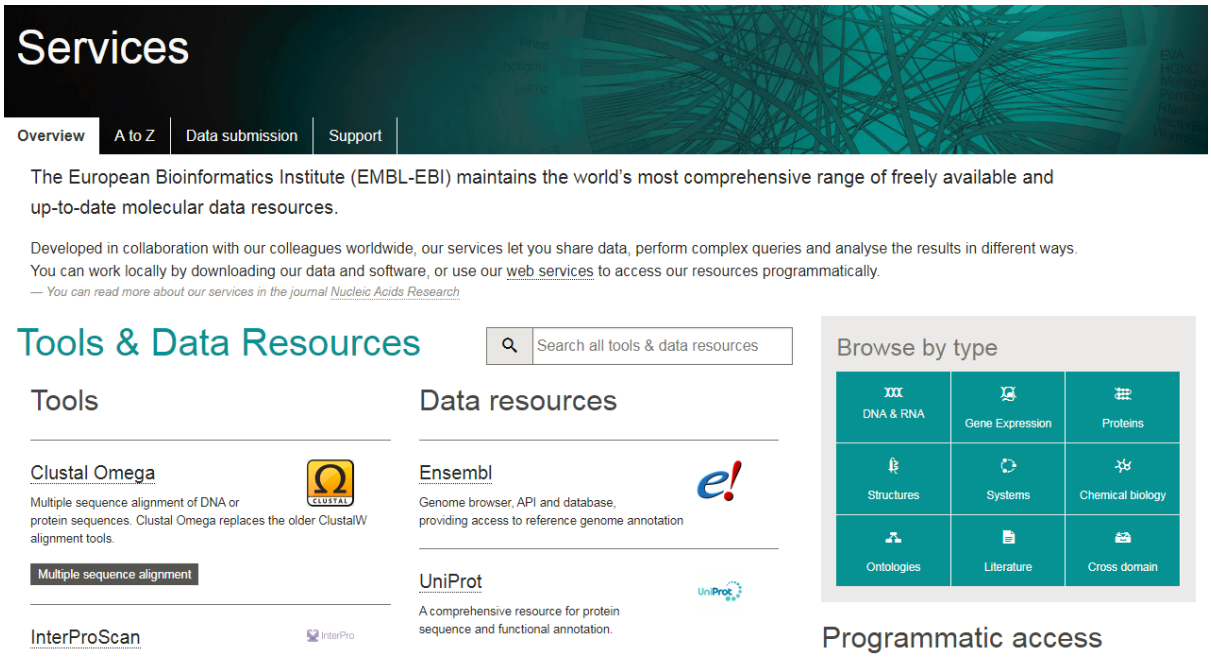


Figure 4: Services page of the EBI web portal.

Both NCBI and EBI offer essentially secondary databases – over 60 each which are generally updated on at least a monthly basis, some daily. Those databases result from the combination, analysis and processing of data stored in databases only reporting experiment results, such as amino acid sequences. A few important examples are briefly described:

- The NCBI's Reference Sequence (RefSeq) database aims to deliver for each individual species, a complete set of non-redundant DNA, RNA, and protein sequences. RefSeq release 88 comprises data from 79,448 organisms: 27,325,628 DNA, 22,461,378 RNA and 110,333,800 protein sequences.
- EBI's Ensembl is a browser for vertebrate genomes incorporating a wealth of relevant information such as genetic variations, splice variants, molecular functions, phenotypes, gene expression and associated pathways, and offering results generated by its integrated comparative genomics tools, including multiple sequence alignments and phylogenetic trees. Ensembl Release 92 (April 2018) supports the analysis of 125 vertebrate genomes.
- The UniProt Knowledgebase (UniProtKB) offered by EBI provides for all know protein sequences not only their sequence, name, taxonomic data and literature references, but also as much annotation information as possible. They may include expression and structural data, associated diseases and phenotypes, and functional annotations that come from both manual curation and computational predictions. In UniProt release 2018_05, UniProtKB contained 557,491 human and 115,678,811 computer generated entries.
- The Online Mendelian Inheritance in Man® (OMIM) database was developed by NCBI to provide an exhaustive catalogue of human genetic phenotypes involving nearly 16,000 genes. In particular, for many human genetic conditions, associated mutations are listed with referenced descriptions of related medical aspects.
- The Pfam database, hosted by the EBI, makes available a collection of multiple sequence alignments and hidden Markov models (HMMs) describing functional regions, also called domains. Using the associated Pfam search tool, the identification of a Pfam domain within a protein sequence of interest may contribute to its functional annotation. Pfam 31.0 (March 2017) contains 16712 domain descriptors.
- Relying on the Protein Data Bank (PDB), a primary database that is "the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies", the NCBI's Molecular Modeling Database (MMDB) offers derived structural information such as secondary structures, structural domains and molecular interactions, as well as evidence of the identification of functional regions. Such combined information assists, for example, further investigation of sequence-structure-function relationships.

All those databases are associated to bioinformatics tools which may have been used to generate some of their content or could be used to investigate further the data they provide. While NCBI makes available over 60 different tools to users, EBI lists almost twice that. It is important to note that, although most tools can be used using a web interface taking advantage of the processing facilities offered by those organisations, it is generally possible to install them (and their associated databases, if relevant) on one's own computer. There are many reasons why one may want to process one's data on one's local machine. First, one must be aware that nothing is ever secure on the internet. Although those organisations are trustworthy and deal with millions of requests every day, the data sent for processing could be intercepted by a third party. Second, databases and tools are regularly updated. Therefore, results of experiments conducted on the internet may vary with time. By fully managing the experimental set up, one can ensure that one's own results are consistent and reproducible. Finally, by running the experiments on one's own computer, one can be in better control of processing times: one may have a particularly powerful machine or one may want to make sure that one's job's processing is not going to be delayed by staying in a slow moving queue on the server hosting the tool.

Most bioinformatics tools on offer are designed to process sequence data. Arguably, the most important task is to retrieve similar sequences from a database. Indeed, it is an essential first step in many bioinformatics pipelines. The most popular tool, Basic Local Alignment Search Tool (BLAST), compares a sequence to a sequence database of choice returning the best hits with their statistical

significance. BLAST and its specialised versions, such as genome, nucleotide or protein BLAST, are available from both NCBI and EBI. Once similar sequences have been identified, pairwise or multiple sequence alignment is often required to highlight informative mutations or conserved sequence regions. Pairwise sequence alignments are usually performed by either the Needleman-Wunsch or the Smith–Waterman algorithms, or some of their variations. Tools based on those algorithms are provided by both NCBI and EBI. Thinkbox 1 portrays a scenario that can be addressed by the tools under discussion.

Thinkbox 1: Origins of the latest seasonal flu virus

The "flu" is a contagious disease caused every winter by influenza virus A infection. Although type A viruses are found in many animals including pigs, ducks, chickens, horses and seals, their main reservoir is thought to be in birds. This is why one speaks of bird flu. Seasonal flu kills between 250,000 and 500,000 people every year. However, some outbreaks of influenza can be even more lethal: the so-called Spanish Flu pandemic (1918-1919) killed tens of millions of people! As a consequence, it is essential to understand the origin of any new flu virus so that its severity can be estimated and appropriate measures can be taken.
Among the 8 genes of influenza A viruses, two code for proteins essential for their spread: Hemagglutinin (H) and Neuraminidase (N). There are 16 different H genes, labelled H1 - H16 (only H1, H2 and H3 are found in human viruses) and 9 different N genes, labelled N1 - N9 (only N1 and N2 are found in human viruses). The nature of those two genes is used to qualify a type of flu. For example, standard flu jabs aims at protecting from influenza H1N1 and H3N2.
The influenza virus can evolve according to two different processes. While 'antigenic-drift' relies on mutations that happen each time a virus is replicated, 'antigenic-switch' is a rare event which requires chromosome recombination: if two different types of viruses (e.g. H5N4 & H1N1) infect the same animal, chromosomes can be mixed which may lead to the production of an entirely new type of virus (e.g. H5N1).
If the proteins of a new influenza A virus were sequenced, how could the type of flu be identified? How could it be compared to recent seasonal flus and major 20th century flu pandemics, such as Spanish, Asian and Hong Kong flus?

For the most challenging task of multiple sequence alignment, no single approach will always perform best. As a consequence, different tools are made available: in addition to the popular Clustal Omega, EBI proposes Kalign, MAFFT, MUSCLE, T-Coffee, while NCBI offers COBALT. Finally, EBI gives access to a phylogenetic tree generation tool, Simple Phylogeny, which allows producing a tree from a multiple sequence alignment. Thinkbox 2 describes a typical problem which can be resolved by building such a tree.

Thinkbox 2: Phylogenetic analysis of the family elephantidae

Classification of species has traditionally been based on anatomy, where they were grouped according to shared physical characteristics. In the case of extinct species, this process relied mainly on bone and teeth structures because of their durability. DNA analysis* has provided an alternative way of classifying organisms; in some cases, this had led organisms to be reclassified.
Mammoths belong to the family Elephantidae that includes three living species: the African elephant (Loxodonta africana), the forest African elephant (Loxodonta cyclotis) and the Indian elephant (Elephas maximas). Mammoths and elephants belong to the order Proboscidea, which also includes the American mastodon (Mammut americanum).
Could the American mastodon be used as an outgroup for the study of Elephantidae?
Could the DNA analysis of those Proboscidea allow deciphering the evolutionary relationships between mammoths and living elephants?

Which living elephant is the most closely related to Mammoths?

*The cytochrome b gene is widely used for phylogenetic work and is seen as a universal metric, since it allows the comparison of independent studies.

Further reading

Yang H., et al. (1996), Phylogenetic resolution within the Elephantidae using fossil DNA sequence from the American mastodon (Mammut americanum) as an outgroup, PNAS 93:1190-4

Specialised on-line resources for 3D structure prediction and analysis

While NCBI and EBI provide quite a complete set of databases and tools allowing sophisticated sequence analysis, their provision in terms of structural data is mainly limited to those deposited in the PDB. Note that it has been estimated that fewer than 10% of human proteins have had their structure deciphered. Although the first protein structure was determined in 1958, experimental techniques remain time and cost consuming, moreover the 3D conformation of some proteins can still not be revealed by existing wetlab technologies. Since determination of a protein's native structure is a crucial step in the process of rational drug design, computational techniques are essential to predict the 3D structure of protein targets. If the 3D structure of a protein belonging to the family of the target is available, it can be used as a template to build an atomic-resolution model from the target sequence: homology modelling software is usually – but see Thinkbox 3 - able to infer very accurate conformations. While there is quite a few highly performing homology based 3D structure predictors, the SWISS-MODEL offered by Swiss Institute of Bioinformatics (SIB) is worth mentioning. Note that the SIB bioinformatics portal is also particularly rich in tools supporting 3D structure analysis. SWISS-MODEL is a fully automated protein structure homology-modelling server that has been cited tens of thousands of times and has been serving the international community for more than 25 years.

Thinkbox 3: Homology modelling; beware of chameleon sequences!

Homology modelling is the computational method of choice to predict the 3D structure of a protein the sequence of which is similar to a protein the structure of which is known. When their sequences display above 30% identity, their structures are expected to be highly similar. Predictions are usually highly accurate: structure qualities are often comparable to those determined by costly experimental methods. As a consequence, they play a crucial role in rational drug design.
However, some sequences, the so-called chameleon sequences, display different 3D conformations when their local environment or a small number of critical amino acids are modified. Yang et al. demonstrated that mutation of a single amino acid could be sufficient to convert a β-strand into an α-helix.
How would secondary or 3D structure prediction software handle chameleon sequences?
How does the existence of chameleon sequences affect the perceived validity of predictions produced by bioinformatics tools?
How cautious should one be of the outputs of computational methods?
Should the approach used to make sure that results from wetlab experiments are trustworthy be applied to results produced in silico?

Further reading

Yang W.Z., et al. (1998) Conversion of a B-strand to an α-Helix induced by a single-site mutation observed in the crystal structure of Fis mutant Pro26ala. Protein Sci. 7(9), 1875-1883

When a template from a related protein is not available, 3D structure prediction becomes extremely challenging: this is an active research field. Still, some tools perform better than others, even if their predictions require to be treated with both care and some scepticism. Among them, one should mention Iterative Threading ASSEmbly Refinement (I-TASSER) and QUARK from the Zhang Lab at the University of Michigan, and Rosetta (initially developed in the Baker lab at the University of Washington). Regarding further analysis of protein 3D structures, the Zhang Lab's bioinformatics portal is recommended because it provides a large variety of services for 3D structure analysis including structure refinement using molecular dynamics, protein binding site prediction and prediction of configuration of protein-ligand/protein-protein interactions, i.e. docking.

Domains of application

Since the completion of the sequencing of the first human genome in 2001 at a cost of $3 billion, biotechnology has made such progress that genome sequencing can now be produced for a fraction that amount. The era of the '$1,000 genome' has started and promises extraordinary advancement in predictive and personalised medicine since acquiring the genome of every single patient could soon become a reality in the most developed countries. Although there are enormous ethical and privacy issues associated to this, one can be confident that genome analysis will be routinely part of the future of medicine. Already, it has informed the treatment of some forms of cancer. Moreover, it is regularly used in genealogy studies and forensics, and mutation models inform the production of new vaccines.
The ability to sequence 'old DNA', i.e. several tens of thousands of years old, has led to the production of the genome of Neanderthal. Its analysis has impacted on the understanding of human evolution and anthropology: modern humans interbred with Neanderthals and most living humans contain Neanderthal genes. Rational drug design is delivering new treatments faster at much lower research and development costs. Genetic data analysis keeps benefiting more domains, even new ones such as nutrigenomics, which studies the relationship between food and genes. Some of which are particularly unexpected (see Thinkbox 4): recently, a new paradigm for video processing was proposed, where bioinformatics tools are applied to images captured by CCTV systems (Kazantzidis et al., 2018)!

We may be living in a unique moment of history: a bioinformatics revolution is starting, where most data, tools, educational resources and even processing facilities required to take part are available online. Let's join that revolution for the benefit of both humanity and indeed the whole planet.

Thinkbox 4: DNA analysis; a new dating method for geologists.

Nasikabatrachidae are purple frogs that were formally discovered in 2003 in Kerala (South West India). DNA analysis revealed that their closest relatives belong to Sooglossidae, a frog family that lives in the Seychelles, an Indian Ocean archipelago which is closer to the African coast than India. Would a frog have swum 2000 miles in the India Ocean to migrate from the Seychelles to India? Further DNA investigation has suggested that those two families separated from each other around 100 million years ago. On the other hand, plate tectonics theory describes Gondwana as a supercontinent where Africa and India were in contact. Gondwana eventually broke up during the Jurassic (160 to 180 million years ago) leading to the separation of the Seychelles and India 65 million years ago.
Does DNA analysis support the plate tectonics theory?
Should geology be added to the list of the application domains of genetics analysis?
Are there other unanticipated scientific domains which could benefit from such analysis?

<u>Further reading</u>

Janani S.J., et al. (2017) A new species of the genus Nasikabatrachus (Anura, Nasikabatrachidae) from the eastern slopes of the Western Ghats, India. Alytes 34 (1-4): 1-19

References

Needleman, S.B. and Wunsch, C.D. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins Journal of Molecular Biology, 48(3), pp.443–53.

Dayhoff, M.O., Eck, R.V., Chang, M.A. and Sochard, M.R. (1965): Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, Maryland.

Pevsner, J. (2015): Bioinformatics and Functional Genomics, Wiley Blackwell, 3rd ed., ISBN-10: 1118581784 - ISBN-13: 9781118581780

Lesk, A. (2013): Introduction to Bioinformatics, OUP Oxford, 4th ed., ISBN-10: 0199651566 - ISBN-13: 978-0199651566

Davies, K. (2010): The $1,000 Genome, New York: Free Press, ISBN 1-4165-6959-6

Richards, J.E. and Hawley, R.S. (2010): The Human Genome Academic Press, 3rd ed., ISBN-10: 0123334454, ISBN-13: 978-0123334459

Kazantzidis, I., Florez-Revuelta, F., M. Dequidt, M., Hill, N. and J.-C. Nebel (2018): Vide-omics: A Genomics-inspired Paradigm for Video Analysis. Computer Vision and Image Understanding, 166:28-40

Further Reading

Abbass,J., Nebel. JC and Mansour, M. (2014): Ab initio Protein Structure Prediction: Methods and Challenges. in Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data, Wiley-Blackwell, ISBN: 978-1-118-13273-9, 32, pp. 703-724

Selzer, P.M., Marhoefer, R.J. and Koch, O (2018): Applied Bioinformatics: An Introduction. Springer International Publishing AG, 2nd ed., ISBN-10: 3319682997 - ISBN-13: 978-3319682990

Abbass, J. and J.-C. Nebel, J-C. (2015): Customised fragments libraries for protein structure prediction based on structural class annotations BMC Bioinformatics, 16:136

Ko, K. H. (2016): Hominin interbreeding and the evolution of human variation. J Biol Res (Thessalon), 23: 17, doi: 10.1186/s40709-016-0054-7

Pavlidis, C. Nebel, J-C. Katsila, T. and Patrinos, G.P. (2016): Nutrigenomics 2.0: The Need for Ongoing and Independent Evaluation and Synthesis of Commercial Nutrigenomics Tests' Scientific Knowledge Base for Responsible Innovation. OMICS: A Journal of Integrative Biology, 20(2): 65-68

Esmaielbeiki, R. Krawczyk, K., Knapp, B., Nebel J-C and Deane, C.M. (2016): Progress and Challenges in Predicting Protein interfaces. Briefings in Bioinformatics, 17(1):117-131, 2016

Useful web-links

Ab initio protein structure prediction with Rosetta: https://www.rosettacommons.org/software

Ab initio protein structure prediction with QUARK: https://zhanglab.ccmb.med.umich.edu/QUARK

Basic Local Alignment Search Tool, BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi

Clustal Omega, https://www.ebi.ac.uk/Tools/msa/clustalo

European Bioinformatics Institute, EBI: https://www.ebi.ac.uk

Fully automated protein structure homology-modelling server, SWISS-MODEL: https://swissmodel.expasy.org

Molecular Modeling Database, MMD: https://www.ncbi.nlm.nih.gov/structure

National Center for Biotechnology Information, NCBI: https://www.ncbi.nlm.nih.gov

Online Mendelian Inheritance in Man, OMIM®: https://omim.org

Optimal Global Pairwise Sequence Alignment, EMBOSS Needle: https://www.ebi.ac.uk/Tools/psa/emboss_needle

Optimal Local Pairwise Sequence Alignment, EMBOSS Water: https://www.ebi.ac.uk/Tools/psa/emboss_water

Phylogenetic tree generation, Simple Phylogeny: https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny

Protein Data Bank, PDB: http://www.wwpdb.org

NCBI Reference Sequence Database, RefSeq: https://www.ncbi.nlm.nih.gov/refseq

Iterative Threading ASSEmbly Refinement, I-TASSER: https://zhanglab.ccmb.med.umich.edu/I-TASSER

UniProt Protein Knowledgebase, UniProtKB: https://www.uniprot.org