Journal of Biological Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjbe20

Bioinformatics: a history of evolution in silico

Vladan Ondřej [a] & Petr Dvořák [a]

[a] Department of Botany, Faculty of Science, Palacký University in Olomouc, Šlechtitelů 11, Olomouc, 783 71, Czech Republic

Version of record first published: 24 Sep 2012.

To cite this article: Vladan Ondřej & Petr Dvořák (): Bioinformatics: a history of evolution in silico , Journal of Biological Education, DOI:10.1080/00219266.2012.716776

To link to this article: http://dx.doi.org/10.1080/00219266.2012.716776

Practical

Bioinformatics: a history of evolution *in silico*

Vladan Ondřej and Petr Dvořák

Department of Botany, Faculty of Science, Palacký University in Olomouc, Šlechtitelů 11, Olomouc, 783 71, Czech Republic

Bioinformatics, biological databases, and the worldwide use of computers have accelerated biological research in many fields, such as evolutionary biology. Here, we describe a primer of nucleotide sequence management and the construction of a phylogenetic tree with two examples; the two selected are from completely different groups of organisms: hominids and cyanobacteria. The hominid group involves genetic information not only of recent species, but also from fossilised ancestors of modern humans. Nucleotide sequences are used to construct phylogenetic trees. In the case of the cyanobacteria group, the nucleotide sequences obtained from fossilised cyanobacteria were searched across databases to find similar sequences and identify recent relatives. This demonstration of evolutionary history provides real examples of the bioinformatics approach in biological research; additionally, it can be utilised as practical exercises in biological studies, suitable for secondary or tertiary-level classes.

Keywords: bioinformatics; nucleotide sequences database; phylogenetic tree; hominids; cyanobacteria

Introduction

Computers, the internet, and databases have dramatically changed biological research. During the last decade, theoretical and computational biology have produced a flood of new biological data, from genomics to evolutionary biology. Research that used to start in the laboratory now starts on the computer, as scientists search databases for information that might suggest new hypotheses.

In the last three decades, personal computers have become accessible not only across all disciplines of science, but they have also become valuable tools for science education. The easy access of students to computers, and recently the free access to biological databases, opens up new possibilities in the teaching of genetics, taxonomy, and evolutionary biology. Bioinformatics, an application of information technology to the assessment of biological data, allows an interdisciplinary approach in the teaching of genetics, taxonomy, and evolutionary biology (Gibas and Jambeck 2001).

In the following paragraphs, we demonstrate some bioinformatics exercises, based on searching for DNA sequences, processing these sequences, and the construction of a phylogenetic tree (Box 1). All three steps not only require basic biological knowledge, but also some basic computer skills. Sequence databases contain large volumes of the DNA, RNA, or protein sequence data of many different kinds of organisms. There are major sequence data collections and deposition sites in Europe, Japan, and the USA (Gibas and Jambeck 2001). These public databases often offer some free software for searching and analysing the sequence data, and aid in finding related scientific publications. For the purpose of these exercises, it is necessary to have selected sequences. Specific genes or genome regions, such as cytochrome b, ribosomal genes, or internal transcribed spacers are used to build phylogenic trees. Here, we propose two simple and inexpensive examples of practical bioinformational tasks, focused on the assumptions of evolution, using sequence-based phylogeny. The

Corresponding author: Vladan Ondřej, Department of Botany, Faculty of Science, Palacký University in Olomouc, Šlechtitelů 11, Olomouc 783 71, Czech Republic. Email: vladan.ondrej@upol.cz

Journal of Biological Education ISSN 0021-9266 print/ISSN 2157-6009 online © 2012 Society of Biology
http://www.tandfonline.com
http://dx.doi.org/10.1080/00219266.2012.716776

educational objectives of these two tasks include the following:

- Students will learn to understand basic facts about the hypotheses of evolution, using DNA sequence data.
- Students will learn to think about particular ways in which the evolution of hominids and cyanobacteria might be investigated.
- Students will gain some elementary knowledge of how researchers assess evolutionary relationships within both fossil and recent living organisms.

Box 1. Phylogenetic tree: a theory

The phylogenetic tree shows the genealogical relationships among different taxa (e.g. species, genera) in a graphical form (Yang 2006). The tree could possibly be constructed on the basis of morphology, the anatomy of the organisms, and DNA sequence data (which currently is the most widely used). The topology of a tree consists of two elementary parts: nodes and branches. A node is a point where the branches merge together (Figure 7); the length of the branches represents the evolutionary distance. The tree with a known root is called a rooted tree, while one without is an unrooted tree. A common way of root determination is by the defining of the outgroup, which is specified as a distantly related species to all others assessed in the tree (these are called the ingroup).

The building of a phylogenetic tree consists of three elementary steps. (1) Acquisition of sequences from either original data or some database (e.g. NCBI). (2) Multiple sequence alignment, using software (e.g. ClustalX) or a manually built matrix. (3) Tree construction in some phylogenetic software.

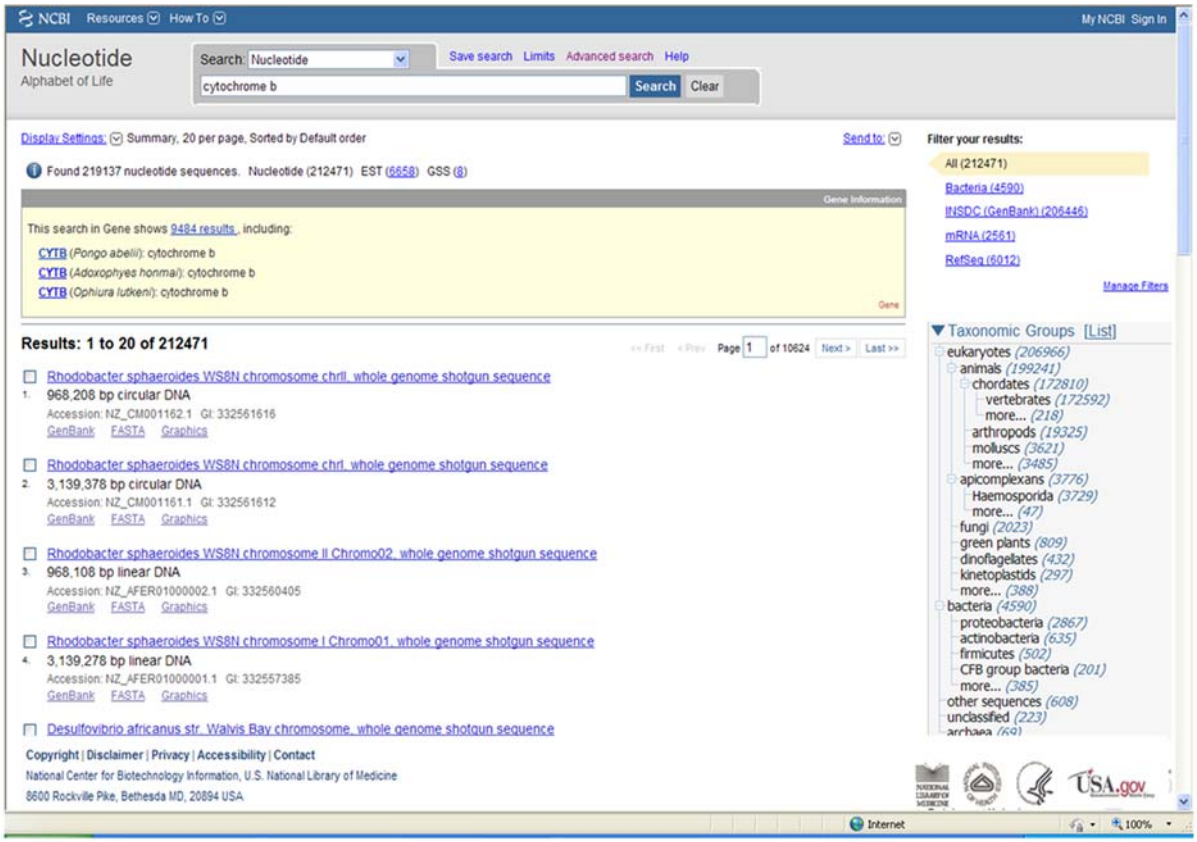

Figure 1. Results of query for cytochrome b on the nucleotide search page of the GenBank http://www.ncbi.nlm.nih.gov/. Going through taxonomic groups the number of results are reduced and sequences of hominids are selected

Table 1. Summarised information of selected sequences for phylogenetic tree construction (cytochrome b) and sequence similarity searching across database (16S rRNA). Information contains accession numbers and species (with isolate origin) or sample names

Cytochrome b		16S rRNA	
Accession number	Species	Accession number	Cyanobacteria sample
JF940522	*Homo sapiens* Sweden	FJ809898	MT3-11 clone 1
JF938916	*Homo sapiens* Portuguese	FJ809904	MTK3-H1 clone 2
JF939049	*Homo sapiens* Armenian		
JF906114	*Homo sapiens* Indian		
253947345	*Homo sapiens neanderthalensis* Mezmaiskaya1		
253947317	*Homo sapiens neanderthalensis* El Sidron		
253947289	*Homo sapiens neanderthalensis* Feldhofer1		
315466581	*Homo sp. altai* Denisova molar		
HM068587	*Pan troglodytes*		
5835135	*Pan paniscus*		
5835149	*Gorilla gorilla*		
5835834	*Pongo abelii* Sumatran orangutan		
5835163	*Pongo pygmaeus* Bornean orangutan		
49146236	*Macaca mulatta*		

```
LOCUS       JF940522               1141 bp    DNA     linear   PRI 21-MAY-2011
DEFINITION  Homo sapiens haplogroup T2f1a mitochondrion, complete genome.
ACCESSION   JF940522 REGION: 14738..15878
VERSION     JF940522.1  GI:333034453
KEYWORDS    .
SOURCE      mitochondrion Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1141)
  AUTHORS   Pike,D.A., Barton,T.J., Bauer,S.L. and Kipp,E.
  TITLE     mtDNA Haplogroup T Phylogeny Based on Full Mitochondrial Sequences
  JOURNAL   J Genet Geneal 6 (1), 1-24 (2010)
REFERENCE   2  (bases 1 to 1141)
  AUTHORS   Greenspan,B.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-MAY-2011) Family Tree DNA - Genealogy by Genetics,
            Ltd., 1445 North Loop West, Suite 820, Houston, TX 77008, USA
FEATURES             Location/Qualifiers
     source          1..1141
                     /organism="Homo sapiens"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /haplogroup="T2f1a"
                     /note="origin_locality:Sweden"
     gene            1..1141
                     /gene="CYTB"
     CDS             1..1141
                     /gene="CYTB"
                     /note="TAA stop codon is completed by the addition of 3' A
                     residues to the mRNA"
                     /codon_start=1
                     /transl_except=(pos:1141,aa:TERM)
                     /transl_table=2
                     /product="cytochrome b"
                     /protein_id="AEF12485.1"
                     /db_xref="GI:333034466"
                     /translation="MTPMRKINPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQ
                     ITTGLFLAMHYSPDASTAFSSIAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGL
                     YYGSFLYSETWNIGIILLLLATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDL
                     VQWIWGGYSVDSPTLTRFFTFHFILPFIIAALAALHLLFLHETGSNNPLGITSHSDKI
                     TFHPYYTIKDALGLLLFILSLMTLTLFSPDLLGDPDNYTLANPLNTPPHIKPEWYFLF
                     AYTILRSVPNKLGGVLALLLSILILAWIPILHMSKQGSMMFAPLSQSLYWLLAADLLI
                     LTWIGGQPVSYPFTIIGQVASVLYFTTILILMPTISLIENKMLKWA"
ORIGIN
        1 atgaccccaa tacgcaaaat taacccccta ataaaattaa ttaaccactc attcatcgac
       61 ctcccaaccc catccaacat ctccgcatga tgaaacttcg gctcactcct tggcgcctgc
      121 ctgatcctcc aaatcaccac aggactattc ctagccatac actactcacc agacgcctca
      181 accgcctttt catcaatcgc cccacatcact cgagacgtaa attatggctg aatcatccgc
      241 taccttcacg ccaatggcgc ctcaatattc tttatctgcc tattcctaca catcggacga
      301 ggcctatatt acggatcatt tctctactca gaaaccctgaa acatcggcat tatcctcctg
      361 cttgcaacta tagcaacagc cttcataggc tatgtcctcc cgtgaggcca aatatcattc
      421 tgaggggcca cagtaatttac aaacttacta tccgccatcc catacattgg gacagacata
      481 gttcaatgaa tctgaggagg ctactcagta gacagtccca ccctcacacg attcttacc
      541 tttcacttca tcttgccctt cattattgca gccctagcgg cactccacct cctattctg
      601 cacgaaacgg gatcaaacaa ccccctagga atcacctccc attccgataa aatcaccttc
      661 caccctatt acacaatcaa agacgccctc ggcttacttc tcttcattct ctcctaatg
      721 acattaacac tattctcacc agacctccta ggcgacccag acaattatac cctagccaac
      781 ccctaaaca ccctcccca catcaagccc gaatgatatt tcctattcgc ctacacaatt
      841 ctccgatccg tccctaacaa gctaggaggc gtccttgccc tattactatc catcctcatc
      901 ctagcaataa tccccatcct ccatatatcc aaacaacaaa gcataatatt tcgcccacta
      961 agccaatcac tttattgact cctagccgca gacctcctca ttctaacctg aatcggagga
     1021 caaccagtaa gctaccctt tacatcatt ggacaagtag catccgtact atactccaca
     1081 acaatcttaa tcctaatacc aactatctcc ctaattgaaa acaaaatact caaatgggcc
     1141 t
//
```

Figure 2. Example of flat file format of the searching result. This format contains not only a sequence and its translation to amino acids, but also information about type of DNA, sample and reference

```
>gb|JF940522.1|:14738-15878 Homo sapiens haplogroup T2f1a mitochondrion, complete genome
ATGACCCCAATACGCAAAATTAACCCCCTAATAAAATTAATTAACCACTCATTCATCGACCTCCCCAACCC
CATCCAACATCTCCGCATGATGAAACTTCGGCTCACTCCTTGGCGCCTGCCTGATCCTCCAAATCACCAC
AGGACTATTCCTAGCCATACACTACTCACCAGACGCCTCAACCGCCTTTTCATCAATCGCCCCACATCACT
CGAGACGTAAATTATGGCTGAATCATCCGCTACCTTCACGCCAATGGCGCCTCAATATTCTTTATCTGCC
TATTCCTACACATCGGACGAGGCCTATATTACGGATCATTTCTCTACTCAGAAACCTGAAACATCGGCAT
TATCCTCCTGCTTGCAACTATAGCAACAGCCTTCATAGGCTATGTCCTCCCGTGAGGCCAAATATCATTC
TGAGGGGCCACAGTAATTACAAACTTACTATCCGCCATCCCATACATTGGGACAGACATAGTTCAATGAA
TCTGAGGAGGCTACTCAGTAGACAGTCCCACCCTCACACGATTCTTTACCTTTCACTTCATCTTGCCCTT
CATTATTGCAGCCCTAGCGGCACTCCACCTCCTATTCTTGCACGAAACGGGATCAAACAACCCCCTAGGA
ATCACCTCCCATTCCGATAAAATCACCTTCCACCCTACTACACAATCAAGACGCCCTCGGCTTACTTC
TCTTCATTCTCTCCTTAATGACATTAACACTATTCTCACCAGACCTCCTAGGCGACCCAGACAATTATAC
CCTAGCCAACCCCTTAAACACCCTCCCCACATCAAGCCCGAATGATATTTCCTATTGGCCTACACAATT
CTCCGATCCGTCCCTAACAAGCTAGGAGGCGTCCTTGCCCTATTACTATCCATCCTCATCCTAGCAATAA
TCCCCATCCTCCATATATCCAAACAACAAAGCATAATATTTCGCCCACTAAGCCAATCACTTTATTGACT
CCTAGCCGCAGACCTCCTCATTCTAACCTGAATCGGAGGACAACCAGTAAGCTACCCTTTTACCATCATT
GGACAAGTAGCATCCGTACTATACTTCACAACAATCTTAATCCTAATACCAACTATCTCCCTAATTGAAA
ACAAAATACTCAAATGGGCCT
```

Figure 3. An example of the FASTA format of selected sequence

Box 2. How does molecular evolution work?

The genetic information of all recently living and fossil organisms is preserved in molecules of DNA inside almost every cell. When the DNA is replicated, some mistakes occur, because of e.g. a DNA polymerase error, or recombination among homologous parts of chromosome during sexual reproduction. This creates some degree of variability within a population. Without this variability, no evolution would occur. A phylogenetic tree based on DNA sequences visualizes such variability by the characterization of differences among the sequences. Essentially, the greater the distances between sequences within a tree, the more evolutionarily distant those sequences are.

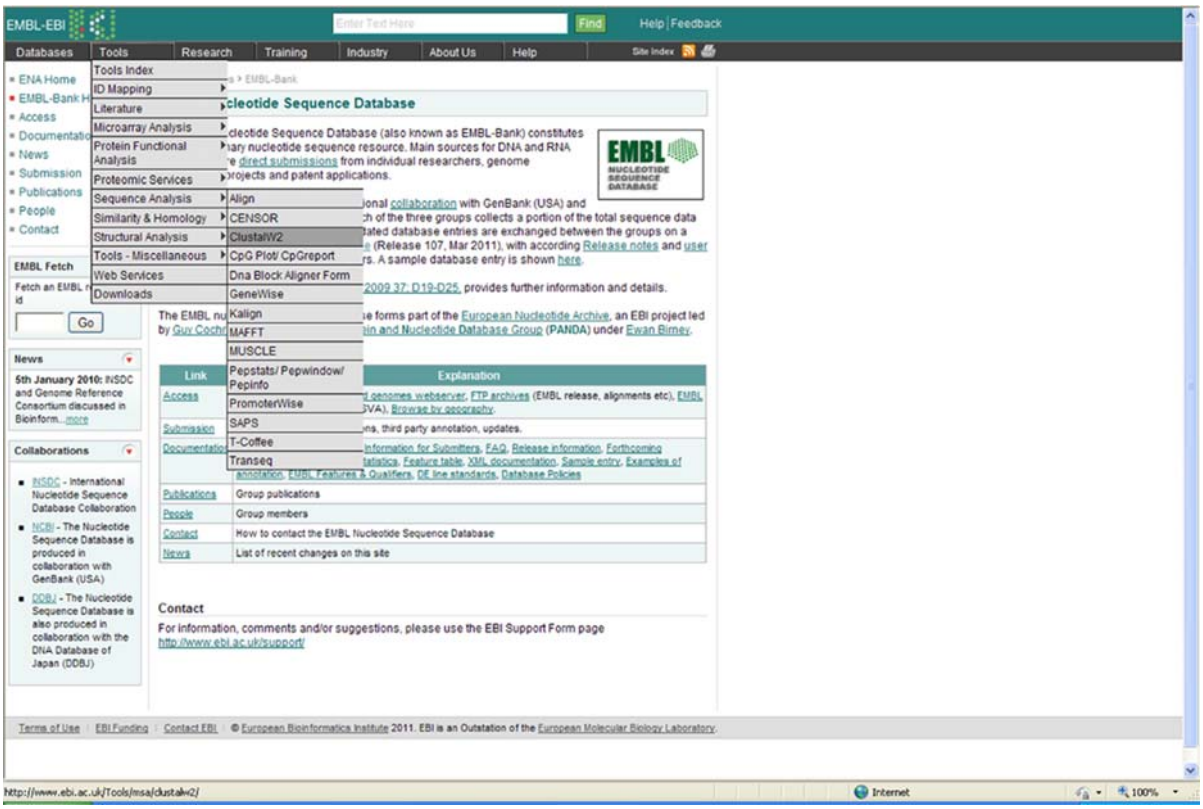

Figure 4. The screenshot of the website http://www.ebi.ac.uk/ with online application of ClustaW2 used for multiple sequence alignment of hominid sequences

Hominid phylogeny: Exercise 1 in managing sequences and the construction of a phylogenetic tree

The first exercise is based on the construction of a phylogenetic tree (for a summary, see Box 1) involving human and near human relatives. The sequences of the cytochrome b gene were chosen. To find the sequences, a query for cytochrome b is sent on the nucleotide search page of GenBank (http://www.ncbi.nlm.nih.gov/nuccore) (Figure 1, Box 3). The search may produce thousands of results, but the selection of the results could be done through a tree of the taxonomic groups: animals, vertebrates, mammals, placental, primates, Hominidae. This approach also leads to a repetition of the systematic biology of organisms and their taxonomy. The following selection depends on the purpose and experience of the scientist, and is time consuming. This is the reason why it is necessary to have prepared selected sequences before the exercise (Table 1). In the group Hominidae, the students will find sequences of cytochrome b from *Gorilla*, *Pongo*, both species of chimpanzees and *Homo sapiens*. Additionally, recent molecular techniques are able to isolate DNA from fossils; thus, the sequences of *Homo sapiens neanderthalensis* is also represented in the database. The exercise could also be extended by the sequence of the poorly described archaic hominid group from Siberia, simply marked as *Homo sp. altai* (Reich et al. 2010). Students could then compare the genetic relationship of this group of humans with both Neanderthal and modern humans. To construct a phylogenetic tree, an outgroup must be added. An outgroup contains sequences of distant relatives, or a possible ancestor of the studied group. In this case, the sequences of the macaque were chosen.

Sequences can be downloaded from the databases in two common formats. The first one, the GenBank flat file format (Figure 2), that was used in GenBank earlier in its history, contains information about gene identity, the conditions under which it was characterised, references, and sequence, together with translation to amino acid sequences. The second format, FASTA (Figure 3), is a simple format containing a single comment line that begins with a > character, followed by a single-character DNA sequence, on as many lines as needed to contain the sequence, without breaks. Of course, some information associated with the gene is lost in the FASTA format, but this format is very useful for following sequence processing.

The sequences selected and prepared by the teacher are used for multiple sequence alignments. One commonly used program for progressive multiple sequence alignment is Clustal X (the stand-alone version) and Clustal W (the online version, see Figure 4). They are able to align medium-sized data sets very quickly, and are easy to use (Larkin et al. 2007). The sequences in FASTA format that have to be aligned are simply copied into the given field of the Clustal W application (Figure 5). The alignments are

Box 3. The most important web pages involved in the basic bioinformatical applications

- http://www.ncbi.nlm.nih.gov/ – NCBI (National Center for Biotechnology and Information). A generally important web page where DNA and protein sequences as well as whole genome projects are stored. All information may be freely searched and downloaded. Moreover, there are a lot of tools for sequence data handling (see more details below).
- http://www.ncbi.nlm.nih.gov/nuccore – NCBI nucleotide search. A basic tool for searching DNA sequence collections. Exercise 1 explains how to perform a search using this tool.
- http://www.ebi.ac.uk/Tools/msa/clustalw2/ – Clustal W represents the most popular multiple sequence alignment algorithm. It is implemented in a variety of online and standalone programs which are freely available. This online tool also allows users to construct a phylogenetic tree using UPGMA and NJ clustering methods. See more detailed description in Exercise 1#. Standalone version ClustalX2 may be downloaded from http://www.clustal.org/.
- BLAST (Basic Local Alignment Search Tool) may be accessed from the following sources: http://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html and http://blast.ncbi.nlm.nih.gov/Blast.cgi. This tool was developed for searching nucleotide or protein databases to find the most similar sequence to your query (Exercise 2).

Other important links:

- http://www.ebi.ac.uk/2can/home.html – A free online project focused on basic concepts in molecular biology and bioinformatics designed in a practical manner. It contains a comprehensive collection of tutorials, and detailed explanations of some tools mentioned within this article.
- http://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html – This web page contains a lot of useful tools for handling DNA and protein sequences. Format Converter tool should be mentioned, because it represents a simple way of converting alignment files into different formats for phylogeny.
- http://www.megasoftware.net/ – MEGA 5 is a new release of standalone user friendly software which allows inexperienced or medium-experienced users to perform all the necessary steps for phylogenetic analysis. It is able to carry out a database search (NCBI), multiple sequence alignment, and inference of a phylogenetic tree.

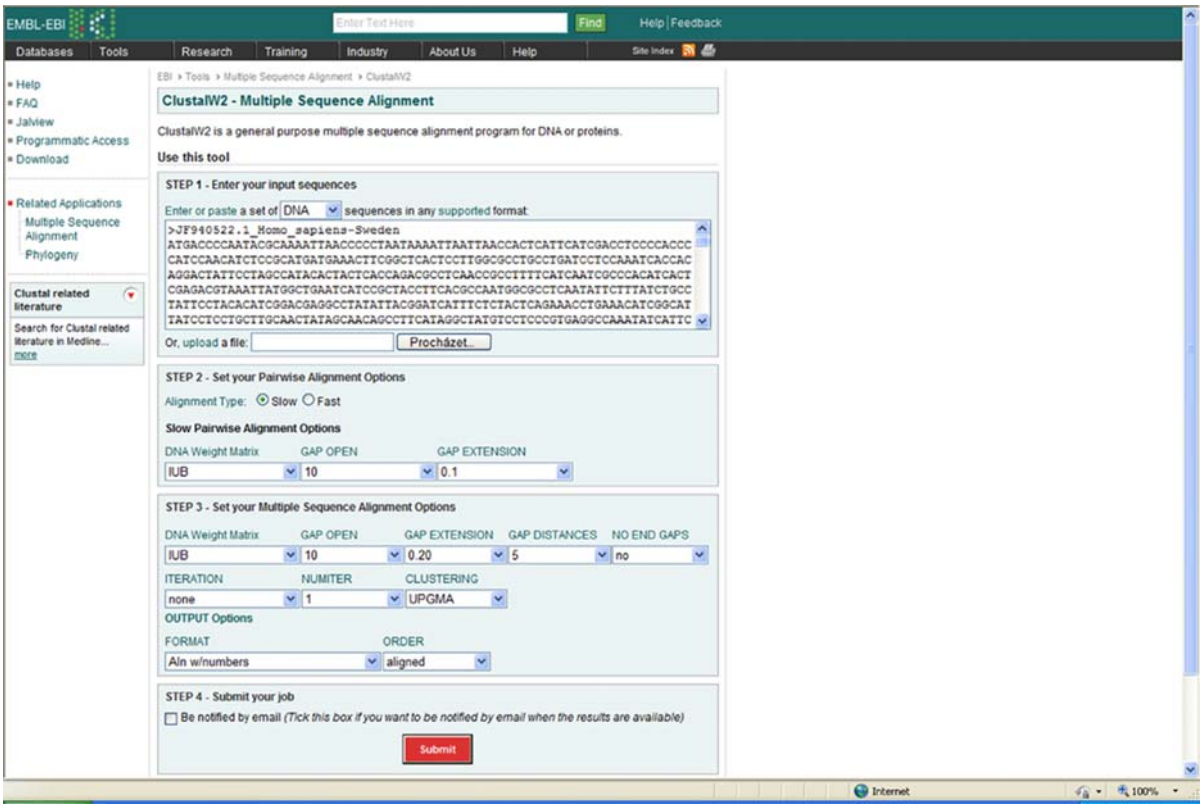

Figure 5. The screenshot http://www.ebi.ac.uk/ of multiple sequence alignment parameters also involving UPGMA algorithm used for hominid sequences

of sufficient quality that they very often do not require manual editing or adjustment, except the ends of used sequences with different lengths. Thus, these ends have to be cut off to get alignment of the

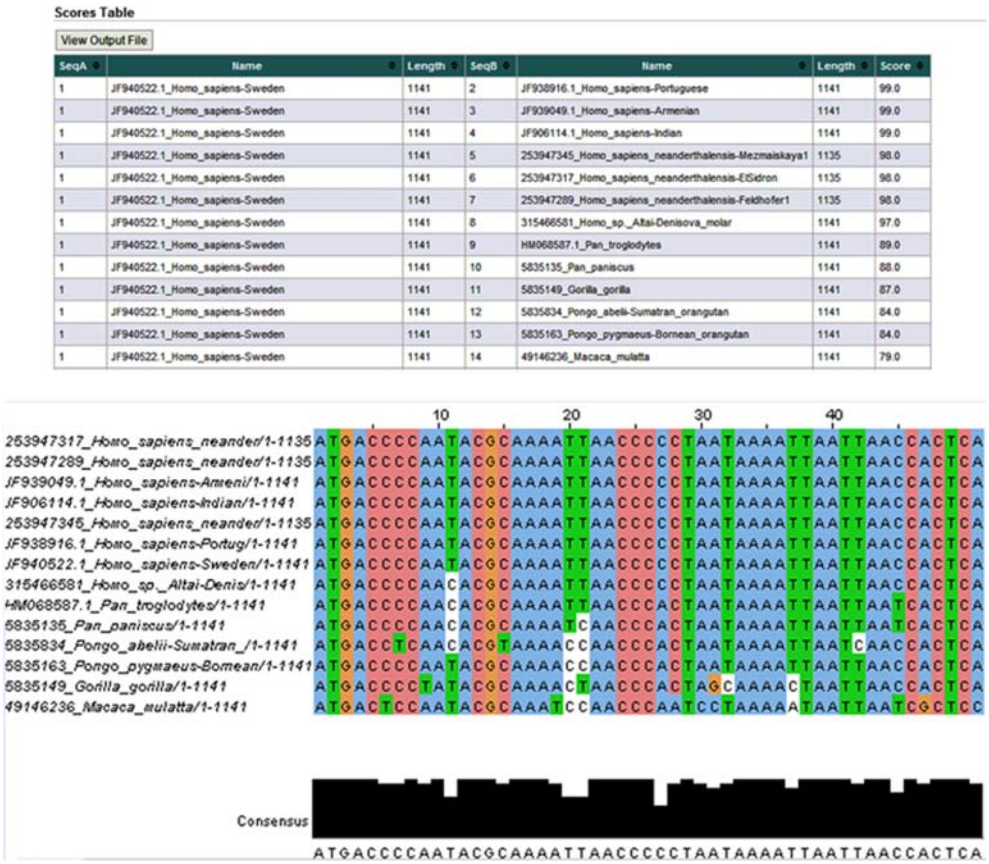

Figure 6. Online view of scores table showing per cent similarity between aligned sequences and coloured multiple alignment result displaying changes in sequences – mutations – between comparing species

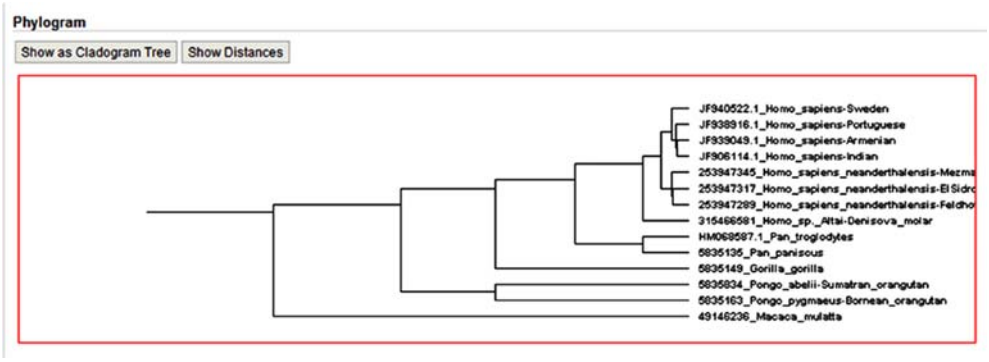

Figure 7. Constructed phylogram originating from selected cytochrome b sequences of hominids

sequences with the same length. Nonetheless, Clustal W and Clustal X continue to be widely used (and increasingly so) on websites. The EBI Clustal site (Box 3) gets literally millions of multiple alignment jobs per year (Larkin et al. 2007), and is used here in the described bioinformatics exercise: http://www.ebi.ac.uk/Tools/msa/clustalw2/ (Figure 4). For this purpose, the pairwise and multiple alignment parameters were set to default. This program uses two alignment algorithms for phylogenic tree construction, based on distance matrix methods: (1) the unweighted pair group method, using arithmetic averages (UPGMA); and (2) the neighbour-joining (NJ) method (Larkin et al. 2007). In the described exercise, UPGMA was used (Figure 5).

The results of multiple sequence alignments (Figure 6) demonstrate to students the basis of evolution, changes in genetic materials, and mutations (Box 2). Students can see the point mutations and insertions/deletions within sequences of related species, and retrieve information about how far away or close the studied species are (by comparison of sequence similarity scores). After that, the students can view the phylogenetic tree (see Figure 7) and start a discussion.

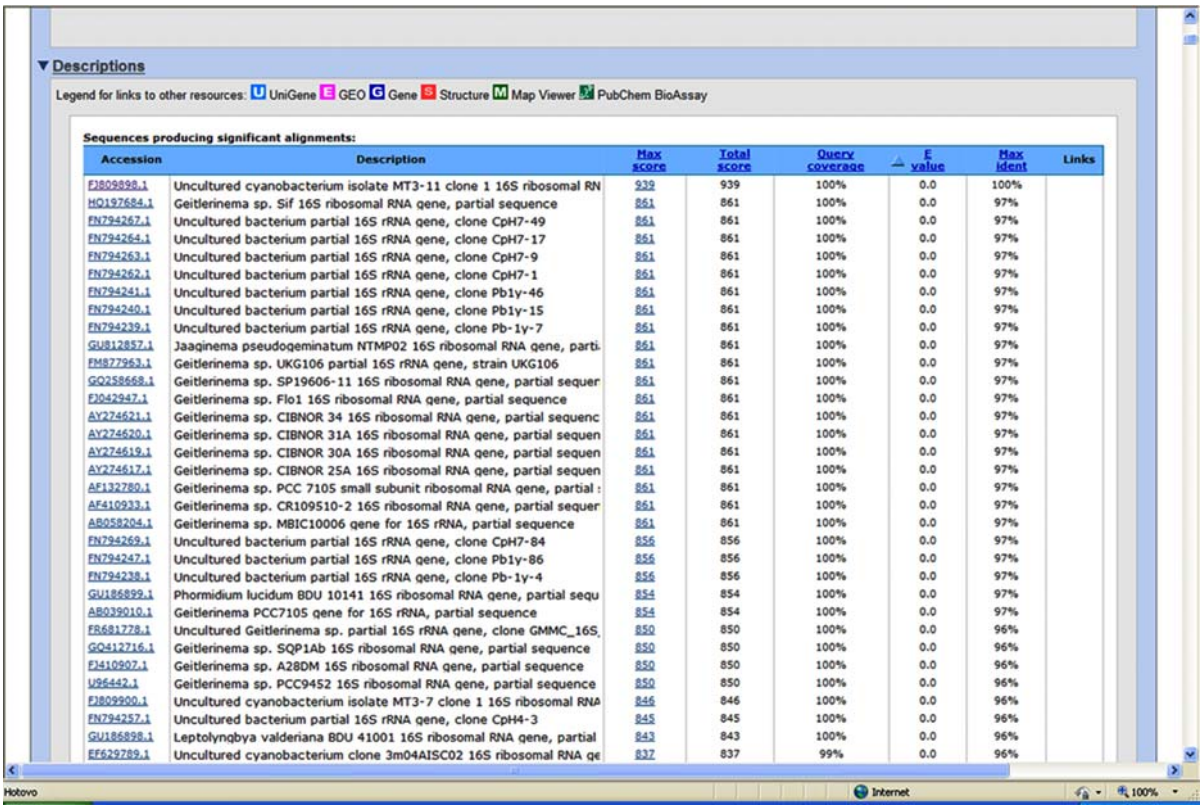

Figure 8. Online view of the BLAST results based on searching for similar sequences in the database to sequences of isolates of the fossilized cyanobacteria. Maximal identity is showed by per cent

The tree obtained (Figure 7) reflects recent knowledge about the evolution of hominids, based on morphological and genetic analyses. In this approach, no significant genetic variations in *H. sapiens*, represented by samples from three different nations, were observed. Different molecular techniques are usually used to describe intraspecific genetic variability. Students will find that the closest living relatives to humans are the chimpanzees. The tree also answers the questions about the archaic hominid group from Siberia, *H. sp. altai*. It has been shown that this group of people have a distinct evolutionary history from both the Neanderthals and modern humans, and that they are further genetically from modern humans than from Neanderthals, in an agreement with the literature (Reich et al. 2010).

Searching relatives to fossilised cyanobacteria: a local alignment searching exercise (Exercise 2)

The second exercise is based on searching for similar sequences. Here, we propose finding recent relatives of fossilised cyanobacteria from which the DNA was isolated from gypsum deposited in the late Miocene. Panieri et al. (2010) successfully extracted and amplified genetic material belonging to ancient cyanobacteria from gypsum crystals dating back as far as 5.910–5.816 Ma (when the Mediterranean had become a giant hypersaline brine pool). This finding represents the oldest ancient cyanobacterial DNA, to date.

A common application of sequence alignment is searching the database for sequences that are similar to a query sequence. The most popular tool for searching sequence databases is a program called BLAST (Basic Local Alignment Search Tool). BLAST is the algorithm at the core of most online sequence search servers, and can perform hundreds, even thousands, of sequence comparisons in a matter of minutes (Gibas and Jambeck 2001).

NCBI BLAST (Box 3), used for the exercise, is the most recent version at the website: http://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html. The teacher selects the sequences that are going to be 'BLASTed'. For the purpose of the exercise described, two sequences of the 16S rRNA gene of fossilised cyanobacteria were chosen. Students put the sequences into the website sequence query application form, mark 'searching' in the nucleotide database (DNA/RNA), and submit the query. The results are shown in a summary table (Figure 8). Here, students find 100% identity with the sequence itself – but what is important is that they also find similar sequences of recent species with high percentages of identity: *Geitlerinema* sp., 97% identity; *Jaaginema pseudogeminatum*, also 97% identity with sequence No. FJ809898; *Chroococcidiopsis* sp., 99% identity with sequence No. FJ809904. By clicking

on the accession number, it is possible to obtain complete information about the identified sequences – the sequence, their origin, and citations in the literature. For example, *Geitlerinema* sp. is a hot spring cyanobacteria, order Oscillatoriales, from Turkey (Izmir) identified by Zeliha Demirel (not published, submitted to the database in 2010); *Jaaginema pseudogeminatum* is a marine cyanobacteria, also order Oscillatoriales, from India, isolated by Thajuddin N., Pandiaraj D. and Mubarak Ali D. (not published, submitted to the database in 2010). *Chroococcidiopsis* sp., order Pleurocapsales, is a hypolithic cyanobacteria occurred in hyperarid deserts (Warren-Rhodes et al. 2007). These BLAST results correspond with the findings of Panieri et al. (2010). The environment of recent relatives of the fossilised cyanobacteria reflects the environment during the formation of the evaporation deposits during the Messinian Salinity Crisis. The recent relatives live in seas and hot springs with high salinity, and in subsurface locations in arid areas.

Conclusions

In these two examples, we were not studying the phylogeny of hominids nor of fossilised cyanobacteria, but they were presented here as motivating examples for use as bioinformatics exercises. Many other such exercises could be designed based on sequence searching and phylogenic tree constructions focused on various organisms – animals, plants, fungi, or pathogenic microbes (as an example). We have shown how easily computers can be used, along with public databases and bioinformatics programs, in the teaching of several of the fundamental pillars of biology: phylogeny and evolution, genetics, taxonomy, as well as ecology. A bioinformatics approach in selected biology lectures has an interdisciplinary character, developing complex thinking in the students, and stimulating them toward discussions and the creation of new hypotheses.

Acknowledgements

This work was supported by IGA UP Agency no. PrF_2011_003 and PrF_2012_001 and by Education for Competitiveness Operational Programme: grant no. CZ.1.07/2.2.00/15.0310.

References

Gibas, C., and P. Jambeck. 2001. *Developing bioinformatics computer skills*. Sebastopol, CA: O'Reilly and Associates.

Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics application notes* 23, no. 21: 2947–8.

Panieri, G., S. Lugli, V. Manzi, M. Roveri, B.C. Schreiber, and K.A. Palinska. 2010. Ribosomal RNA gene fragments from fossilized cyanobacteria identified in primary gypsumfromthe late Miocene, Italy. *Geobiology* 8, no. 2: 101–11.

Reich, D., R.E. Green, M. Kircher, J. Krause, N. Patterson, E.Y. Durand, B. Viola, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468: 1053–60.

Warren-Rhodes, K.A., K.L. Rhodes, L.N. Boyle, S.B. Pointing, Y. Chen, S. Liu, P. Zhuo, and Ch.P. McKay. 2007. Cyanobacterial ecologyacross environmental gradients and spatial scales in China's hot and cold deserts. *FEMS Microbiology Ecology* 61, no. 3: 470–82.

Yang, Z. 2006. *Computional molecular evolution*. Oxford: Oxford University Press.