Bringing bioinformatics into the classroom

Bioinformatics

Food Detective

A PRACTICAL GUIDE

Version: 8 June 2020

Food Detective

Overview

This Practical Guide introduces the idea of computers as tools to help understand aspects of biology. In particular, it looks at how DNA sequences can be used to identify specific organisms, why this is important in the food industry, and how this can be used to help detect food fraud. Analyses are run online using sequence data from the 4273pi project website: **4273pi.org**.

Teaching Goals & Learning Outcomes

This Guide introduces a popular Web-based tool for searching biological sequence databases, and shows how to identify different species based on their specific DNA sequences – their 'barcodes'. On reading the Guide and completing the exercises, you will be able to:

- explain what is meant by DNA barcoding;
- search biological sequence databases using the online program BLAST;
- judge the reliability of database-search results in terms of their statistical significance; and
- evaluate the biological implications of search results with reference to food safety.

1 Introduction

The advent of computers and computer-driven technologies has opened new opportunities for, and has given unprecedented power to, biological investigations. Studies that once took months or years to complete may now be performed in hours or days. In the 1940s and '50s, for example, it took *ten years* to determine the **amino acid** sequence of the first **protein**[1-3] – **insulin** – before automatic methods in the late 1970s made it possible to sequence proteins in more practical time-scales. Technologies for rapid, high-throughput **DNA** sequencing and for *storing* DNA sequences didn't become available until the 1980s. The first DNA sequence database – which was called the EMBL data library – was publicly released in June 1982: at that time, it held 568 sequences[4]. Today, we take it for granted that we have instant access not just to hundreds but to *millions* of biological sequences (both DNA and protein), and to software tools and **algorithms** for analysing them, allowing us to perform comprehensive searches in minutes, and detailed analyses within hours.

The interdisciplinary discipline that encompasses these developments – fusing aspects of molecular biology, statistical analysis and computer science to gather, store, analyse and visualise biological data on a vast scale – is *bioinformatics*[5]. Continued advances in computer technology mean that bioinformatics approaches are now used routinely to underpin evolutionary studies, genetic and genomic research, personalised medicine, environmental and forensic analyses, and much more, in ways that were never possible before.

The ability to sequence DNA and to store the resulting **nucleotide** sequences in databases means that we can compare human **gene** sequences not just with each other but also with those of other animals; this allows us to identify differences between them. However, as already mentioned, the volume of information stored in sequence databases is huge. Just consider that the haploid human **genome** contains around 3 billion base pairs; and the human genome sequence is just one of thousands of completely- and partially-sequenced genomes currently stored in our databases. We therefore need to recruit computers to help us process this information.

This Guide explores how a simple bioinformatics tool can be used to search one of the world's largest nucleotide sequence databases, and to identify the organisms from which specific DNA sequences derive. We start with a set of DNA fragments taken from a sausage, described by a UK butcher as '100% pork sausage'. We use database searches to investigate what meat the sausage actually contains: is it 100% pork, or does it contain material from other animals as well?

2 About this Guide

This Guide introduces a popular software tool for searching DNA sequence databases: it shows how we can use the tool to determine the organisms from which DNA sequences derive, and illustrates the value of this technology for the food industry. Exercises are provided to help use the search tool via its Web interface. Throughout the text, key terms – rendered in **bold** type – are defined in boxes. Additional information is provided in supplementary tables and figures.

KEY TERMS

Algorithm: a set of steps or 'recipe' for solving a particular problem
Amino acid: an organic acid, one of 20 common, naturally occurring building-blocks of proteins
DNA: deoxyribonucleic acid, a molecule comprising two chains that coil together, forming a double-helix that carries genetic information
Gene: a molecular unit of heredity, broadly corresponding to a piece of DNA (or **RNA**) that encodes a protein (or functional RNA)
Genome: the entirety of an organism's genetic information, encoded as either DNA or RNA (in viruses)
Insulin: a hormone that regulates carbohydrate & fat metabolism
Nucleotide: a chemical base (one of 4 building-blocks of DNA & RNA) linked to a molecule of sugar & a molecule of phosphoric acid
Protein: organic compound containing one or more linear polymers of amino acids; existing in globular or membrane-bound forms, proteins participate in virtually all cellular processes
RNA: ribonucleic acid, a molecule comprising a long, unbranched chain of ribonucleotides (*i.e.,* nucleotides whose sugar component is ribose)

3 What is DNA?

DNA is a chemical that carries genetic information – in fact, it contains all the instructions that living organisms need in order to grow, reproduce and function. It's made up of two very long molecules that wind around each other, forming a spiral structure known as a 'double helix', as shown in Figure 1. The building blocks of each molecule of DNA – its chemical 'bases', or nucleotides – are Adenine, Thymine, Guanine and Cytosine: A, T, G and C. Particular regions of DNA sequences – termed **coding regions** – may encode proteins: in these regions, the order of the nucleotide bases determines the nature and order of the protein's constituent amino acids. This relationship, between the order of bases in DNA and the sequence of amino acids in proteins, is the basis of the **genetic code**.

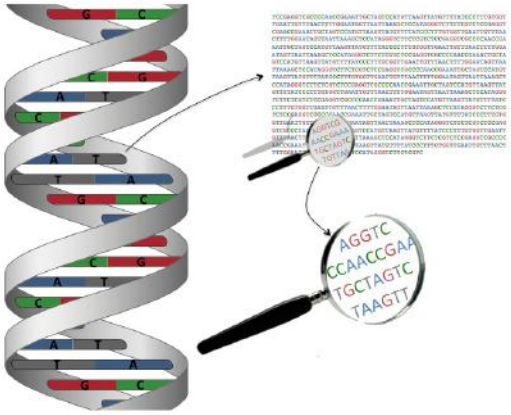

Figure 1 ***DNA molecules winding together to form a double helix****. Each Adenine (A) base from one molecule pairs with each Thymine (T) of the other, & each Guanine (G) pairs with each Cytosine (C). The order of bases in one of the DNA strands can be stored in a database as a sequence of As, Ts, Gs & Cs, which can then be analysed by computers.*

3.1 What can DNA tell us?

Every person is different, essentially because every individual's DNA is unique, as illustrated in Figure 2. Similarly, the DNA of every species differs from every other. This allows us to use DNA as a kind of identifying 'signature' for different species and for different individuals within species.

Figure 2 ***Each individual's DNA is unique.*** *We all look & behave differently because we each have a different genetic make-up; this is true of the DNA of humans & of the DNA of all other species.*

3.2 What is a DNA barcode?

While, overall, each individual has a unique DNA signature, there are nevertheless regions of DNA that are common to all animals, but that vary between species. These regions are termed 'barcodes'. These are rather like the barcodes used in supermarkets: each product has a barcode, but differences between them allow different items to be distinguished when scanned at the tills. QR codes function in a similar way – see Figure 3.

Figure 3 ***DNA barcoding.*** *There are regions of DNA that are present in all species but have distinct differences between species. These can be used to distinguish organisms, rather like the way barcodes & QR codes are used to identify individual entities.*

In just the same way as supermarket barcodes are used to identify individual products, DNA barcodes can be used to identify specific organisms. This has particular applications in the identification of species that look very similar in ecological analyses (especially some larvae that can't be identified in any other way); it's also widely used to test product authenticity: *e.g.*, timber, fish, leather and crops.

4 Food fraud

You may remember the 'horse-meat scandal' reported in parts of Europe in 2013, when foods advertised as containing beef were discovered to contain various amounts of (undeclared or improperly declared) *horse meat* – up to 100% in some cases! At the time, although 'horse meat' made the headlines, some products were also found to contain other undeclared meats, like pork.

The 2013 scandal should have led to sustained improvements in food labelling and advertising; nevertheless, the problem didn't entirely go away. For example, in Scotland in 2017, around 8% of meat in random test samples taken from supermarkets, restaurants and manufacturers, was found to contain DNA from an animal, or animals, not labelled as ingredients[6]. Lamb dishes, kebabs, sausages and some pizza toppings were among the mis-represented products. To give an idea of the type of 'contaminations' discovered, some 'pork' spare ribs were found to be chicken; some 'lamb' curries were shown to contain peanuts and beef, but no lamb; 'pork' sausages were found that contained pork and beef; a 'chicken' stir-fry contained turkey and chicken; 'beef' mince contained beef and pork; and minced 'lamb' was shown to contain lamb and chicken.

KEY TERMS

Coding region: the portion of a DNA sequence that encodes the information required to create a protein or RNA product

Codon: a group of three nucleotides that signals a specific amino acid, or the start or end of a coding region, according to its base sequence

Genetic code: the set of rules cells use to translate information sequestered in genetic sequences (*i.e.*, within **codons**) into proteins

Although, in some cases, the data were regarded as consistent with minor levels of cross-contamination during food processing, in others, the levels were significantly higher – sufficient to indicate *genuine food fraud*. Discoveries like these are made possible by the fact that horses, sheep, cattle, pigs and so on, each have DNA that is specifically and recognisably different, which in turn has allowed government scientists to use tools like DNA barcoding in their efforts to detect food fraud and contamination.

5 Becoming a food detective

Our investigation involves trying to identify DNA sequences found in a sample of allegedly '100% pork sausages' bought from a butcher's shop. DNA was extracted from the sausages; it was then sequenced and the barcode sequences were located. We'll compare the actual DNA barcodes from this sausage sample to known DNA sequences. The barcode sequences have no **annotation**: they're simply called *Sequence_A* to *Sequence_H*. We'll use these sequences to search a public DNA database. The format of the sequences is called FastA format (shown in **Figure 4**): this is just a concise way of storing biological sequences for efficient database searching.

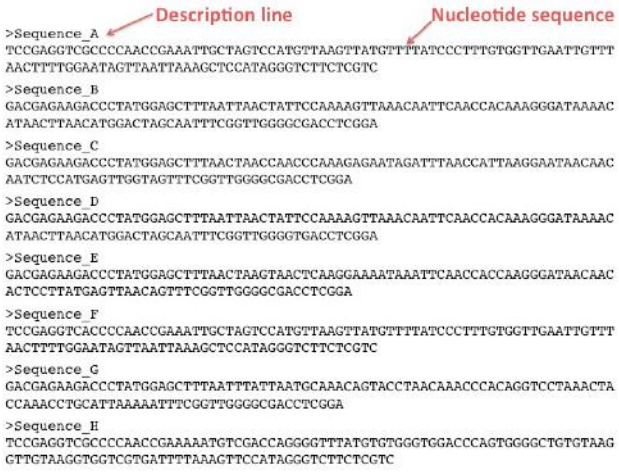

Figure 4 ***Sequences in FastA format***. *The file begins with a short descriptor or label that identifies the sequence, preceded by the '>' symbol, & is followed by the sequence itself in single-letter (A, T, C, G) notation. The sequence of only one strand of the double helix is stored (the sequence of the other strand can be inferred from base-pairing rules).*

Your task is to identify the DNA sequences present in the meat, and hence to answer the question, is the sausage 100% pork?

EXERCISES

1 Open a Web browser.
2 Go to the 4273π website: 4273pi.org. Click on the '*SCHOOLS*' tab (4273pi.org/schools), where you'll find a link to a set of Food Detective sequences, A to H.
3 Copy the first sequence, including its descriptor line ('*>Sequence_A*').
4 Keep the browser window open during the next tasks, so that you can copy & paste any of the sequences whenever you need to.

5.1 The database search tool, BLAST

To try to identify the species to which these sequences belong, we will search a database for sequences that show a high degree of similarity to each DNA barcode sequence. To do this, we'll use a popular online search tool – called BLAST[7,8] – at the **NCBI**. The NCBI maintains vast databases that contain most known DNA sequences, and makes them and their search tools (some of which are shown in **Figure 5**) freely available for public use. The program that we'll use is called blastn, which is designed to search nucleotide sequence databases with nucleotide sequence queries, like sequences A to H.

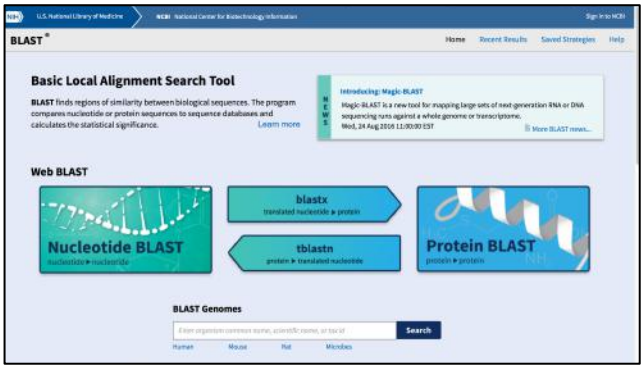

Figure 5 ***The NCBI BLAST homepage***. *There are several BLAST programs, each adapted for a different type of search, depending on whether the starting point is a nucleotide or a protein sequence.*

EXERCISES

1 Open a new tab in your Web browser.
2 Do a Web search for '*NCBI BLAST*'.
3 Check that the first result has a name like '*BLAST: Basic Local Alignment Search Tool*'. If it does, click on the link.
4 Click on '*Nucleotide BLAST*'.
5 In the box labelled '*Enter Query Sequence*', paste in *Sequence_A*.
6 Under '*Program Selection*', '*Optimize for*' (see **Figure 6**), select '*Somewhat similar sequences (blastn)*'.
7 At the bottom of the page, you'll see a button labelled '*BLAST*'. Click on the button to submit the search.

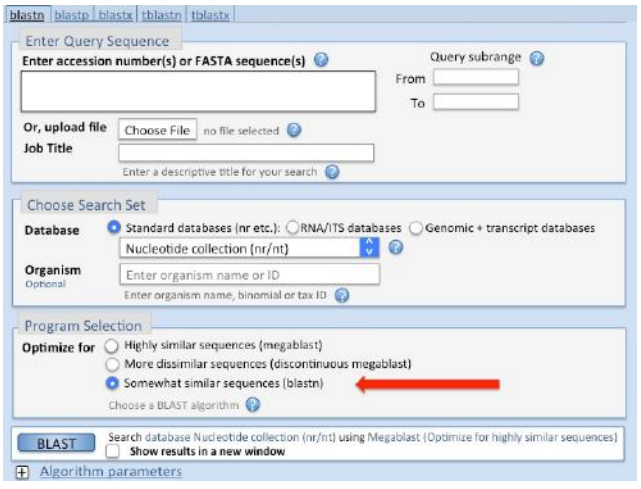

Figure 6 ***The NCBI BLAST blastn program***. *When using blastn for this exercise, select 'Somewhat similar sequences', under Program Selection, as indicated by the red arrow.*

KEY TERMS

Annotation: notes included within database entries to make them both informative & re-usable
NCBI: National Center for Biotechnology Information, part of the National Library of Medicine of the National Institutes of Health, USA

The blastn program may take a few minutes to run, during which time the *Status* on the Web page is shown as *Searching*. Don't refresh the page – please be patient. The results will eventually appear on an output page headed *BLAST >> blastn suite >> results for…*, as illustrated in **Figure 7**.

Initially, the output may appear daunting. However, the results are essentially broken down into two parts. The first section lists the original search **parameters** and a set of filtering options. The second section lists the best matches found in the database, giving more specific details: *e.g.*, the name or *Description* of each sequence; the extent of overlap between the query and the matched sequence (the greater the *Query cover* the better the match); the ***E value***, which indicates the reliability or significance of the match; the % identity (*Per Ident*); and the ***Accession number*** of the matched sequence – clicking on this number directs you to its full entry in the **GenBank**[9-10] database (note: to obtain further, more detailed information about BLAST and how to use it, there are many comprehensive tutorials and guides available online[11-15]).

Note that the descriptions of the best-matching sequences listed in this table (*i.e.*, on the left-hand side of the output) use the scientific (Latin) names of each matched organism (*Sus scrofa*, *Homo sapiens*, *etc.*). For the purpose of this exercise, you'll need to identify the common name for each organism.

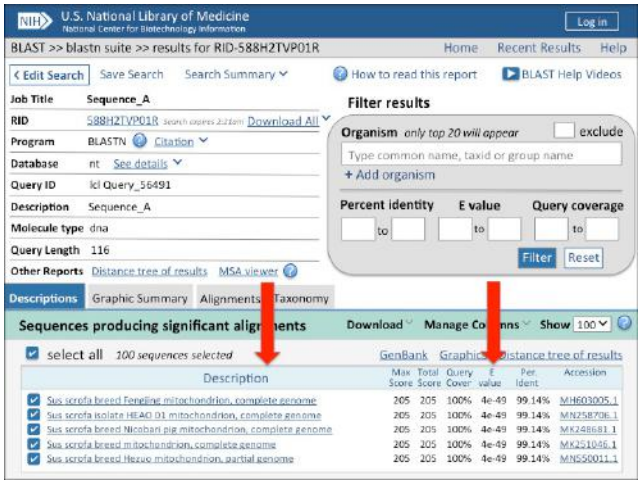

Figure 7 ***Typical features of a blastn output file****. The top of the file lists the original search parameters, alongside a set of filtering options. Beneath this section are listed the descriptions of the best-matching sequences found in the database, together with their respective E-values, as indicated by the red arrows.*

EXERCISES

1 When the search is complete, scroll to the table of results, shown beneath the original search parameters (follow the red arrows in **Figure 7**). The first result in the table is the best match to *Sequence_A* in the database.

2 Make a note of the species from which this best match comes. Record the result in Table 1 below. The species name is found in the '*Description*' column of the BLAST results table (shown beneath the left-hand arrow). Remember, species names are in their scientific/Latin form. If you're unsure about the common name, refer to Table 2 or use a Web search to find out.

3 Also record the E-value of the best match (indicated beneath the right-hand arrow).

4 Repeat the BLAST search for Sequences B to H individually, & complete Table 1.

Table 1 ***BLAST results for sequences A to H.***

Sequence	Species *scientific* name	Species common name	*E-value*
A			
B			
C			
D			
E			
F			
G			
H			

Table 2 ***Examples of scientific names.***

Species *scientific* name	Species common name
Abramites hypselonotus	Marbled headstander fish
Bos taurus	Cattle
Gallus gallus	Chicken
Homo sapiens	Human
Methanothermobacter spp.	Bacteria
Neocalanus cristatus	Copepod
Ovis aries	Sheep
Scinax sp.	Snouted tree frog
Sus scrofa	Pig
Vitis vinifera	Common grape vine

EXERCISES

1 What do your results in Table 1 tell us about the DNA in the sausage? Does the meat seem to be 100% pork?

2 Do any of your results seem unexpected? Explain.

3 Are your results *really* unexpected? Think about how sausages are made & how DNA is extracted.

Consider why we might find other animal DNA in a pork sausage. Our sausages came from a butcher's shop. Butchers handle many different animal products, so traces of their meat could have been picked up from the counter-tops, mincing machines, chopping knives, *etc*. Also, sausages are handmade, so some traces of human DNA may have been transferred during the manufacturing process. In fact, *any food* – not just sausages – is likely to contain traces of human DNA, albeit at sufficiently low levels not to cause alarm! Finally, the DNA samples themselves could have been contaminated in the lab during the extraction and sequencing processes. In the sausage used for this study, almost all sequences were found to be from pig (note that sequences A to H are a non-random sample).

KEY TERMS

Accession number: a unique (generally invariant) computer-readable number or code that identifies a particular entry in a database

E-value: or Expect value, the number of matches of this quality (or better) that are expected to occur simply by chance (the closer the value to 0, the better the match)

GenBank: the nucleotide sequence database maintained at the NCBI; the database was first released in December 1982

Parameter: a value given to a program to tell it what to do (*e.g.*, the type & name of database to be searched, or which algorithm to use)

To recap, you have used a program from the suite of BLAST tools – specifically blastn – to search one of the world's largest DNA sequence databases (GenBank). The blastn algorithm took each sequence that you pasted into the search box and compared it with every sequence in GenBank, then listed the results in a table showing the best matches at the top. The question is, how is the order of matches in this table calculated? How can we know that the results are reliable and biologically meaningful?

GenBank is an enormous database: it contains more than 200 million sequences! As a result, when we perform searches, our query sequence may sometimes match one or more of the database sequences by chance. We measure this effect (*i.e.*, the number of times we'd expect to see a match of the same quality between our query sequence and a sequence in the database simply by chance) using a parameter called the *Expect* or E value. The higher the E-value, the more unreliable the match; the smaller the E-value, the more reliable the match: an E-value of 0, or close to zero, is extremely reliable.

BLAST results show E-values in a way that may be unfamiliar to you: *e.g.*, 5.2e-15. But this is just a convenient and more efficient way of representing numbers that have a lot of digits: 5.2e-15 is equivalent to $5x10^{-15}$, which takes up a lot less space than having to write 0.000000000000005!

Ultimately, the reliability of database search results is very important, and E-values are one of the ways in which we can get an indication of how reliable BLAST results really are.

5.2 Exploring the reliability of database matches

In the first part of this practical exercise, we discovered the species from which DNA sequences A to H came by using the database search tool, blastn. In the next part of the exercise, we'll consider the reliability of database matches by looking more closely at their E-values. There is a difference between (perhaps) unexpected results with *high* reliability and unexpected results with *low* reliability.

EXERCISES

1 Return to the 4273π website: 4273pi.org. Click on the '*SCHOOLS*' tab (4273pi.org/schools), & follow the link to the set of Food Detective sequences labelled I to M (illustrated in **Figure 8**).

2 As before, copy each sequence in turn, including its descriptor line ('*>Sequence_I', etc.*) & run blastn at the NCBI. Recap: paste each sequence into the box labelled '*Enter Query Sequence*'; under '*Program Selection*', '*Optimize for*', select '*Somewhat similar sequences (blastn)*'; then click on the '*BLAST*' button to submit the search.

3 When the search is complete, scroll to the table of results shown beneath the original search parameters, & identify the best match (the first match in the results table).

4 For each best match, record its scientific & common names, alongside its E-value, in Table 3.

Table 3 ***BLAST results for sequences I to M.***

Sequence	Species *scientific* name	Species common name	*E-value*
I			
J			
K			
L			
M			

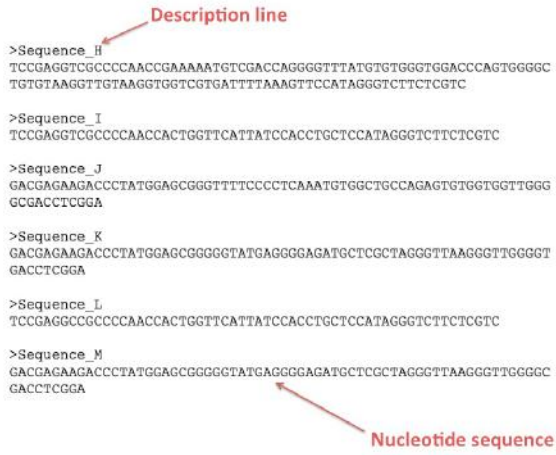

Figure 8 ***Sequences I to M in FastA format****. Remember to include the description line when you cut & paste the sequences for input to blastn.*

EXERCISES

1 Compare the E-values of the sequences in Table 1 to those you've recorded for the sequences in Table 3. Which table contains sequences with the highest E-values?

2 Which table allows us to identify the species from with the DNA sequences originate with the greatest reliability? Why?

Following the first component of our investigation, we have considerable confidence that the various animals identified – cow, human, *etc.* – really *do* have some DNA in the pork sausage. By contrast, in the second component, we *do not* believe that the various organisms identified have DNA in the sausage.

The reason for the difference in our confidence in the reliability of the two sets of results is that the E-values in the second task are not reliable – they are not *statistically significant.* We draw this conclusion because the E-values of the matched sequences listed in Table 3 are very much higher than those in Table 1. Remember, the closer the value to 0, the better (more significant) the match – a value of 0 would indicate an extremely good match.

A likely explanation for the difference in significance between the results in Tables 1 and 3 is that the sequences used in the second task contain errors – the bottom line is, DNA-sequencing machines aren't perfect! So, whenever you search a sequence database (like GenBank, or any other sequence database), always pay close attention to the E-values of returned matches, and use these to determine whether the results are statistically reliable, and whether the results are biologically meaningful or relevant, given the context of your study.

TAKE HOMES

Having completed these exercises, you can now:

1 Search a biological sequence database with a DNA barcode;

2 Use the NCBI's *BLAST* program, blastn, to search one of the world's largest nucleotide sequence databases, GenBank;

3 Determine the reliability of BLAST search results in terms of their statistical significance – *i.e.*, with respect to their E-values;

4 Use DNA sequences to identify their species of origin;

5 Evaluate the biological implications of search results with reference to food safety.

6 References & further reading

1 Sanger F. (1945) **The free amino groups of insulin.** *Biochem. J.*, **39**, 507-515.
2 Sanger F *et al.* (1955) **The amide groups of insulin.** *Biochem. J.*, **59**(3), 509–518.
3 Ryle AP *et al.* (1955) **The disulphide bonds of insulin.** *Biochem. J.*, **60**(4), 541–556.
4 Hamm GH & Cameron GN. (1986) **The EMBL data library.** *Nucleic Acids Res.*, **14**(1), 5-9.
5 Torrance J *et al.* (2012) **Higher Biology.** Hodder Gibson.
6 Howell D. (2018) **Curry and pizza among failed Scottish meat tests.** BBC News, **www.bbc.co.uk/news/uk-scotland-45800465**
7 Altschul SF *et al.* (1990) **Basic local alignment search tool.** *J. Mol. Biol.*, **215**(3), 403-410.
8 Altschul SF *et al.* (1997) **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.*, **25**(17), 3389-3402.
9 Burks C *et al.* (1985) **The GenBank nucleic acid sequence database.** *Comput. Appl. Biosci.*, **1**(4), 225-233.
10 Sayers EW *et al.* (2020) **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res.*, 48(D1), D9-D16.
11 **BLAST Guide: Overview of the various NCBI BLAST services & reports: ftp.ncbi.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf**
12 **NCBI Tutorials: www.ncbi.nlm.nih.gov/home/tutorials**
13 **The New BLAST Results Page: Enhanced graphical presentation and added functionality: ftp.ncbi.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf**
14 Pertsemlidis A & Fondon JW 3rd. (2002) **Having a BLAST with bioinformatics (and avoiding BLASTphemy).** *Genome Biology*, **2**, reviews 2002.1–2002.10.
15 Attwood TK & the GOBLET Foundation. (2018) **A Critical Guide to BLAST.** [version 1; not peer reviewed]. *F1000Research*, **7**, 1435 (document) (**doi.org/10.7490/f1000research.1116052.1**).

7 Acknowledgements & funding

GOBLET Practical Guides build on GOBLET's Critical Guide concept, using layout ideas from the Higher Apprenticeship specification for college-level students in England. The contents expand on materials made freely available by the 4273π project (**4273pi.org**).

This Guide was developed with the support of a donation from EMBnet to the GOBLET Foundation.

Design concepts and the Guide's front-cover image were contributed by CREACTIVE.

8 Licensing & availability

9 Disclaimer

About the organisations

GOBLET

GOBLET (Global Organisation for Bioinformatics Learning, Education & Training; **www.mygoblet.org**) was established in 2012 to unite, inspire and equip bioinformatics trainers worldwide; its mission, to cultivate the global bioinformatics trainer community, set standards and provide high-quality resources to support learning, education and training.

GOBLET's ethos embraces:

- ***inclusivity**: welcoming all relevant organisations & people*
- ***sharing**: expertise, best practices, materials, resources*
- ***openness**: using Creative Commons Licences*
- ***innovation**: welcoming imaginative ideas & approaches*
- ***tolerance**: transcending national, political, cultural, social & disciplinary boundaries*

Further information can be found in the following references:

- Attwood *et al.* (2015) **GOBLET: the Global Organisation for Bioinformatics Learning, Education & Training.** *PLoS Comput. Biol.*, **11**(5), e1004281.
- Corpas *et al.* (2014) **The GOBLET training portal: a global repository of bioinformatics training materials, courses & trainers.** *Bioinformatics*, **31**(1), 140-142.

GOBLET is a not-for-profit foundation, legally registered in the Netherlands: CMBI Radboud University, Nijmegen Medical Centre, Geert Grooteplein 26-28, 6581 GB Nijmegen. For general enquiries, contact **info@mygoblet.org**.

EMBnet

EMBnet, the Global Bioinformatics Network, is a non-profit organisation, founded in 1988 to establish and maintain bioinformatics services in Europe. Eventually expanding beyond European borders, EMBnet created an international network to support and deliver bioinformatics services across the life sciences: **www.embnet.org**.

Since its foundation, EMBnet has had a keen interest in Education and Training (E&T), and has delivered tutorials and courses worldwide. Perceiving a need to unite and galvanise international E&T activities, EMBnet was a principal founder of GOBLET. For more information and general enquiries, contact **info@embnet.org**.

The 4273π Project

4273π provides a freely available, customised distribution of Raspbian GNU/Linux for the Raspberry Pi computer. 4273π is for those wishing to teach, learn or use bioinformatics on the Raspberry Pi. The project focuses on bioinformatics education, computational science, open educational resources, public engagement and democratisation of science: **4273pi.org**.

CREACTIVE

CREACTIVE, by Antonio Santovito, specialises in communication and Web marketing, helping its customers to create and manage their online presence: **www.gocreactive.com**.

About the authors

Stevie Anne Bain

Stevie A Bain is a Postdoctoral Research Associate in Bioinformatics Education at the University of Edinburgh. She is currently responsible for the running, delivery and development of the 4273π bioinformatics education project. Stevie is an evolutionary biologist with a research background in genomics and epigenetics.

Daniel Barker

Daniel Barker is a Reader in Bioinformatics and Director of the MSc Bioinformatics programme at the University of Edinburgh. He is particularly interested in bioinformatics education, computational phylogeny and philosophy of science. With colleagues, he released the 4273π SD card for Raspberry Pi as an Open Educational Resource in 2013. As part of the 4273π project, he has been helping bring a version of the workshop presented in this GOBLET resource to schools and teachers around Scotland since 2016.

Affiliation

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh (UK).

Teresa K Attwood (orcid.org/0000-0003-2409-4235)

Teresa (Terri) Attwood is emeritus Professor of Bioinformatics, having more than 25 years' experience teaching bioinformatics in undergraduate and postgraduate degree programmes, and in *ad hoc* courses, workshops and summer schools, in the UK and abroad.

Focusing on protein sequence analysis (particularly G protein-coupled receptors), she created the PRINTS database, co-founded InterPro, and co-developed tools for sequence analysis, and for linking research data and scientific articles (Utopia Documents).

She wrote the first introductory bioinformatics text-book; her third book was published in 2016:

- Attwood TK & Parry-Smith DJ. (1999) ***Introduction to Bioinformatics.*** Prentice Hall.
- Higgs P & Attwood TK. (2005) ***Bioinformatics & Molecular Evolution.*** Wiley-Blackwell.
- Attwood TK, Pettifer SR & Thorne D. (2016) ***Bioinformatics challenges at the interface of biology and computer science: Mind the Gap.*** Wiley-Blackwell.

Affiliation

Department of Computer Science, The University of Manchester, Oxford Road, Manchester (UK).