Visualising DNA, RNA & Proteins in Jalview
School Workbook
### About this workbook:

The workbook contains 4 practical web-based bioinformatic projects:

**Project 1** views DNA and RNA sequences and their 3D structures.

**Project 2** views the sequence and 3D structure of human myoglobin protein, then compares this protein with myoglobin sequences from other animals using a tree based on sequence similarity.

**Project 3** views a range of proteins and considers how their 3D structures affects protein function.

**Project 4** views the exons and introns in the gene involved in sickle cell anaemia, and identifies the genetic mutations in the coding DNA sequence and its effect on the haemoglobin protein.

The workbook, the web links to run the exercises, as well as additional resources such as videos are available at the ‘Resources for Schools’ web page on the Jalview website: [http://www.jalview.org/school-resources](http://www.jalview.org/school-resources)

---

#### Who is this workbook for:
- **Secondary school biology pupils (aged 16-18 years old)**

#### Knowledge required:
- Moderate computer literacy

#### Equipment needed:
- A desktop or laptop computer with a web browser and internet access

---

### Getting Started

<table>
<thead>
<tr>
<th>Getting Started</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project 1-Viewing DNA &amp; RNA</td>
<td>3</td>
</tr>
<tr>
<td>Exercise 1: What is DNA?</td>
<td>5</td>
</tr>
<tr>
<td>Exercise 2: What are codons?</td>
<td>8</td>
</tr>
<tr>
<td>Exercise 3: What is RNA?</td>
<td>10</td>
</tr>
<tr>
<td>Project 2-Viewing Myoglobin Proteins</td>
<td>11</td>
</tr>
<tr>
<td>Exercise 4: Viewing human myoglobin protein</td>
<td>12</td>
</tr>
<tr>
<td>Exercise 5: Comparing myoglobin proteins from different species</td>
<td>13</td>
</tr>
<tr>
<td>Project 3-Viewing Proteins</td>
<td>15</td>
</tr>
<tr>
<td>Exercise 6: Viewing different classes of proteins</td>
<td>16</td>
</tr>
<tr>
<td>Project 4-Viewing the genetic mutation involved in sickle cell anaemia</td>
<td>20</td>
</tr>
<tr>
<td>Exercise 7: What are exons &amp; introns?</td>
<td>22</td>
</tr>
<tr>
<td>Exercise 8: View the coding sequence of the HBB gene &amp; its protein product</td>
<td>23</td>
</tr>
<tr>
<td>Appendix</td>
<td>27</td>
</tr>
<tr>
<td>Glossary</td>
<td>29</td>
</tr>
</tbody>
</table>

---

### What is Jalview?

Jalview is free-to-use computer software developed at the University of Dundee. It is designed to allow research scientists to visualise and analyse DNA, RNA, and proteins. It uses an interactive multi-window interface for viewing sequences, alignments, annotations, trees, and three-dimensional structures. Jalview can read files directly from public biological databases and has a number of analysis tools for aligning sequences, producing trees, measuring similarities, and comparing structures.

[www.jalview.org](http://www.jalview.org)

The workbook was produced by Dr Suzanne Duce with help from Mungo Carstairs, Benedict Soares, Bob Hanson, Dmitry Finkelbergs, Charlotte Campbell, Jim Procter & Geoff Barton

We would like to acknowledge the BBSRC and Wellcome Trust who have funded this work.

Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, DD1 5EH © 5th May 2021
Getting Started

Learning Objectives:
• Access the Jalview schools web page
• Open the Jalview Schools Workbook
• Launch JalviewJS

1. Open a web browser such as Chrome or Firefox. We suggest you avoid Internet Explorer as JalviewJS does not always work in older browsers.

2. Search using keywords 'Jalview Schools'. From the list of results, select 'Resources for Schools' at http://www.jalview.org/school-resources. This will open the Jalview Schools web page.

3. Select the 'View the School Workbook' link in Section (1) on the Jalview Schools web page.

4. The Schools Workbook will open in an adjacent tab. Scroll through the pages to view the different exercises.

5. Return to Jalview Schools web page. Click the link named 'View DNA fragment' in Section (2): Project 1-Exercise 1.

6. JalviewJS viewer opens in an adjacent window in the browser.

Change the appearance of the windows in JalviewJS:
(i) To move a window, place the mouse on title panel on the top of the window, then click-and-drag.
(ii) To enlarge a window, place the mouse on the lower right-hand corner of the window, then click-and-drag.
(iii) To close a window, click the ‘X’ in the top right-hand corner of the window.
Navigating Jalview's Windows

- **Desktop window**
- **Alignment window**
- **Split screen window**
- **DNA Sequences**
- **Protein Sequences**
- **Status bar**
- **Overview window**
- **Structure window**
- **Tree window**

Overview: http://www.jalview.org/tutorial/myoglobin.mfa

Overview: http://www.jalview.org/tutorial/myoglobin.mfa

Overview: http://www.jalview.org/tutorial/myoglobin.mfa

Overview: http://www.jalview.org/tutorial/myoglobin.mfa

Overview: http://www.jalview.org/tutorial/myoglobin.mfa

Overview: http://www.jalview.org/tutorial/myoglobin.mfa
Project 1: 'Viewing DNA & RNA'
DNA (deoxyribonucleic acid) is made up of two strands in a double helix. The two DNA strands are anti-parallel with respect to each other.

A DNA strand is a polymer. Its sub-units are called nucleotides. A nucleotide consists of a phosphate group, a 5-ring sugar, and a nitrogenous base (for more information see Table 1 in the Appendix).

Each carbon in the sugar ring is assigned a number. The base is attached to the 1’ carbon (reads 1 prime). The hydroxyl group is attached to the 3’ carbon. The phosphate group is attached to the 5’ carbon.

The backbone of a strand of DNA is made of deoxyribose sugars linked to phosphates by phosphodiester bonds. The bases are attached to the backbone as sidechains. There are four different bases: guanine (G), thymine (T), cytosine (C) and adenine (A).

The nucleotide bases between the two intertwining strands of DNA form weak hydrogen bonds. The adenine base aligns with the thymine base. They form 2 hydrogen bonds. The cytosine base aligns with the guanine base. They form 3 hydrogen bonds. These pairings are called complementary base pairings.

A DNA strand has polarity. One end is called the 3’ end and other is 5’ end. This relates to the position of the 3’ and 5’ sugar carbons.

In the figure opposite, the 3’ carbons beside the hydroxyl groups are coloured cyan. The 5’ carbons beside the phosphate groups are coloured purple.
Exercise 1: What is DNA?

Learning Objectives:
• Open DNA sequence and its 3D structure in Jalview
• Colour the nucleotide bases in the sequence
• View the 3D structure

1. Click the link named 'View DNA fragment' in Section (2): Project 1-Exercise 1 of the Schools web page.

2. JalviewJS with the DNA sequences and its 3D structure opens in an adjacent tab. This is the sequence from a fragment of B-DNA (PDB id 3BSE).

3. Select the Colour menu in the alignment window. Select the Nucleotide colour scheme.

4. The Jalview nucleotide colour scheme: adenine bases are green, cytosine bases are yellow, guanine bases are red and thymine bases are blue.

5. Click in the 3D structure window with the mouse, then drag the mouse to change the view. Notice how the adenine bases align with the thymine bases. The cytosine bases align with the guanine bases. These are complementary base pairings.

7. Jmol 3D mouse commands:
   (i) To rotate the structure place mouse on the structure, then click-and-drag.
   (ii) To zoom press the shift key, then click-and-drag.
   For more information about Jmol visit: http://wiki.jmol.org/index.php/Main_Page

Q: What are the names of the 4 different DNA bases, and their single letters identifier?
Q. How many nucleotide base pairs are there in this fragment of DNA? (Tip: count the base pairs in the 3D structure window)
Exercise 1: What is DNA?

8. Select the Colours menu in the Jmol structure window. Select the Purine/Pyrimidine colour scheme.

9. The purines (adenine and guanine bases) are coloured pink. The pyrimidines (thymine and cytosine bases) are coloured cyan.

10. In the alignment window, hover the mouse over the first base (adenine) in the top strand (A) and view its location on the structure. Next, hover the mouse over the first base (adenine) in the second strand (B) of the alignment and view its location on the structure.

Note where the first base in each sequence is located in the Jmol structure window (step 10).
Q. Are the DNA strands parallel or anti-parallel?
Q. Look at the shape of purine and pyrimidine, how are they different? (see Appendix Table 1)

11. In the Jmol structure window, Select the View menu. Select Show Chain. Uncheck 3BSE:B.

12. Drag the mouse to change the view, so you are looking down the spiral. (The first adenine base on the strand A should be at the top, check this by hovering the mouse over this base in the alignment). Review the direction in which the spiral turns.

Q. In Step 12, what direction is the strand turning, clockwise or anticlockwise?
Exercise 1: What is DNA?

The 3D structure of a molecule can be displayed in several different ways. View 1: DNA is commonly displayed as a ribbon cartoon. View 2: Alternatively, the atoms and bonds can be represented as balls and sticks. View 3: The shape and size of a molecule can be represented by the space-fill model which displays the atoms as spheres with the atoms’ radii equal to their Van der Waals radius.

**View 1**
Ribbon cartoon model
coloured with the Jalview nucleotide colour scheme.

**View 2**
Ball-and-stick model
atoms are displayed as balls coloured with the CPK colour scheme.

**View 3**
Space-fill model
displays atoms as spheres with atom radii equal to their Van der Waals radius.

Click link to view the 3D model in SketchFab:
https://skfb.ly/6ZzzZ

Click link to view the 3D model in SketchFab:
https://skfb.ly/6ZDzx

Click link to view the 3D model in SketchFab:
https://skfb.ly/6ZDzz

For more information about the DNA fragment 3BSE visit the Protein Data Bank website at:
www.rcsb.org/structure/3BSE
Exercise 2: What are codons?

Learning Objectives:
• View the DNA coding sequence for myoglobin protein alongside its protein product in a split-frame window
• Identify codons (triplet of nucleotide) that code for an amino acid residue during protein synthesis

1. Click the link named 'View coding sequence of DNA and codons' in Section (2) Project 1-Exercise 2 on the Schools web page.

Note: It may take a little while for Jalview to open the files depending on the speed of the internet.

2. A split-screen window opens containing the coding DNA sequence (upper panel) and the myoglobin protein sequence (lower panel).

3. The DNA and protein sequence panels are linked. Place the mouse over an amino acid in the lower panel and the associated nucleotide triplet or codon is highlighted in a black box in the upper panel. Note: Use scroll bar to move the alignment from left to right.

4. Place the mouse on the leucine amino acid L at residue 3 in the lower panel, and note the associated 3 nucleotide bases (codon) highlighted by the black box in the upper panel. Repeat for leucine at residue 10, and leucine at residue 12. Note: Status box provides additional information.
Exercise 2: What are codons?

Q. When an amino acid residue is selected in the protein sequence, why are three nucleotide bases highlighted by a black box in the DNA coding sequence panel?
Q. What DNA triplet bases are associated with leucine at residue 3, 10 and 12?
Q. Would you expect them to be the same? (see Codon Table below)

5. **Click on the 3D structure window** to bring it to the front. **Place the mouse over any amino acid residues** in the lower panel of the split-frame alignment window.

Can you see the location of the amino acid in the 3D structure?

What triplet bases (codon) codes for this amino acid residue?

### Codon Table: A codon is a set of three nucleotides, or triplet, that code for a specific amino acid residue during protein synthesis.

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>F</td>
</tr>
<tr>
<td>TTC</td>
<td></td>
</tr>
<tr>
<td>TTA</td>
<td>L</td>
</tr>
<tr>
<td>TTG</td>
<td></td>
</tr>
<tr>
<td>TCT</td>
<td>S</td>
</tr>
<tr>
<td>TCC</td>
<td></td>
</tr>
<tr>
<td>TCA</td>
<td></td>
</tr>
<tr>
<td>TCG</td>
<td></td>
</tr>
<tr>
<td>TAT</td>
<td>Y</td>
</tr>
<tr>
<td>TAC</td>
<td></td>
</tr>
<tr>
<td>TAA</td>
<td>stop</td>
</tr>
<tr>
<td>TAG</td>
<td></td>
</tr>
<tr>
<td>TGT</td>
<td>C</td>
</tr>
<tr>
<td>TGC</td>
<td></td>
</tr>
<tr>
<td>TGA</td>
<td>stop</td>
</tr>
<tr>
<td>TGG</td>
<td></td>
</tr>
<tr>
<td>CTT</td>
<td>L</td>
</tr>
<tr>
<td>CTC</td>
<td></td>
</tr>
<tr>
<td>CTA</td>
<td></td>
</tr>
<tr>
<td>CTG</td>
<td></td>
</tr>
<tr>
<td>CCT</td>
<td></td>
</tr>
<tr>
<td>CCC</td>
<td></td>
</tr>
<tr>
<td>CCA</td>
<td></td>
</tr>
<tr>
<td>CCG</td>
<td></td>
</tr>
<tr>
<td>CAT</td>
<td></td>
</tr>
<tr>
<td>CAC</td>
<td></td>
</tr>
<tr>
<td>CAA</td>
<td></td>
</tr>
<tr>
<td>CAG</td>
<td></td>
</tr>
<tr>
<td>CGT</td>
<td></td>
</tr>
<tr>
<td>CGC</td>
<td></td>
</tr>
<tr>
<td>CGA</td>
<td></td>
</tr>
<tr>
<td>CGG</td>
<td></td>
</tr>
<tr>
<td>ATT</td>
<td>I</td>
</tr>
<tr>
<td>ATC</td>
<td></td>
</tr>
<tr>
<td>ATA</td>
<td></td>
</tr>
<tr>
<td>ATG</td>
<td>M start</td>
</tr>
<tr>
<td>ACT</td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>ACA</td>
<td></td>
</tr>
<tr>
<td>ACG</td>
<td></td>
</tr>
<tr>
<td>AAT</td>
<td></td>
</tr>
<tr>
<td>AAC</td>
<td></td>
</tr>
<tr>
<td>AAA</td>
<td></td>
</tr>
<tr>
<td>AAG</td>
<td></td>
</tr>
<tr>
<td>AGT</td>
<td></td>
</tr>
<tr>
<td>AGC</td>
<td></td>
</tr>
<tr>
<td>AGA</td>
<td></td>
</tr>
<tr>
<td>AGG</td>
<td></td>
</tr>
<tr>
<td>GTT</td>
<td>V</td>
</tr>
<tr>
<td>GTC</td>
<td></td>
</tr>
<tr>
<td>GTA</td>
<td></td>
</tr>
<tr>
<td>GTG</td>
<td></td>
</tr>
<tr>
<td>GCT</td>
<td></td>
</tr>
<tr>
<td>GCC</td>
<td></td>
</tr>
<tr>
<td>GCA</td>
<td></td>
</tr>
<tr>
<td>GCG</td>
<td></td>
</tr>
<tr>
<td>GAT</td>
<td></td>
</tr>
<tr>
<td>GAC</td>
<td></td>
</tr>
<tr>
<td>GAA</td>
<td></td>
</tr>
<tr>
<td>GAG</td>
<td></td>
</tr>
<tr>
<td>GGT</td>
<td></td>
</tr>
<tr>
<td>GGC</td>
<td></td>
</tr>
<tr>
<td>GGA</td>
<td></td>
</tr>
<tr>
<td>GGG</td>
<td></td>
</tr>
</tbody>
</table>
Exercise 3: What is RNA?

Learning Objectives:
• Open an RNA sequence and its 3D structure in Jalview
• Colour the nucleotide bases in the sequence
• View the 3D structure

1. Click the link named 'View RNA' in Section (2): Project 1-Exercise 3 on the Schools web page.

2. JalviewJS with the RNA sequence and its 3D structure opens in an adjacent tab.

3. Select the Colour menu in the alignment window. Select the Nucleotide colour scheme.

4. The Jalview nucleotide colour scheme: adenine bases are green, cytosine bases are yellow, guanine bases are red and uracil bases are blue.

Jmol 3D mouse commands:
(i) To rotate the structure place mouse on the structure, then click-and-drag.
(ii) To zoom press the shift key, then click-and-drag.
For more information about Jmol visit: http://wiki.jmol.org/index.php/Main_Page

6. Click in the 3D structure window with the mouse to bring it to the front. Move the mouse across the model to change the view.

This is the SAM responsive riboswitch mRNA (PDB id 2GIS). For more information go to https://en.wikipedia.org/wiki/SAM_riboswitch_(S-box_leader).

Q: What are the names of the four different RNA bases?
Q. How does RNA and its nucleotides differ from those of DNA? (see Appendix Table 1)
Project 2: 'Viewing Myoglobin Proteins'
Exercise 4: Viewing Human Myoglobin Protein

**Background:** Myoglobin is a protein located in muscle; it complexes with iron for oxygen storage. For example, the concentration of myoglobin in muscle cells affects how long an organism can hold its breath. In 1958, whale myoglobin was the first protein ever to have its 3D structure revealed by X-ray crystallography. Max Perutz and John Kendrew won a Nobel Prize in chemistry for this work.

**Learning Objectives:**
- Open the human myoglobin protein sequences in Jalview
- Colour the amino acid residues
- View the 3D structure

1. **Click the link** named 'View human myoglobin protein' in Section (2): Project 2-Exercise 4 on the Schools web page

2. JalviewJS with the myoglobin sequence and its 3D structure opens in an adjacent tab. **Select the Colour menu** in the alignment window. **Select the Taylor colour scheme.**

3. **Select the Colour menu** in the alignment window. **Select the Taylor colour scheme.**

4. In the Taylor colour scheme, each amino acid residue has its own individual colour. (For the key to the 1-letter amino acid codes see Appendix Table 3).

5. **Use the horizontal scroll bar** to scroll to the end of the sequence. **Click on the last residue.** View the information in the status box in the lower left-hand corner of the alignment window.

6. **Click in the 3D structure window** with the mouse to bring it to the front. Move the mouse across the model to change the view of the human myoglobin protein.

Q. How many amino acid residues are there in the human myoglobin protein? (see step 5)
Q. Identify the alpha helix regions in the structure?
Exercise 5: Comparing Myoglobin Protein from Different Species

Learning Objectives:
- Compare myoglobins from different animals
- Produce protein similarity tree

1. Click the link named 'View myoglobin from different species and create a Tree' in Section (2): Project 2-Exercise 5 on the Schools web page.

2. JalviewJS with the protein alignment opens in an adjacent window. Select the Colour menu. Select the Zappo colour scheme.

3. In the Zappo colour scheme, the amino acids are coloured based on their physicochemical properties. (For the key to the 1-letter amino acid codes see Appendix Table 3).

4. Select the Calculate menu. Select Calculate Tree or PCA.

5. Click Calculate in the 'Choose Calculation' box. By default, Neighbour Joining and Blosum62 should be selected, and the Principal Component Analysis option not selected. (Close the 'Choose Calculation' box once the tree has appeared).

6. A tree window opens. Move the tree window to the right placing the mouse on the top title bar then click-and-drag the mouse. The tree reflects the similarity between the human, gorilla, whale, cow, mouse, dolphin, chicken, ostrich, dog, sheep, and elephant myoglobin protein sequences.

7. Click the mouse on the tree, and a red vertical line appears. The red line groups sequences. Each group has its own randomly generated colour.
Exercise 5: Comparing Myoglobin Protein from Different Species

8. Move the mouse to a different location on the tree, this will change the grouping and colours of groups.

Each branch has a number, they can be added together to determine the similarity scores.

9. Select the View menu in the tree window. Select Sort Alignment by Tree.

10. The sequences in the alignment window are reordered to reflect the tree. Note how the sequence names in the alignment window have the same colours as those in the tree window.

Q. Looking at the tree window, are the groupings what you might expect? (Tip: The shorter the length of the branches between species, the more similar they are. The length between each branch points is shown in brackets).

Q: From the tree (step 8), which animal has the most similar myoglobin to humans?

Q: From the tree (step 8), compared to humans which animals have the least similar myoglobin?

Uniprot identifier codes used in this exercise

<table>
<thead>
<tr>
<th>Uniprot ID</th>
<th>Species</th>
<th>Uniprot ID</th>
<th>Species</th>
<th>Uniprot ID</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>P02144</td>
<td>Human</td>
<td>P04247</td>
<td>Mouse</td>
<td>P63113</td>
<td>Dog</td>
</tr>
<tr>
<td>P02147</td>
<td>Gorilla</td>
<td>P68276</td>
<td>Dolphin</td>
<td>P02190</td>
<td>Sheep</td>
</tr>
<tr>
<td>P02185</td>
<td>Whale</td>
<td>P02197</td>
<td>Chicken</td>
<td>P02187</td>
<td>Elephant</td>
</tr>
<tr>
<td>P02192</td>
<td>Cow</td>
<td>P85077</td>
<td>Ostrich</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P02144; P02147; P02185; P02192; P04247; P68276; P02197; P85077; P63113; P02190; P02187</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Project 3: 'Viewing Proteins'
Exercise 6: Viewing Different Classes of Proteins

Background: Proteins are compounds made up of a long chain of amino acid molecules. The amino acids and their order in the protein determine the shape and chemical characteristics of a protein. This in turn, influences the function (role) of a protein.

Biologists use four terms to describe protein structure:-
- **Primary structure** is the sequence or order of the amino acids that making up the protein.
- **Secondary structure** describes the folding pattern of the polypeptide backbone, they are stabilised by hydrogen bonds.
- **Tertiary structure** describes the 3-dimensional shape of the protein. The tertiary structure is stabilised by interactions such as hydrogen bonds, ionic bonds, disulphide bridges, hydrophobic and Van der Waals interactions.
- **Quaternary structure** refers to the structure that forms when several individual proteins link together to form a larger protein complex.

1. **Click the link** named 'View Haemoglobin protein' in Section (2): Project 3-Exercise 6 on the Schools web page

2. In an adjacent tab, the haemoglobin sequence and its 3D structure open in JalviewJS. Haemoglobin is made up of 4 sub-units. The protein strands are displayed as ribbons.

3. **Click in the 3D structure window** with the mouse, then **drag the mouse** to change the view.

4. **Zoom to view one of the subunits** in the 3D molecule. Observe the heme group, that is displayed as ball-and-stick model. The heme consists of an iron atom (orange) within a heterocyclic ring.

- **β-Sheet (3 strands)**
- **α-helix**

---

16
Exercise 6: Viewing Different Classes of Proteins

Learning Objectives:
- Open the protein sequences in Jalview
- Colour the amino acid residues
- View the 3D structure

1. Click the link named 'View collagen protein' in Section (2): Project 3-Exercise 6 on the Schools web page.

2. In an adjacent window, the collagen protein sequence and its 3D structure opens in JalviewJS.

3. Select the Colour menu. Select a colour scheme eg Zappo.

4. In the Jmol structure window, Select the Colour menu. Select By Chain to view each of the chains in the 3D structure.

5. Click in the 3D structure window with the mouse. Drag the mouse to change the view.

Repeat this process for the other proteins listed on the web page: Note: Some are quite large file so may take a little while to open.

Q. How many different protein amino acids are found in nature? (see Appendix Table 3)
Q. What is the name of the amino acid that is represented by the letter A?
Q. What is the chemical formula of the amino acid glycine (G)?

<table>
<thead>
<tr>
<th>Protein type</th>
<th>Function</th>
<th>Example</th>
<th>PDB ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme</td>
<td>Catalysis of chemical reactions</td>
<td>Amylase [<a href="http://www.rcsb.org/pdb/101/motm.do?momID=74">www.rcsb.org/pdb/101/motm.do?momID=74</a>]</td>
<td>1SMD</td>
</tr>
<tr>
<td>Structure</td>
<td>Provides mechanical support to cells and tissues</td>
<td>Collagen [<a href="http://www.rcsb.org/pdb/101/motm.do?momID=4">www.rcsb.org/pdb/101/motm.do?momID=4</a>]</td>
<td>1CAG &amp; 1K6F</td>
</tr>
<tr>
<td>Storage</td>
<td>Stores small molecules or ions</td>
<td>Ferritin [<a href="http://www.rcsb.org/pdb/101/motm.do?momID=35">www.rcsb.org/pdb/101/motm.do?momID=35</a>]</td>
<td>5XB1 &amp; 1MFR</td>
</tr>
<tr>
<td>Signalling</td>
<td>Regulates body metabolism and the nervous system</td>
<td>Insulin [<a href="http://www.rcsb.org/pdb/101/motm.do?momID=14">www.rcsb.org/pdb/101/motm.do?momID=14</a>]</td>
<td>1TRZ</td>
</tr>
<tr>
<td>Transport</td>
<td>Carry substances around the body</td>
<td>Myoglobin [<a href="http://www.rcsb.org/pdb/101/motm.do?momID=1">www.rcsb.org/pdb/101/motm.do?momID=1</a>]</td>
<td>3RGK &amp; 1MBO</td>
</tr>
</tbody>
</table>
Exercise 6: Additional Information

<table>
<thead>
<tr>
<th>Protein</th>
<th>Link to view model in SketchFab</th>
<th>3D Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amylase</td>
<td><a href="https://skfb.ly/6ZDBE">https://skfb.ly/6ZDBE</a></td>
<td></td>
</tr>
<tr>
<td>Collagen</td>
<td><a href="https://skfb.ly/6ZDDR">https://skfb.ly/6ZDDR</a></td>
<td></td>
</tr>
<tr>
<td>Antibody</td>
<td><a href="https://skfb.ly/6ZDDZ">https://skfb.ly/6ZDDZ</a></td>
<td></td>
</tr>
<tr>
<td>Ferritin</td>
<td><a href="https://skfb.ly/6ZDE6">https://skfb.ly/6ZDE6</a></td>
<td></td>
</tr>
<tr>
<td>Insulin</td>
<td><a href="https://skfb.ly/6ZDEo">https://skfb.ly/6ZDEo</a></td>
<td></td>
</tr>
<tr>
<td>Myoglobin</td>
<td><a href="https://skfb.ly/6ZDEp">https://skfb.ly/6ZDEp</a></td>
<td></td>
</tr>
</tbody>
</table>

Alternative method: Open files directly from biological databases

Open sequence files by reading them from public databases
1. Go to the File menu in the desktop window. Select Fetch Sequences.

2. In the 'New Sequence Fetcher' box, select PDB from the 'Select Database' list. Select/Click OK.

3. Enter the PDB ID code eg 1IGT in the 'PDB Sequences Fetcher' box. Press [return]. Select the PDB ID code from the list eg 1smd. Click OK. (Close the 'PDB Sequence Fetcher' box once the sequence has loaded).

4. Select the View menu in the alignment window. Uncheck Show Sequence Features. (Note: This needs to be toggled off otherwise it can mask the residue colour schemes).
Alternative method: Open files directly from biological databases

5. Click and drag the mouse to select all the names of the sequences in the alignment. 
Right click the mouse to open the pop-up menu. 
Select 3D Structure Data.

Troubleshooting Note: to open the context menu, place the mouse cursor over the sequence names and right click, not the sequence itself.

6. In the ‘Structure Chooser’ box, select the PDB Id entries. 
Click New View. 
(Click and drag the mouse or use the Shift key if the IDs are not selected).

7. A 3D structure window opens. 
In the Jmol structure window,

8. Select the Colour menu in Jmol window. 
Select By Chain to view each of the chains in the 3D structure.

9. Click in the 3D structure window with the mouse. 
Drag the mouse to change the view.
Project 4: 'Viewing the Genetic Mutation involved for Sickle Cell Anaemia'
**What are Chromosomes & Genes?**

A chromosome is made up of DNA tightly coiled around proteins called histones. Each human cell normally contains 23 pairs of chromosomes, i.e., a total of 46 chromosomes. This is an image of a set of 46 stained metaphase chromosomes from a male human.

Chromosomes are often represented in an idealized arrangement with their centromeres aligned. In the image both the male (XY) and female (XX) versions of the 23rd chromosome are shown.

This is a pictorial representation of human chromosome 11. Each chromosome contains several genes. Genes are regions of DNA that produce specific proteins during translation.

The location of the haemoglobin-beta (HBB) gene, the gene involved in sickle cell disease, is marked with an arrow.

Genes are made up of protein coding regions called exons, separated by regions called introns. Introns do not encode proteins but appear to play a role in regulation and gene expression.

The figure showing the relationship between the DNA in chromosomes, genes, exons, and introns.
Exercise 7: What are Exons & Introns?

Background: Sickle cell anaemia is a genetic disease where the body produces crescent-shaped red blood cells. A single nucleotide mutation in the HBB gene causes the disease. The HBB gene is located on chromosome 11 and is involved in the production of beta-globin protein. Two beta- and two alpha-globin proteins combine to produce haemoglobin present in red blood. The mutation changes the amino acid at residue 7 of the beta-globin protein. The disease is autosomal recessive which means that only people with 2 copies of the sickle cell gene have the condition. A study in 2015 suggests that about 4.4 million people have this disease.

Learning Objectives:
- Identify the exon and intron regions in the HBB gene

1. Click the link named 'View Exons & Introns' in Section (2): Project 4-Exercise 7 on the Schools web page.

2. JalviewJS with the HBB gene opens in an adjacent window. The pink DNA regions are introns. The green DNA regions are exons.

3. Move the overview window away from the alignment. Increase the window's size. The red box shows the part of the sequence visible in the alignment window. Use the mouse to drag the red box to the right.

Note: The bases in the haemoglobin beta (HBB) gene are labelled with the 1-letter identifier: G (guanine), T (thymine), C (cytosine) and A (adenine).

Note: The Overview window can be opened from the View menu in the alignment window. Select the Overview Window.

Q. How many exons are in the HBB gene?
Q. What is the main difference between exons and introns?
Exercise 8: Viewing coding DNA sequence & its protein product

**Learning Objectives:**
- View the coding DNA sequence of the *HBB* gene alongside its protein product
- Identify the location of the mutation involved in sickle cell anaemia

1. **Click the link** named 'View the coding DNA & its protein product' in Section (2): Project 4-Exercise 8 on the Schools web page.

   Note: It may take a little while for Jalview to open the files depending on the speed of the internet.

2. A split-frame window opens that contains the *HBB* coding DNA sequence in the top panel and *HBB* protein in the lower panel. The panels are inter-linked.

3. **Place the mouse over an amino acid** in the lower panel, **View the associated nucleotide triplet or codon** in the upper panel.

4. **Use the scroll bar to move the sequences** to the start of the sequence. The sickle cell mutation is highlighted in red on both the DNA and protein sequences.

5. In the upper coding panel, place the mouse over the **adenine (A) at base number 20** (highlighted in red).
   
   A tooltip opens, note its information.

   In the lower protein panel, repeat with the **glutamic acid amino acid (E) at residue number 7** (highlighted in red).
   
   A second tooltip opens, note its information.

   If the tooltip doesn’t open, right click the mouse and select **Feature Details** instead.
   
   Additional information is available in **Status Bar** in lower left-hand corner.

**Questions:**

Q. What nucleotide triplet codon code produces the glutamic acid at residue 7?
Q. From the information in the tooltip, what is the change in the nucleotide base at base number 20 that is responsible for sickle cell anaemia?
Q. In the mutated HBB protein, what amino acid replaces the glutamic acid at residue 7?
The location of the mutation (highlighted in red) on the haemoglobin model can be viewed in SketchFab. Click here to view: [https://skfb.ly/6ZDNG](https://skfb.ly/6ZDNG).

Q. How does a single mutation at base 20 in the HBB gene result in a person getting sickle cell disease?

Q. Do you know the evolutionary advantage that heterozygote sickle cell mutation infers (only 1 of the HHB genes has the mutation), that has resulted in an increased prevalence of sickle cell anaemia in Africa? (see [https://en.wikipedia.org/wiki/Heterozygote_advantage](https://en.wikipedia.org/wiki/Heterozygote_advantage)).

To Review:
- The DNA can be sub-divided into genes.
- Genes code for specific proteins.
- Alongside sequence of nucleotides that code for a protein, there are nucleotide sequences that are non-coding.
- The coding regions of a gene are called exons and the non-coding regions are called introns.
- Intron are transcribed but then removed during the production of messenger RNA (mRNA).

Exercise 8: Advanced version

For students who want to experiment for themselves, this version illustrates all the steps involved. They include reading files from public biological database, launching the split-frame view, loading features, and opening in 3D structures from the PDB. Once a student has mastered these steps then they can view other genes and proteins for themselves.

1. Click the link named 'ADVANCED VERSION....View the coding DNA & its protein product' in Section (2): Project 4-Exercise 8 on the Schools web page.

2. In the empty JalviewJS desktop window, select the File menu. Select Fetch Sequences.

3. Select EMBLCDS from the 'Select Database' list in the 'New Sequence Fetcher' box.
4. Enter the ID **CAG46711** in the New Sequence Fetcher box. Click **OK** to retrieve the coding DNA sequence. (Close the New Sequence Fetcher box once the sequence has loaded).

5. Select the **Calculate** menu in the alignment window. Select **Get Cross References**. Select **Uniprot**. (Note: You may have to wait a little whilst Jalview fetches the data).

6. A split-screen window opens that contains the **HBB coding DNA sequence** (top panel) and **HBB protein** (lower panel). The panels are inter-linked. Place the mouse over an amino acid in the protein and view the associated codon in the DNA sequence.

7. There are two links in the top left hand corner of the JalviewJS web page. Click-and-drag the top link Drag 'this link on DNA sequence to add features' onto the **HBB gene sequence panel** in the upper DNA alignment window. This opens the features file and colours the DNA sequence.

8. Click-and-drag the second link Drag 'this link on protein sequence to add features' onto the **protein sequence panel** in the lower protein alignment window. This opens the features file and colours the protein sequence.

9. The sickle cell mutation is highlighted in red on both the DNA and protein sequences.

10. In the upper DNA panel, select the **View** menu. Select the **Overview Window** to view all 3 exons.
Exercise 8: Advance version

11. In the upper panel, place the mouse over the adenine (A) at base number 20 in the DNA sequence. A tooltip opens, note its information. In the lower protein panel, repeat with the glutamic acid amino acid (E) at residue number 7. Another tooltip opens, note its information. Additional information is available in Status Bar in lower left-hand corner.

12. Click the mouse cursor on the protein sequence name. Right click the mouse to open the pop-up menu. Select 3D Structure Data.

13. In the 'Structure Chooser' box, select PDB id 3nmm entry. Select New View.

14. A 3D structure window opens containing the HBB protein.

15. In the protein sequence panel, select the View menu in the alignment window. Select Features Settings... .

16. In the 'Sequence Feature Settings' box, click the mouse on the red Sickle_cell_variant feature name and drag it to the top of the list, above the green RESNUM feature name. Click OK.

17. Rotate the 3D structure to locate residue 7 (coloured red) that is mutated in sickle cell anaemia. Mouse over the protein residue 7 in the protein sequence (lower panel of the split-screen viewer). See the effect in the 3D viewer.

Note: If you can't see the red coloured residue return to step 16 and change the order of features in the list.
Appendix

Table 1: Composition of DNA & RNA subunits
DNA and RNA are polymers made up of nucleotide sub-units. The nucleotide consists of a phosphate group, a 5-ring sugar, and a nitrogenous base. DNA contains a deoxyribose sugar and has a thymine base. RNAs contains a ribose sugar and has an uracil base. RNAs are usually single-stranded.

Table 2: Codon Table
A codon is a set of three nucleotides, or triplet, that code for a specific amino acid residue during protein synthesis.

Amino Acids
Amino acids are the building blocks (sub-units) of proteins. Attached to the central carbon is a hydrogen, a NH₂ amino group (this can have a positive charge depending on pH), a COOH carboxylic group (this can have a negative charge depending on pH) and variable side group R. The R sidechain influences whether an amino acid is polar or non-polar, acidic, or basic.
Amino acids link together by forming peptide bonds between the COOH and NH₂ of neighbouring amino acids during a condensation reaction that releases water. Multiple amino acids join together to form a polypeptide. A protein is produced when the amino acid strand contains more than 50 residues.

![Peptide bond]

Table 3: Amino acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3 letter code</th>
<th>1 letter code</th>
<th>Sidechain polarity</th>
<th>Linear</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>nonpolar</td>
<td>(\text{CH}_3\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}_3\text{H}_7\text{NO}_2)</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>basic polar</td>
<td>(\text{HN=CH(NH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>6\text{H}</em>{14}\text{N}_4\text{O}_2)</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
<td>polar</td>
<td>(\text{H}_2\text{N-CO-CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}_6\text{H}_7\text{O}_3)</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
<td>acidic polar</td>
<td>(\text{HOOC-CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}_4\text{H}_7\text{NO}_3)</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td>nonpolar</td>
<td>(\text{HS-CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}_3\text{H}_7\text{NO}_2)S</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
<td>acidic polar</td>
<td>(\text{HOOC-(CH}_2\text{)_2-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>5\text{H}</em>{10}\text{N}_2\text{O}_4)</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
<td>polar</td>
<td>(\text{H}_2\text{N-CO-(CH}_2\text{)_2-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>5\text{H}</em>{10}\text{N}_2\text{O}_3)</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
<td>nonpolar</td>
<td>(\text{H}-\text{CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}_2\text{H}_4\text{N}_2\text{O})</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td>basic polar</td>
<td>(\text{NH}-\text{CH=NH}-\text{CH}-(\text{CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>6\text{H}</em>{14}\text{O}_2)</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td>nonpolar</td>
<td>(\text{CH}_3\text{-CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>6\text{H}</em>{13}\text{O}_2)</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
<td>nonpolar</td>
<td>(\text{(CH}_2\text{)_2-CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>5\text{H}</em>{13}\text{N}_2\text{O}_2)</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
<td>basic polar</td>
<td>(\text{H}_2\text{N-(CH}_2\text{)_2-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>6\text{H}</em>{14}\text{N}_2\text{O}_2)</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
<td>nonpolar</td>
<td>(\text{CH}_3\text{S-(CH}_2\text{)_2-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>5\text{H}</em>{14}\text{NO}_2)S</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
<td>nonpolar</td>
<td>(\text{Ph-CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>9\text{H}</em>{11}\text{NO}_2)</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
<td>nonpolar</td>
<td>(\text{-NH-(CH}_3\text{)_2-CH-COOH)})</td>
<td>(\text{C}<em>6\text{H}</em>{12}\text{O}_2)</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
<td>polar</td>
<td>(\text{HO-CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}_3\text{H}_7\text{NO}_3)</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
<td>polar</td>
<td>(\text{CH}_2\text{-CH(OH)-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}_4\text{H}_9\text{NO}_3)</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
<td>nonpolar</td>
<td>(\text{Ph-NH-CH=CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>{11}\text{H}</em>{12}\text{N}_2\text{O}_2)</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
<td>polar</td>
<td>(\text{HO-Ph-CH}_2\text{-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>9\text{H}</em>{11}\text{NO}_3)</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
<td>nonpolar</td>
<td>(\text{(CH}_2\text{)_2-CH-CH(NH}_2\text{-COOH)})</td>
<td>(\text{C}<em>6\text{H}</em>{11}\text{NO}_2)</td>
</tr>
</tbody>
</table>

Additional Help:
Help can be accessed from the Help menu in the Jalview’s desktop window, and select Documentation.
If you are interested in learning more, the Jalview manual contains several hands-on exercises, available at [www.jalview.org/about/documentation](http://www.jalview.org/about/documentation).
Or visit the Jalview YouTube training channel at [https://www.youtube.com/channel/UCJpnyZB770yz7ftbrJ0tfw](https://www.youtube.com/channel/UCJpnyZB770yz7ftbrJ0tfw) and view the videos and playlists.

If you enjoyed the exercises in this workbook, we would appreciate it if you would post us a message on Twitter or Facebook.
Please include @Jalview, #STEM, #genetics
Glossary

Amino acid: molecular sub-units of peptides and proteins.

Bioinformatics: the application of computer and statistical techniques to the management of biological data.

cDNA (complementary DNA): cDNA sequence is synthesized from an RNA template by reverse transcription. It contains 5' and 3' untranslated regions (UTRs) as well as coding regions.

CDS (protein-coding sequence): the portion of the mRNA transcript that is translated by ribosomes into proteins.

Chromosome: located in the cell nucleus, it contains the cellular DNA along with a number of proteins (eg histones) that compact and package the DNA.

Codon: a set of three adjacent nucleotides (triplet) that code for a specific amino acid residue during protein synthesis.

DNA (deoxyribonucleic acid): the molecule that encodes genetic information. It carries the instructions for all aspects of an organism's functions such as growth, metabolism and reproduction. These chains can be over 100,000,000 molecules in length.

Exon: the sections of a gene that are translated into proteins, they remain in the transcript (mRNA) after introns have been spliced out of the genomic sequence.

Gene: a region of DNA that encodes a specific protein or protein subunit.

Genetic code: sets of triplet nucleotides that encodes specific amino acids.

Genome: all the genetic material in the chromosomes of a particular organism.

Genomic DNA (gDNA): all the DNA residing in the chromosomes.

Genotype: all the genes in a particular individual.

Intron: the noncoding part of the genome that is transcribed then spliced out of the RNA.

Phenotype: the observable characteristics or features of a living organism.

Phylogenetic tree: an evolutionary tree for organismal species or cellular macromolecules that is built using inheritance or molecular sequence information.

Protein: a biological macro-molecule composed of a string of amino acids joined together by peptide bonds.

Protein sequence: the sequence of amino acids in a protein.

Nucleotide: building blocks of RNA and DNA made up of a nitrogenous base, a molecule of sugar and phosphoric acid.

Multiple sequence alignment: an alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column.

RNA (ribonucleic acid): RNA are similar to DNA but containing the ribose sugar rather than deoxyribose sugar and the base uracil (U) rather than thymine (T). Typically they are single-stranded.

Replication: process by which DNA makes a copy of itself during cell division.

Sequence alignment: arranging the sequences of protein, RNA or DNA to identify regions of similarity. The similarity could be a consequence of functional, structural, or evolutionary relationships.

Translation: process where mRNA is decoded by ribosomes to produce specific amino acids and polypeptides.

Transcription: process where a segment of DNA is copied into RNA by the enzyme RNA polymerase.

Free Public Biological Databases:

• UniProt is a database of protein sequences (http://www.uniprot.org/).

• Protein Data Bank (PDB) is a database of crystallographic, three-dimensional structural data of large biological molecules (http://www.rcsb.org/).

• Ensembl is a genomic database (http://ensemblgenomes.org/).

• EMBL (CDS) data originates from the European Nucleotide Archive (ENA) database of annotated DNA and RNA sequences (https://www.ebi.ac.uk/ena).