Jalview

Visualising DNA, RNA & Proteins in Jalview

School Workbook

University of Dundee

About this workbook:

The workbook contains 4 easy-to-run web-based bioinformatics projects:
Project 1 views DNA and RNA sequences and their 3D structures.
Project 2 views a human myoglobin protein sequence and its 3D structure, then use a similarity tree to compare the myoglobin sequences from different animals.
Project 3 views the sequences and 3D structures of proteins with a range of different biological functions.
Project 4 views the exons and introns in the *HBB* gene, and identifies the genetic mutation on the DNA that is linked to sickle cell anaemia, then views its location on the 3D structure of the haemoglobin protein.

The workbook, the 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:
https://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 computer with a web browser and internet access. As exercises run in a web browser, no software has to be downloaded.

What is Jalview?

Jalview is free-to-use computer software developed at the University of Dundee. It is designed to allow 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

Getting Started

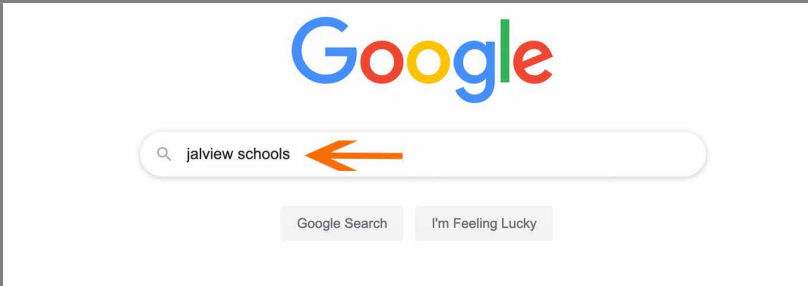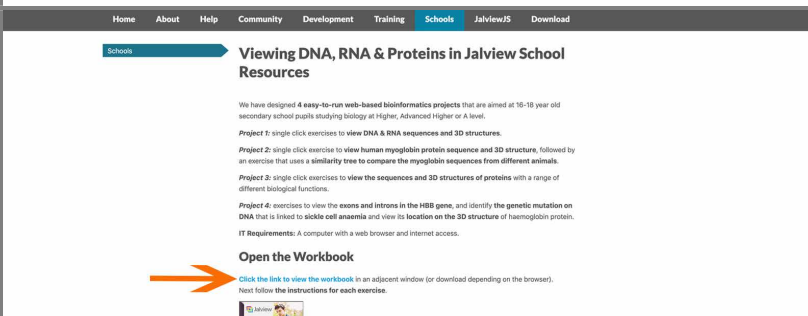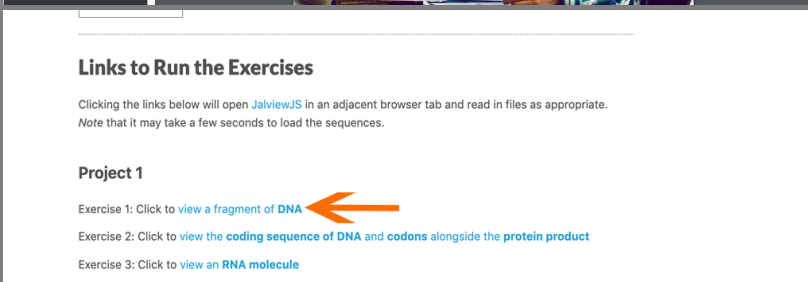|	
1. **Open a web browser** such as Chrome or Firefox. We suggest you avoid Internet Explorer as JalviewJS does not always work in older browsers.	
2. **Search** using keywords '***Jalview Schools***'. From the list of results, **select** '**Resources for Schools**' at https://www.jalview.org/school-resources. This will open the Jalview Schools web page.	
3. **Select the 'Click the link to view the Workbook**' link 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 'view a fragment of DNA**' in Project 1-Exercise 1.	
6. JalviewJS viewer **opens in an adjacent window** in the web browser.	

Change the appearance of the windows in JalviewJS:
(i) **To move a window**, place the mouse on the 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 the Jalview's Windows

Project 1:
'Viewing DNA & RNA'

What is DNA?

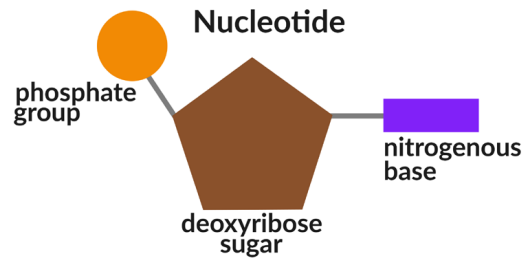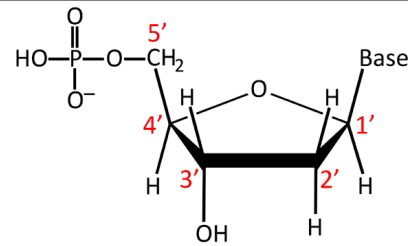|	
DNA (deoxyribonucleic acid) is made up of **two strands** in a **double helix**. The two DNA strands are **anti-parallel** with respect to each other.	strand of DNA
A DNA strand is a polymer. Its sub-units are called **nucleotides**. A nucleotide consists of a **phosphate group**, a **5-ring sugar,** and a **nitrogenous base** (for more information see Table 1 in the Appendix).	Nucleotide — phosphate group, deoxyribose sugar, nitrogenous base
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)**.	Sugar phosphate backbone. Nitrogenous base: Guanine (G), Thymine (T), Cytosine (C), Adenine (A)
The **nucleotide bases** between the two intertwining strands of DNA form weak **hydrogen bonds**. The **adenine** aligns with the **thymine**, forming 2 hydrogen bonds. The **cytosine** aligns with the **guanine**, forming 3 hydrogen bonds. These pairings are called **complementary base pairings**.	Complementary base pairings — hydrogen bonds
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.	C5' carbon (purple), 5' (prime) end, Phosphorus atoms (orange), Sugar phosphate backbone, Bases on sidechains, 3' (prime) end, C3' carbon (cyan)
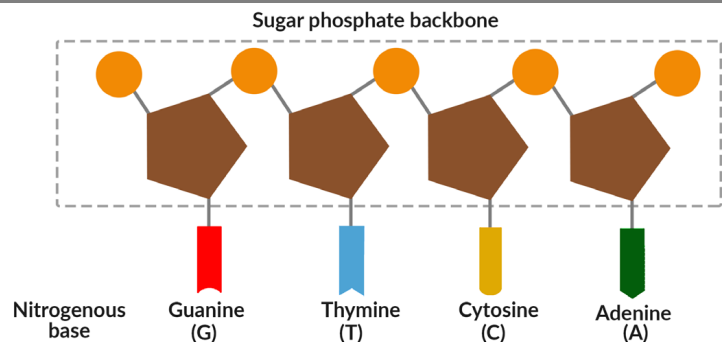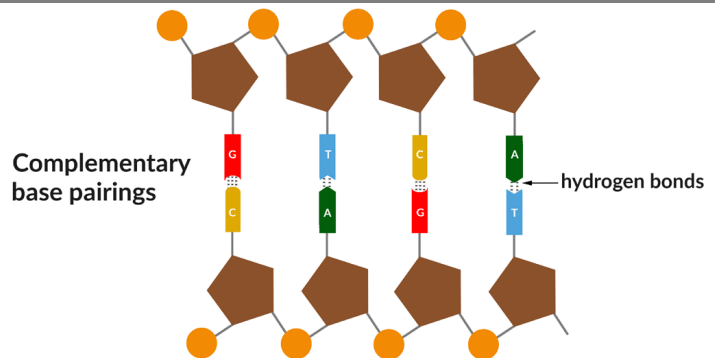
4
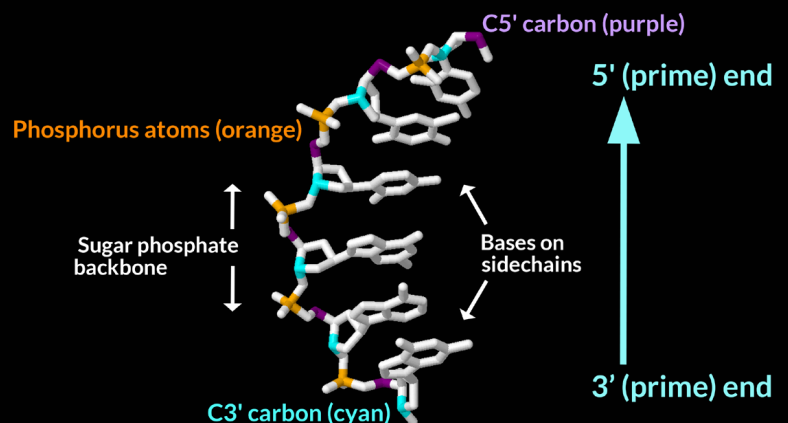
Exercise 1: What is DNA?

> **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** 'view a fragment of DNA' in Project 1-Exercise 1 on the Schools web page.	
2. JalviewJS with the DNA sequences and its 3D structure opens in an adjacent tab of the web browser. 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**.	
6. *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** the mouse. For more information about Jmol visit: https://wiki.jmol.org/index.php/Main_Page	
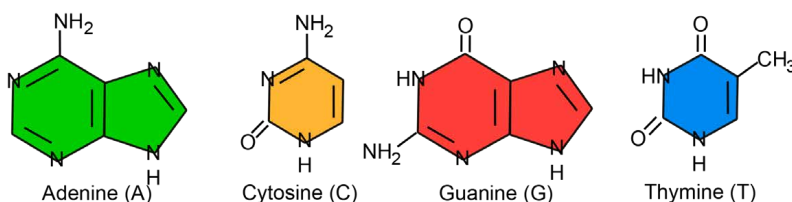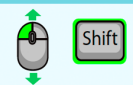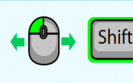
Q: What are the names of the 4 different DNA bases, and their single letters identifier?
Q. How many nucleotide base pairs are there in this fragment of DNA? (Tip: count the base pairs in the 3D structure window)

Exercise 1: What is DNA?

7. **Select the Colours** menu in the Jmol structure window. **Select the Purine/Pyrimidine** colour scheme.	
8. The **purines** (adenine and guanine bases) are coloured pink. The **pyrimidines** (thymine and cytosine bases) are coloured cyan.	
9. 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. In the Jmol structure window, note the location of the first base in each sequence.	
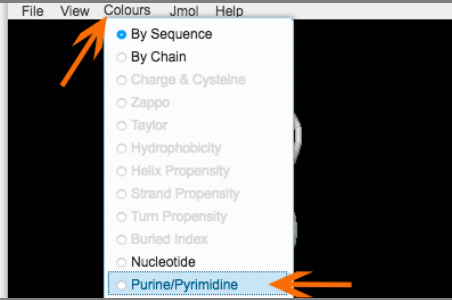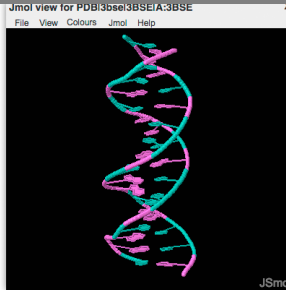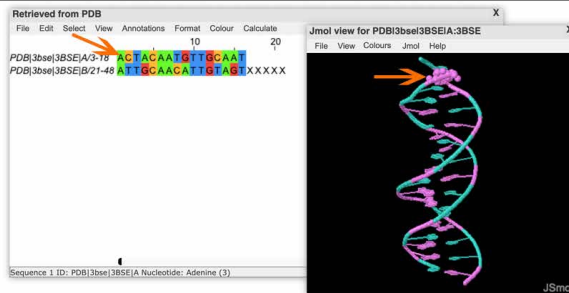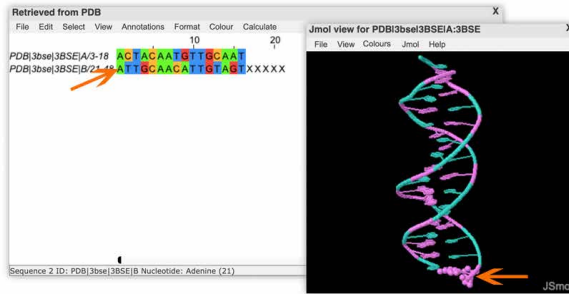
Q. From step 9, are the DNA strands parallel or anti-parallel?
Q. Look at the shape of the purines and pyrimidines in the structure viewer, how are they different to each other? (see Table 1 in the Appendix)

10. In the Jmol structure window, **select the View** menu. **Select Show Chain**. **Uncheck 3BSE:B**.	
11. **Drag the mouse to change the view** to 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. What is the direction of the spiral turning?	
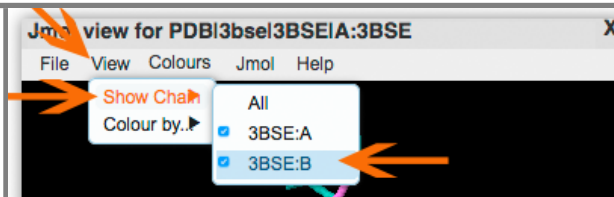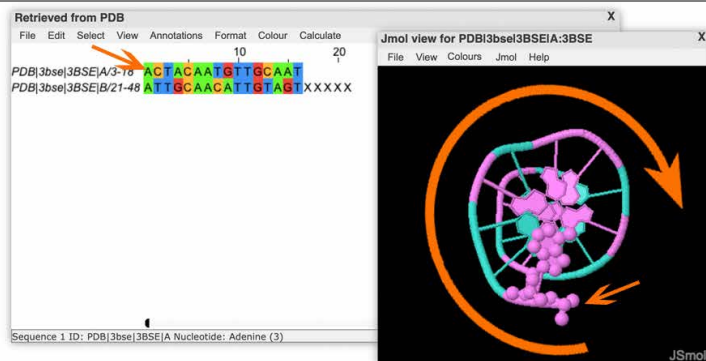
Q. In Step 11, what direction is the strand turning, clockwise or anticlockwise?

Exercise 1: What is DNA?

The 3D structure of a molecule can be displayed in several different ways.
View 1: DNA is 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.
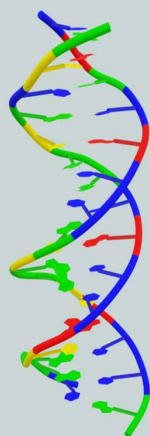
View 1 **Ribbon cartoon model** coloured with the Jalview nucleotide colour scheme.	*View 2* **Ball-and-stick model** atoms are displayed as balls coloured with the CPK colour scheme.	*View 3* **Space-fill model** displays atoms as spheres with atom radii equal to their Van der Waals radius.
Click link 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*
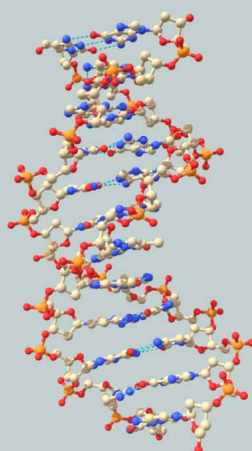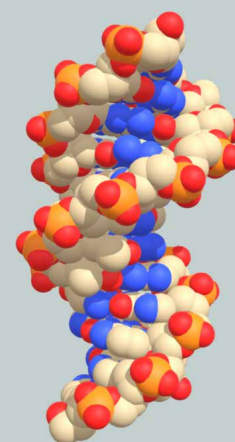
For more information about the DNA fragment 3BSE visit the Protein Data Bank website at www.rcsb.org/structure/3BSE

Exercise 2: What are codons?

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

1. **Click the link** 'view the coding sequence of DNA & codons alongside the protein product' in Project 1-Exercise 2 on the Schools web page.	Visualising DNA, RNA & Proteins in Jalview School Workbook University of Dundee **Links to Run the Exercises** Clicking the links below will open JalviewJS in an adjacent browser tab and read in files as appropriate. *Note* that it may take a few seconds to load the sequences. **Project 1** Exercise 1: Click to view a fragment of **DNA** Exercise 2: Click to view the **coding sequence of DNA** and **codons** alongside the **protein product** Exercise 3: Click to view an **RNA molecule**
Note: It may take a little while for Jalview to open the files depending on the speed of the internet.	
2. In an adjacent browser tab, a split-frame window opens containing the **coding DNA sequence (in the upper panel)** and the **myoglobin protein sequence (in the lower panel)**.	Linked CDS and protein view File Edit Select View Annotations Format Colour Calculate CDS|Coding_Seq|CAG46747 **Coding DNA sequence** File Edit Select View Annotations Format Colour Calculate Myoglobin_Protein|UNIPROT|P02144 MGL SDGEWQL VLNVWGKVEAD I PGHGQEVL I RL FKGHPETL EKFDKFKHL KSED **Protein sequence** Sequence 1 ID: UNIPROT|P02144 Residue: MET (1)
3. The DNA and protein sequence panels are linked. **Place the mouse over an amino acid residue in the lower pane**l 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 to the left and to the right.	Linked CDS and protein view File Edit Select View Annotations Format Colour Calculate CDS|Coding_Seq|CAG46747 atggggctcagcgacggggaatggcagttggtgctgaac ← **Scroll bar** Sequence 1 ID: CDS|Coding_Seq|CAG46747 Nucleotide: Guanine (11) File Edit Select View Annotations Format Colour Calculate Myoglobin_Protein|UNIPROT|P02144 MGLSDGEWQL VLNVWGKVEAD I PGHGQEVL I RL FKGHPE ← **Scroll bar** Sequence 1 ID: UNIPROT|P02144 Residue: SER (4)
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: The status box provides additional information.	Linked CDS and protein view File Edit Select View Annotations Format Colour Calculate CDS|Coding_Seq|CAG46747 gggctcagcgacgggaatggcagttggtgctgaacg Sequence 1 ID: EMBLCDS|CAG46747 Nucleotide: Cytosine (7) ← **Status box** File Edit Select View Annotations Format Colour Calculate Myoglobin_Protein|UNIPROT|P02144 MGLSDGEWQL VLNVWGKVEAD I PGHGQEVL I RL FKGH Sequence 1 ID: Myoglobin_Protein|UNIPROT|P02144 Residue: LEU (3) ← **Status box**

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 is the triplet bases (codon) codes for the selected amino acid residue?

Linked CDS and protein view
File   Edit   Select   View   Annotations   Format   Colour   Calculate

CDS|Coding_Seq|CAG46747

Sequence 1 ID: EMBLCDS|CAG46747 Nucleotide: Guanine (178)
File   Edit   Select   View   Annotations   Format   Colour   Calculate

Myoglobin_Protein|UNIPROT|P02144   KVEADIPGHGQEVL IRLFKGHPETLEKFDKFKHLKSEDEMKAS EDLKKHGATVL

Jmol view for Myoglobin_Protein|UNIPROT|P02144:3rgk   X
File   View   Colours   Jmol   Help

Sequence 1 ID: Myoglobin_Protein|UNIPROT|P02144 Residue: GLU (60)

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

TTT TTC } F	CTT CTC CTA CTG } L	ATT ATC ATA } I	GTT GTC GTA GTG } V
TTA TTG } L		ATG — M start	
TCT TCC TCA TCG } S	CCT CCC CCA CCG } P	ACT ACC ACA ACG } T	GCT GCC GCA GCG } A
TAT TAC } Y	CAT CAC } H	AAT AAC } N	GAT GAC } D
TAA TAG } stop	CAA CAG } Q	AAA AAG } K	GAA GAG } E
TGT TGC } C	CGT CGC CGA CGG } R	AGT AGC } S	GGT GGC GGA GGG } G
TGA } stop TGG } W		AGA AGG } R	

9

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 its 3D structure

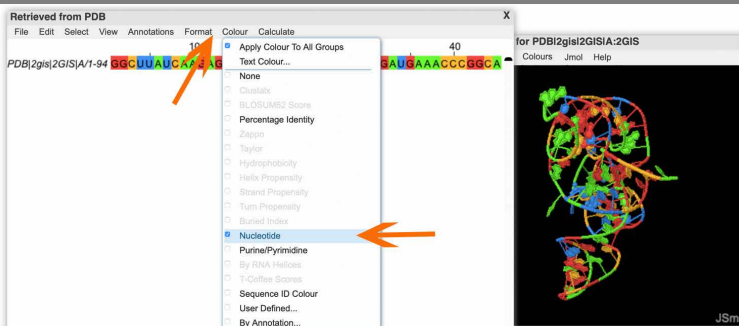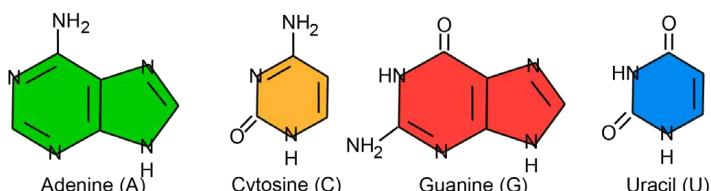|	
1. **Click the link** 'view an **RNA molecule**' in Project 1-Exercise 3 on the Schools web page.	**Links to Run the Exercises** Clicking the links below will open JalviewJS in an adjacent browser tab and read in files as appropriate. *Note* that it may take a few seconds to load the sequences. **Project 1** Exercise 1: Click to view a fragment of **DNA** Exercise 2: Click to view the **coding sequence of DNA** and **codons** alongside the **protein product** Exercise 3: Click to view an **RNA molecule**
2. JalviewJS with the RNA sequence and its 3D structure opens in an adjacent web browser 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.	Adenine (A) Cytosine (C) Guanine (G) Uracil (U)
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** the mouse. For more information about Jmol visit: https://wiki.jmol.org/index.php/Main_Page	**Mouse Movements -- for Jmol** Rotate on the X-Y axes: Zoom in and out: Shift Translate the Molecule: Ctrl Alt Rotate on the Z axis: Shift
5. **Click in the 3D structure window** with the mouse to bring it to the front. **Move the mouse** across the window 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).	
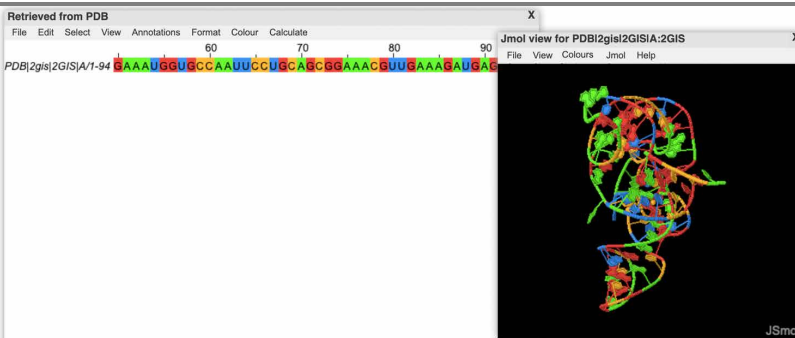
Q: What are the names of the four different RNA bases?
Q. How does RNA and its nucleotides differ from those of DNA? (see Table 1 in the Appendix)

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 animal 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 sequence in Jalview
- Colour the amino acid residues
- View its 3D structure

1. **Click the link** 'view Human Myoglobin protein and 3D structure' in Project 2-Exercise 4 on the Schools web page.	Exercise 3: Click to view an **RNA molecule** **Project 2** Exercise 4: Click to view **Human Myoglobin protein sequence and 3D structure** ← Exercise 5: Click to view **Myoglobin protein sequences** from different species and create a **Tree** **Project 3**
2. JalviewJS with the myoglobin sequence and its 3D structure opens in an adjacent tab of the web browser.	
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 the mouse on the last residue** in the sequence. **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** within the 3D structure window 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. How many alpha helix regions are there in the structure? (see page 17 for more information)

Exercise 5: Compare Myoglobin Proteins from Different Species

> **Learning Objectives:**
> • Compare myoglobin sequences from different animals
> • Produce a protein sequence similarity tree

1. **Click the link** 'view Myoglobin protein sequences from different species and create a Tree' in Project 2-Exercise 5 on the Schools web page.	Exercise 3: Click to view an **RNA molecule** **Project 2** Exercise 4: Click to view **Human Myoglobin protein sequence and 3D structure** Exercise 5: Click to view **Myoglobin protein sequences** from different species and create a **Tree** **Project 3** Exercise 6: View protein sequences and 3D structures
2. JalviewJS with the protein alignment opens in an adjacent window. **Select the Colour** menu in the alignment window. **Select the Zappo** colour scheme.	http://www.jalview.org/tutorial/myoglobins.mfa File Edit Select View Annotations Format Colour Calculate Apply Colour To All Groups / Text Colour... / None / Clustalx / BLOSUM62 Score / Percentage Identity / Zappo / Taylor / Hydrophobicity

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 Table 3 in the Appendix).	Aliphatic/hydrophobic — ILVAM
	Aromatic — FWY
	Positive — KRH
	Negative — DE
	Hydrophilic — STNQ
	conformationally special — PG
	Cysteine — C

> Using the horizontal scroll bar to view the amino acid residues in the different myoglobin sequences in the alignment window.
> Q. Which animal has the most similar sequence compared to the human myoglobin?

4. **Select the Calculate** menu in the alignment window. **Select Calculate Tree or PCA**.	File Edit Select View Annotations Format Colour Calculate Sort ▶ Calculate Tree or PCA... Pairwise Alignment AutoCalculate Consensus Sort Alignment With New Tree Show flanking regions Extract Scores
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).	**Choose Calculation** X ○ Principal Component Analysis Tree ● Neighbour Joining ○ Average Distance BLOSUM62 Calculate Close
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.	

Exercise 5: Compare Myoglobin Proteins from Different Species

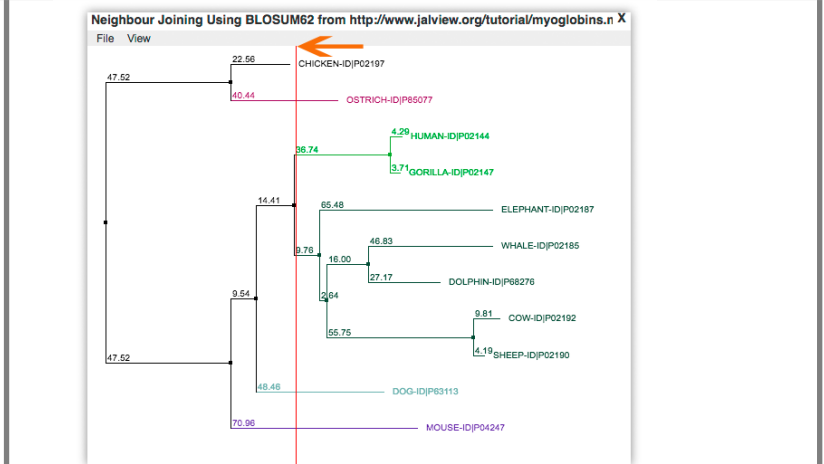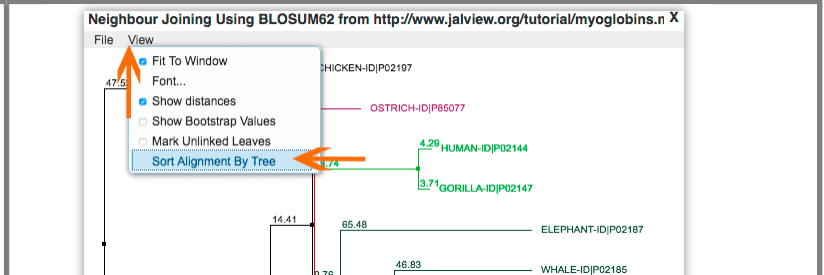|	
7. **Click the mouse on the tree**, and a red vertical line appears. The red line groups sequences. Each group has its own randomly generated colour.	Neighbour Joining Using BLOSUM62 from https://www.jalview.org/tutorial/schools
8. **Click the mouse at different locations in the tree**, this will change the grouping and the colours of groups. Each branch has a number, they can be added together to determine the similarity scores.	Neighbour Joining Using BLOSUM62 from http://www.jalview.org/tutorial/myoglobins.n
9. **Select the View** menu in the tree window in the tree window. **Select Sort Alignment by Tree**.	Neighbour Joining Using BLOSUM62 from http://www.jalview.org/tutorial/myoglobins.n
10. The sequences in the alignment window are reordered to reflect the sequence order in the tree. Note how the sequence names in the alignment window have the same colours as those in the tree window.	
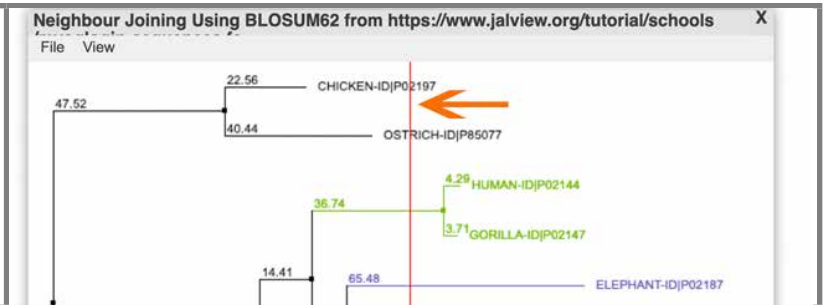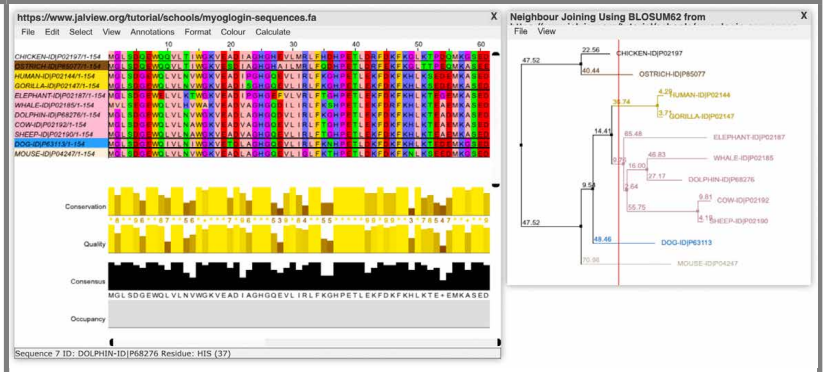
Q. Looking at the tree window, are the groupings what you might expect?
(The shorter the length of the branches between species, the more similar they are. The length between each branch points is shown in brackets).
Q: From the tree (step 8), which animal has the most similar myoglobin to humans?
Q: From the tree (step 8), compared to humans which animal has the least similar myoglobin?

Uniprot identifier codes used in this exercise

Uniprot ID	Species	Uniprot ID	Species	Uniprot ID	Species
P02144	Human	P04247	Mouse	P63113	Dog
P02147	Gorilla	P68276	Dolphin	P02190	Sheep
P02185	Whale	P02197	Chicken	P02187	Elephant
P02192	Cow	P85077	Ostrich		
P02144; P02147; P02185; P02192; P04247; P68276; P02197; P85077; P63113; P02190; P02187					

Project 3:
'Viewing Proteins'

Exercise 6: Viewing Proteins with Different Functions

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.

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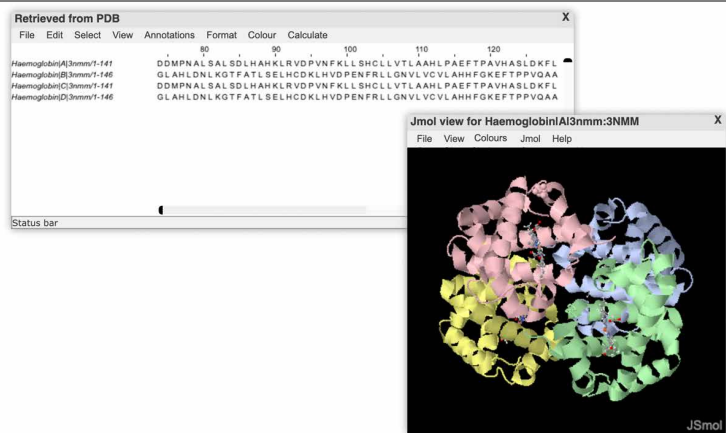- Colour the amino acid residues
- View the 3D structures

1. **Click the link 'view Haemoglobin protein**' in Project 3-Exercise 6 on the Schools web page.	**Project 3** Exercise 6: View protein sequences and 3D structures • Click to view **Haemoglobin** protein (transport) • Click to view **Collagen** protein (structure) • Click to view **Amylase** protein (enzyme) • Click to view **Antibody** protein (defence) • Click to view **Insulin** protein (signalling) • Click to view **Ferritin** protein (storage) • Click to view **Myoglobin** protein (transport)
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. 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 a ball-and-stick model. The heme consists of an iron atom (orange) within a heterocyclic ring.	
5. **Click the link** '**view Collagen protein**' in Project 3-Exercise 6 on the Schools web page.	**Project 3** Exercise 6: View protein sequences and 3D structures • Click to view **Haemoglobin** protein (transport) • Click to view **Collagen** protein (structure) • Click to view **Amylase** protein (enzyme) • Click to view **Antibody** protein (defence)
6. In an adjacent window, the collagen protein sequence and its 3D structure opens in JalviewJS.	
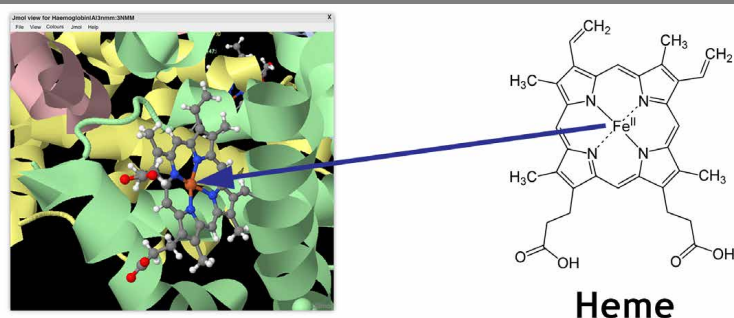
Exercise 6: Viewing Proteins with Different Functions

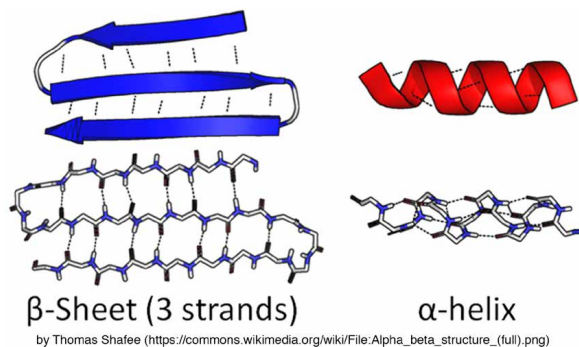|	
7. **Select the Colour** menu. **Select a colour scheme** eg Zappo.	
8. **Select the Colour menu** in the Jmol structure window. **Select By Chain** to view each of the chains in the 3D structure.	
9. **Click in the 3D structure window with the mouse**. **Drag the mouse** across the window to change the view.	
Repeat steps 5-9 for the other proteins listed on the web page (amylase, antibody, insulin, ferritin and myoglobin). **Note: Some of the files in these exercises are quite large so they may take a little while to open.**	**Project 3** Exercise 6: View protein sequences and 3D structures • Click to view **Haemoglobin** protein (transport) • Click to view **Collagen** protein (structure) • Click to view **Amylase** protein (enzyme) • Click to view **Antibody** protein (defence) • Click to view **Insulin** protein (signalling) • Click to view **Ferritin** protein (storage) • Click to view **Myoglobin** protein (transport)

Q. How many different protein amino acids are found in nature? (see Table 3 in the Appendix)
Q. What is the name of the amino acid that is represented by the letter A?
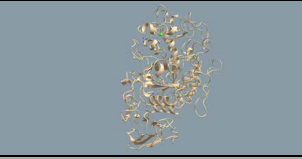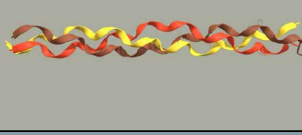Q. What is the chemical formula of the amino acid glycine (G)?
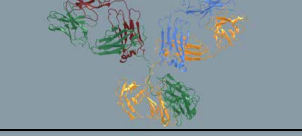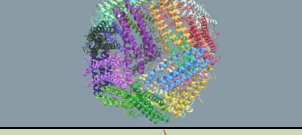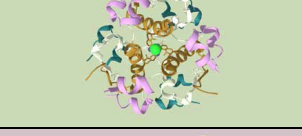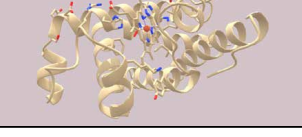
Biologists use four terms to describe protein structure:-

- ***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 (eg alpha helix and beta sheet). 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.

β-Sheet (3 strands)   α-helix

by Thomas Shafee (https://commons.wikimedia.org/wiki/File:Alpha_beta_structure_(full).png)

Q. Review the secondary structure of each protein, are there alpha-helix or beta-sheet regions present in the 3D structures?

Exercise 6: Viewing Proteins with Different Functions

Protein Type	Protein Function (role)	Examples	PDB IDs	SketchFab* Model Links	3D Model in SketchFab
Enzyme	Catalysis of chemical reactions	**Amylase** https://pdb101.rcsb.org/motm/74	**1SMD**	*https://skfb.ly/6ZDBE*	
Structure	Provides mechanical support to cells & tissues	**Collagen** https://pdb101.rcsb.org/motm/4	**1CAG & 1K6F**	*https://skfb.ly/6ZDDR*	
Defence	Protection against disease	**Antibody** https://pdb101.rcsb.org/motm/21	**1IGT & 1IGY**	*https://skfb.ly/6ZDDZ*	
Storage	Stores small molecules or ions	**Ferritin** https://pdb101.rcsb.org/motm/35	**5XB1 & 1MFR**	*https://skfb.ly/6ZDE6*	
Signalling	Regulates body metabolism & the nervous system	**Insulin** https://pdb101.rcsb.org/motm/14	**1TRZ**	*https://skfb.ly/6ZDEo*	
Transport	Carry substances around the body	**Myoglobin** https://pdb101.rcsb.org/motm/1	**3RGK & 1MBO**	*https://skfb.ly/6ZDEp*	

* 3D models of each of the proteins above have been uploaded in SketchFab, click the links in the table to view them.

Project 4:
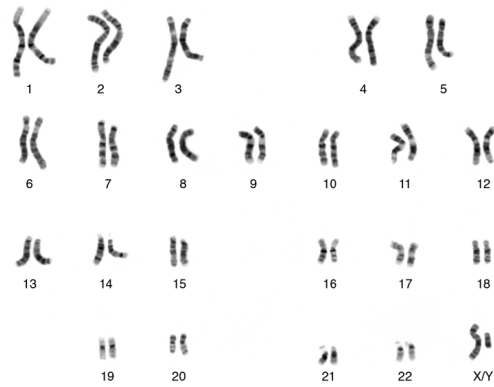# 'Viewing the Genetic Mutation involved for Sickle Cell Anaemia'

What are Chromosomes & Genes?

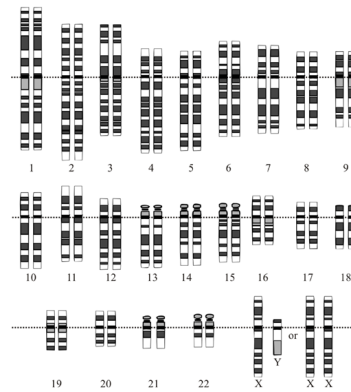|	
A **chromosome** is made up of **DNA** tightly coiled around proteins called histones. Each human cell normally contains 23 pairs of chromosomes, i.e. a total of 46 chromosomes. This is an image of a set of 46 stained metaphase chromosomes from a male human. [Image courtesy of National Human Genome Research Institute]	
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. [Image courtesy of NCBI Genome Decoration Page]	
This is a pictorial representation of human chromosome 11. Each chromosome contains several **genes**. Genes are regions of DNA that code for specific proteins. [Image courtesy of National Library of Medicine]	
The location of the haemoglobin-beta (*HBB*) gene, the gene involved in sickle cell disease, is marked with an orange arrow. [Image courtesy of National Library of Medicine]	
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. [Image courtesy of National Human Genome Research Institute]	
Figure showing the relationship between the DNA in chromosomes, genes, exons, and introns. [Image courtesy of Thomas Splettstoesser]	
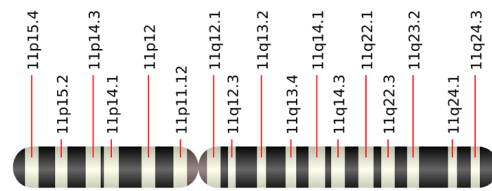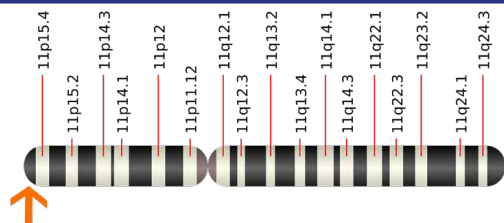
Exercise 7: What are Exons & Introns?

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.
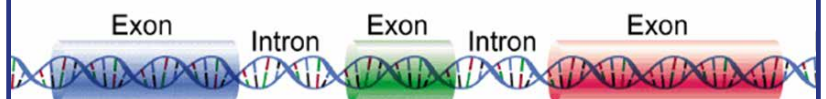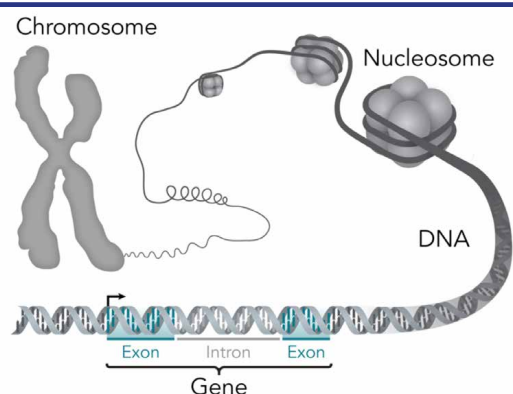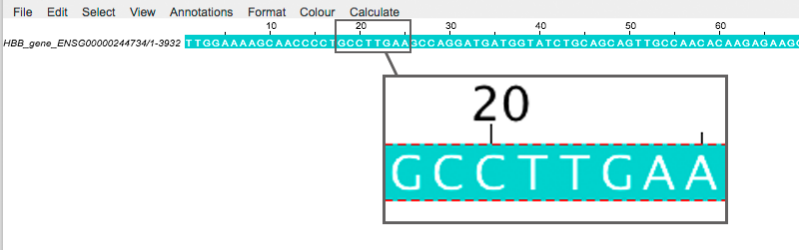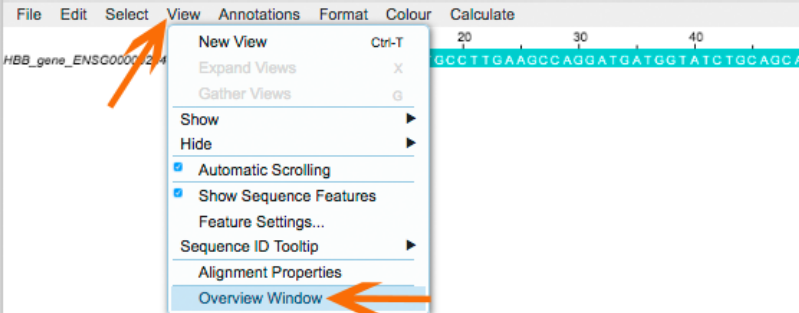The disease is autosomal recessive which means that only people with 2 copies of the sickle cell gene have the condition. A study in 2015 suggests that about 4.4 million people have this disease.

Learning Objectives:
• Identify the exon and intron regions of the *HBB* gene

1. **Click the link 'view Exons & Introns in the DNA of the HBB gene**' in Project 4-Exercise 7 on the Schools web page.	• Click to view **Insulin** protein (signalling) • Click to view **Ferritin** protein (storage) • Click to view **Myoglobin** protein (transport) **Project 4** Exercise 7: Click to view **Exons & Introns** in the DNA of the *HBB* gene Exercise 8: Click to view the **coding DNA** of the *HBB* gene & its **protein product** Exercise 8 (*Advanced Version*): Click to view the **coding DNA** of the *HBB* gene & its **protein product** reading files from public biological databases (*EMBLCDS ID = CAG46711*)
2. JalviewJS with the *HBB* gene opens in an adjacent browser 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** in the overview window shows the part of the sequence visible in the alignment window. **Drag the red box to the right** using the mouse and view the effect in the alignment window.	Overview http://www.jalview.org/tutorial/HBBgene.fa
4. The bases in the haemoglobin beta (*HBB*) gene are labelled with the 1-letter identifier: G (guanine), T (thymine), C (cytosine) and A (adenine).	20 GCCTTGAA
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 & its protein product

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

1. **Click the link** '**view the coding DNA of the *HBB* gene & its protein product**' in 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 residue** 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 mutation linked to sickle cell anaemia is highlighted in red on both the DNA and protein sequences.	
5. In the upper 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 residue (E) at residue 7 (highlighted in red).** A second tooltip opens, note its information. Additional information is available in **Status Bar** in lower left-hand corner. If the tooltip doesn't open, right click the mouse to open a context menu and select **Feature Details** instead.	
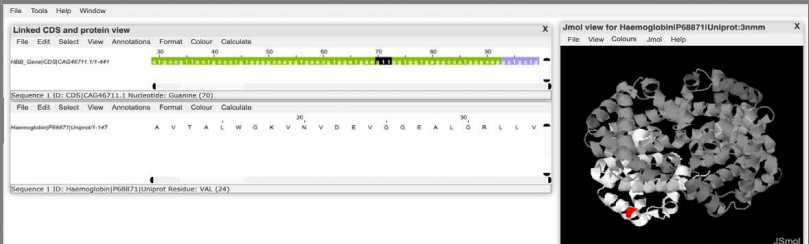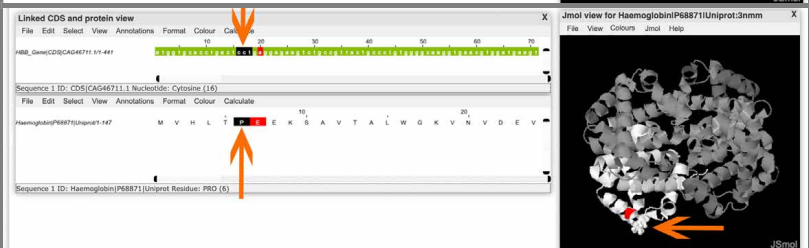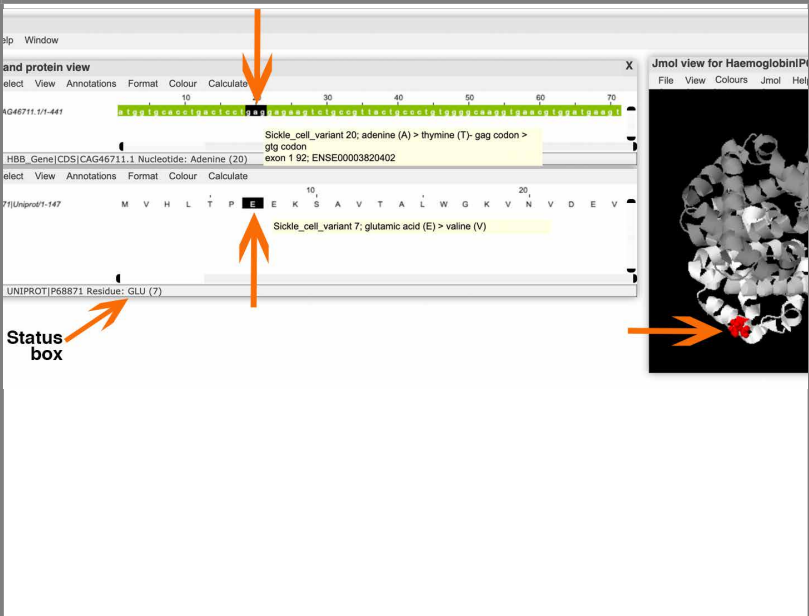
Q. What nucleotide triplet codon code produces the glutamic acid at residue 7?

Q. In the mutated HBB protein, what amino acid replaces the glutamic acid at residue 7?

Exercise 8: Viewing coding DNA sequence & its protein product

The location of the mutation (highlighted in red) on the haemoglobin model can be viewed in SketchFab (https://skfb.ly/6ZDNG).
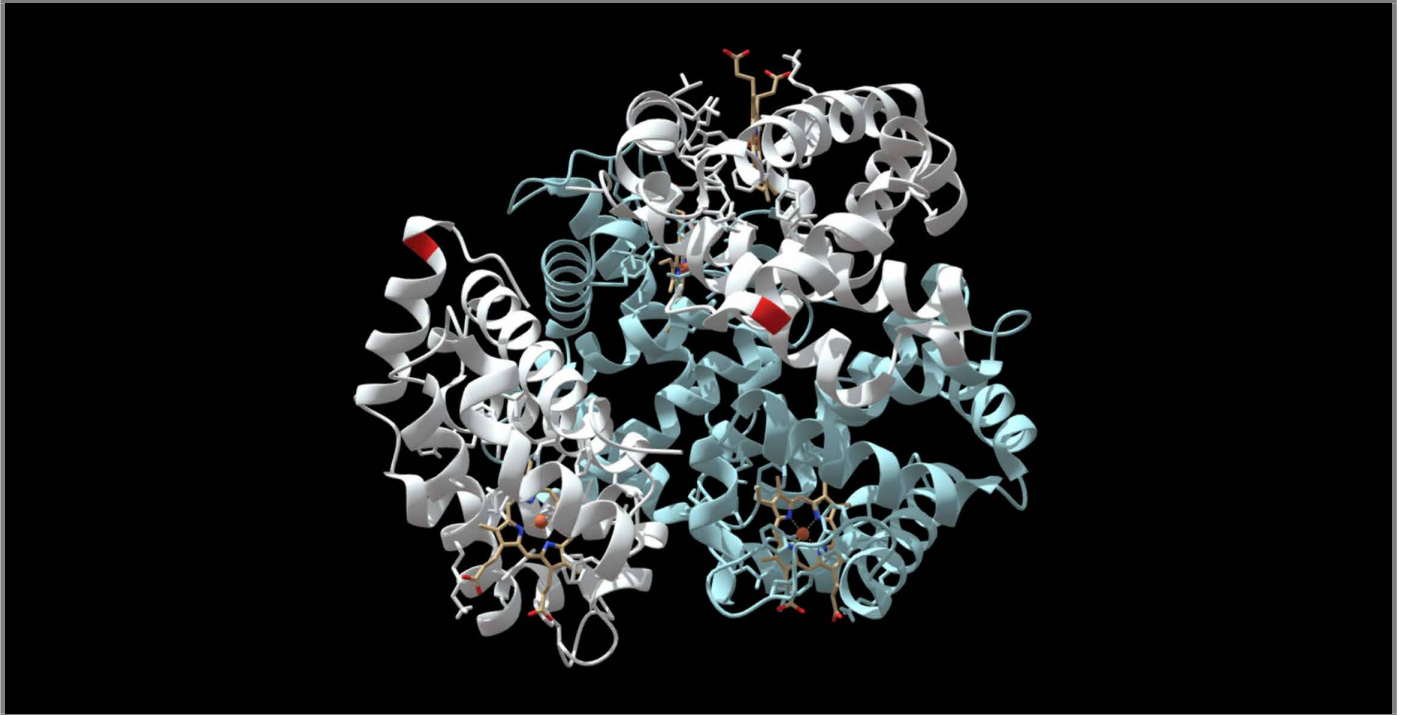
Q. Does every person with this *HBB* gene mutation get sickle cell anaemia?

Q. What happens to the red blood cells of people with sickle cell anaemia?

Q. What evolutionary advantage does heterozygote sickle cell mutation (only 1 of the *HHB* genes have the mutation) infers on people in countries such as Africa?

(see https://en.wikipedia.org/wiki/Heterozygote_advantage).

Advanced Exercise 1: Fetching files from Biological Databases

> **Learning Objectives:**
> • Fetch protein sequences from a public biological database and open them in JalviewJS

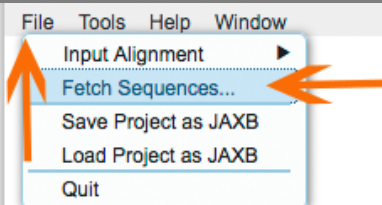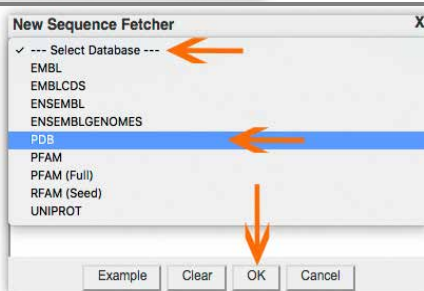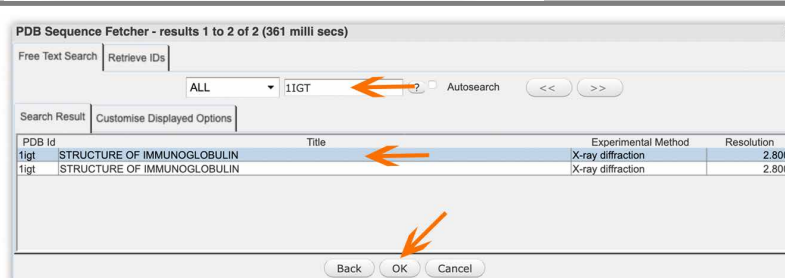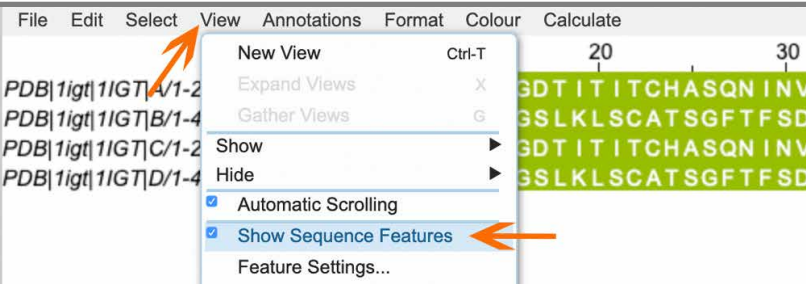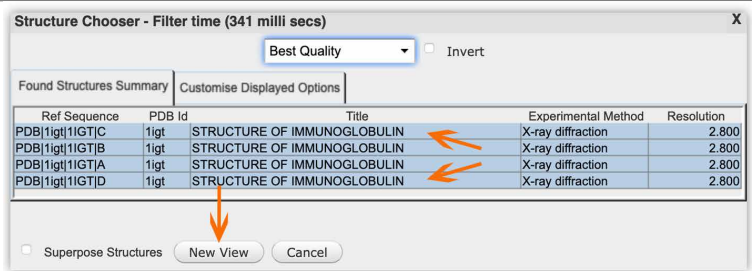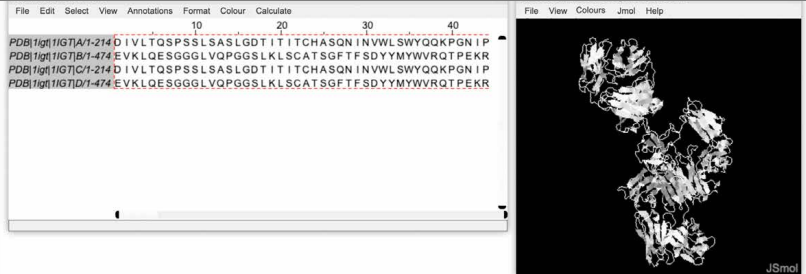|	
1. **Click the link** 'JalviewJS' at the top of the 'Links to Run Exercises' section on the Schools web page.	**Links to Run the Exercises** Clicking the links below will open JalviewJS in an adjacent browser tab and read in files as appropriate. *Note* that it may take a few seconds to load the sequences. **Project 1** Exercise 1: Click to view a fragment of DNA Exercise 2: Click to view the coding sequence of DNA and codons alongside the protein product Exercise 3: Click to view an RNA molecule
2. **Select the File menu** in the JalviewJS desktop window. **Select Fetch Sequences**.	File Tools Help Window Input Alignment ▶ Fetch Sequences... Save Project as JAXB Load Project as JAXB Quit
3. In the 'New Sequence Fetcher' box, **select PDB** from the 'Select Database' list. **Select/Click OK**.	**New Sequence Fetcher** X ✓ --- Select Database --- EMBL EMBLCDS ENSEMBL ENSEMBLGENOMES PDB PFAM PFAM (Full) RFAM (Seed) UNIPROT Example Clear OK Cancel
4. **Enter the PDB ID code** eg **1IGT** in the 'PDB Sequences Fetcher' box. Press the **return key**. **Select the PDB ID code** from the list ie **1Igt**. **Click OK.** (Close the 'PDB Sequence Fetcher' box once the sequence has loaded).	PDB Sequence Fetcher - results 1 to 2 of 2 (361 milli secs) X Free Text Search Retrieve IDs ALL 1IGT ? Autosearch << >> Search Result Customise Displayed Options PDB Id | Title | Experimental Method | Resolution 1igt | STRUCTURE OF IMMUNOGLOBULIN | X-ray diffraction | 2.800 1igt | STRUCTURE OF IMMUNOGLOBULIN | X-ray diffraction | 2.800 Back OK Cancel
5. **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).	File Edit Select View Annotations Format Colour Calculate New View Ctrl-T Expand Views X Gather Views G Show ▶ Hide ▶ ✓ Automatic Scrolling ✓ Show Sequence Features Feature Settings... PDB|1igt|1IGT|A/1-2... PDB|1igt|1IGT|B/1-4... PDB|1igt|1IGT|C/1-2... PDB|1igt|1IGT|D/1-4...
6. **Click and drag the mouse to select all the names of the sequences** in the alignment window. **Right click the mouse** to open the pop-up menu. **Select 3D Structure Data**.	File Edit Select View Annotations Format Colour Calculate 10 20 PDB|1igt|1IGT|A/1-214 DIVLTQSPSSLSASLGDTITITCHAS PDB|1igt|1IGT|B/1-474 EVKLQESGGGLVQPGGSLKLSCATSG PDB Selection ▶ VLTQSPSSLSASLGDTITITCHAS PDB PDB|1igt|1IGT|B ▶ KLQESGGGLVQPGGSLKLSCATSG 3D Structure Data... Hide Insertions Hide Sequences
Troubleshooting Note: **Place the mouse cursor over the sequence names** and **right click**, <u>not</u> the sequence itself, **to open the context menu.**	This context menu opens when the mouse cursor is placed on the names of the sequences ✓ This context menu opens when the mouse cursor is placed on the sequences ✗
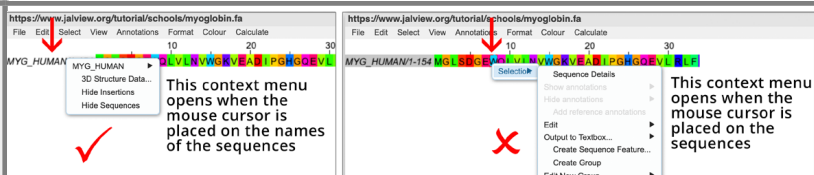
Advanced Exercise 1: Fetching files from Biological Databases

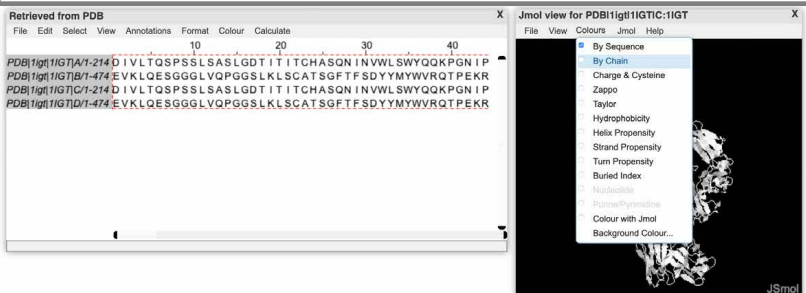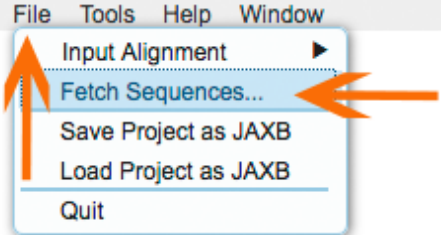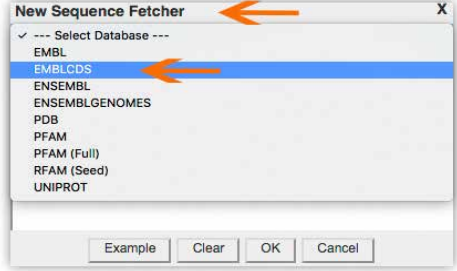|	
7. In the 'Structure Chooser' box, **select the PDB Id entries**. **Click New View**. (Click and drag the mouse or use the Shift key if the IDs are not selected).	
8. A 3D Jmol structure window opens.	
9. **Select the Colour menu** in Jmol window. **Select By Chain** to colour each of the chains in the 3D structure seperately.	
10. The protein sequences of the IgG2a monoclonal antibody is visualized in JalviewJS alongside its 3D structure. Visit the PDB web site for more information: www.rcsb.org/structure/1igt	
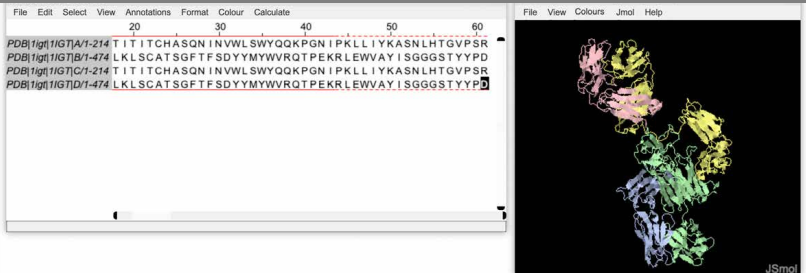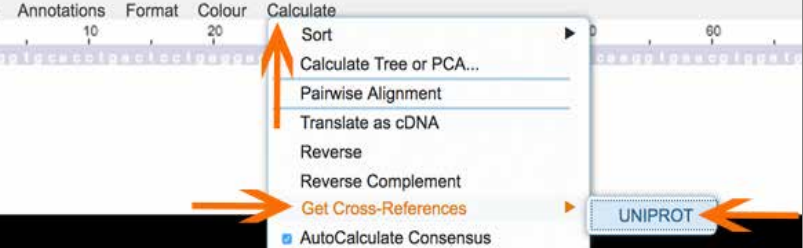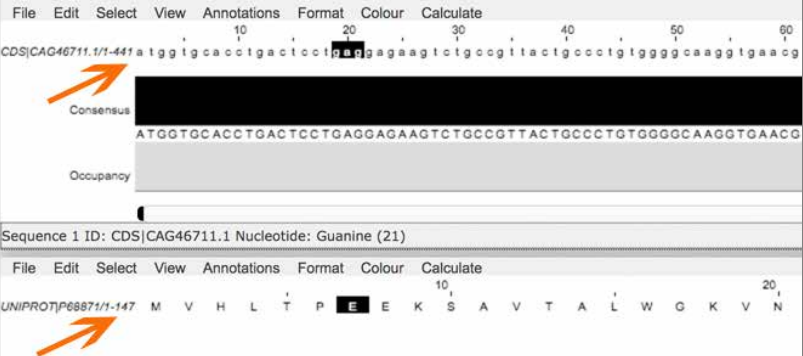
Free Public Biological Databases:

- **UniProt** is a database of protein sequences (https://www.uniprot.org/).
- **Protein Data Bank (PDB)** is a database of crystallographic, three-dimensional structural data of large biological molecules (https://www.rcsb.org/).
- **Ensembl** is a genomic database (https://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).

Advanced Exercise 2: View the *HBB* Gene & HBB Protein

> **Learning Objectives:**
> · Fetch the DNA sequence of the *HBB* gene from a biological database
> · Open the HBB protein sequence in the split-frame window
> · Load the sickle cell anaemia feature files onto the sequences
> · Open the 3D structure of human haemoglobin into the Jmol structure viewer
> · Identify the location of the point mutation on the haemoglobin structure

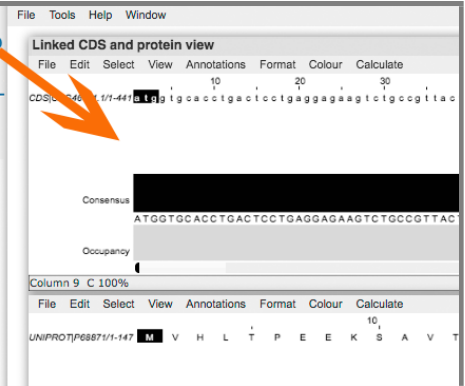|	
1. **Click the link 'view the coding DNA of the HBB gene & its protein product -**' in the Advanced version of Exercise 8.	· Click to view **Myoglobin** protein (transport) **Project 4** Exercise 7: Click to view **Exons & Introns** in the DNA of the *HBB* gene Exercise 8: Click to view the **coding DNA** of the *HBB* gene & its **protein product** Exercise 8 (*Advanced Version*): Click to view the **coding DNA** of the *HBB* gene & its **protein product** reading files from public biological databases (*EMBLCDS ID = CAG46711*)
2. In the empty JalviewJS desktop window, **select the File** menu. **Select Fetch Sequences**.	File Tools Help Window Input Alignment ▶ Fetch Sequences... Save Project as JAXB Load Project as JAXB Quit
3. **Select EMBLCDS** from the 'Select Database' list in the 'New Sequence Fetcher' box.	**New Sequence Fetcher** X ✓ --- Select Database --- EMBL EMBLCDS ENSEMBL ENSEMBLGENOMES PDB PFAM PFAM (Full) RFAM (Seed) UNIPROT Example Clear OK Cancel
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).	**New Sequence Fetcher** X EMBLCDS Example query: CAA37824 *Enter one or more accession IDs separated by a semi-colon ";"* ☐ *Replace commas with semi-colons* CAG46711 Example Clear OK Cancel
5. **Select the Calculate** menu in the alignment window. **Select Get Cross References**. **Select Uniprot**. (Note: You may have to wait a little while Jalview fetches the data).	Annotations Format Colour Calculate Sort ▶ Calculate Tree or PCA... Pairwise Alignment Translate as cDNA Reverse Reverse Complement Get Cross-References ▶ UNIPROT ☑ AutoCalculate Consensus
6. A split-frame 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 residue in the protein** and **view the associated codon in the DNA sequence**.	File Edit Select View Annotations Format Colour Calculate CDS|CAG46711.1/1-441 atggtgcacctgactcct**gag**gagaagtctgccgttactgccctgtggggcaaggtgaacg Consensus ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACG Occupancy Sequence 1 ID: CDS|CAG46711.1 Nucleotide: Guanine (21) File Edit Select View Annotations Format Colour Calculate UNIPROT|P68871/1-147 M V H L T P **E** E K S A V T A L W G K V N

Advanced Exercise 2: View the *HBB* Gene & HBB Protein

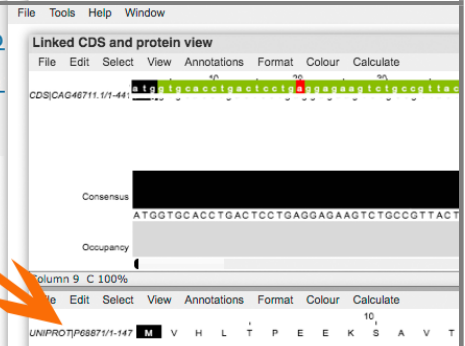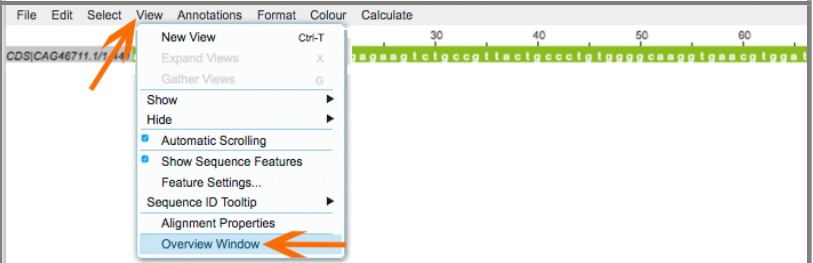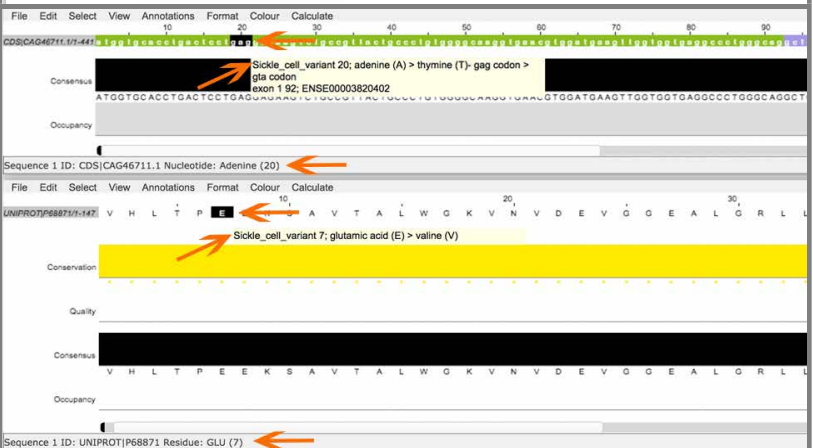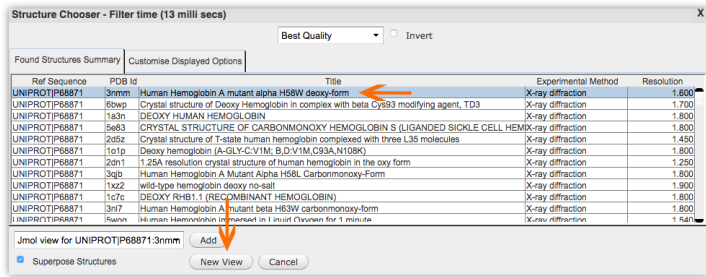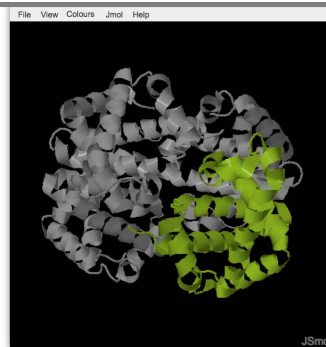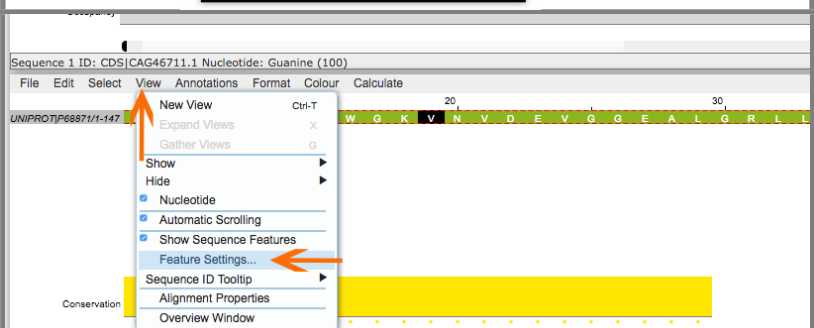|	
7. There are <u>two links</u> in the top left hand corner of the JalviewJS web page. **Click-and-drag the top link 'Drag this link on DNA sequence to add features'** onto the *HBB* gene sequence panel in the upper DNA alignment window. This opens the features file and colours the DNA sequence.	
8. **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.	
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**. A 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**.	
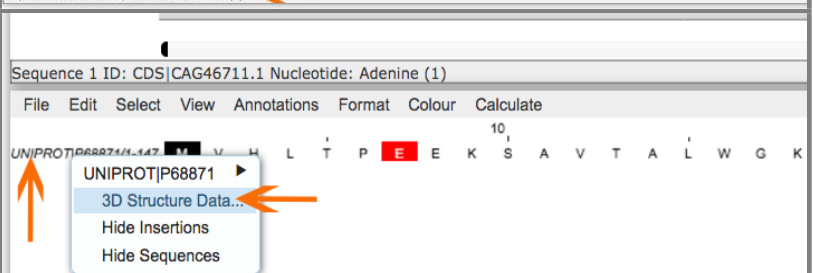
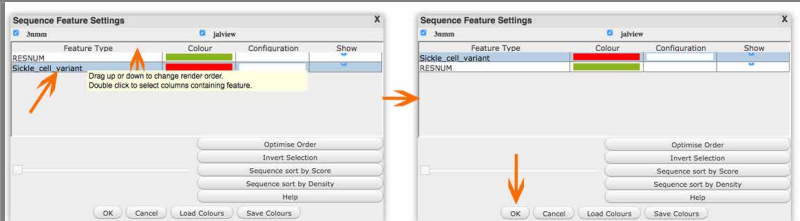|	
13. In the 'Structure Chooser' box, **select PDB id 3nmm** entry. **Select New View**.	
14. A 3D structure window opens containing the HBB protein.	
15. In the protein sequence panel, **select the View** menu in the alignment window. **Select Features Settings...**	
16. In the 'Sequence Feature Settings' box, **click the mouse on the red Sickle_cell_variant feature name**. **Drag it to the top of the list**, <u>above</u> 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-frame 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.	

These two advanced exercises develop a range of skills needed to run Jalview, including:
• launching JalviewJS
• reading sequence files from public biological database
• opening 3D structures files in the structure viewer
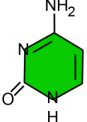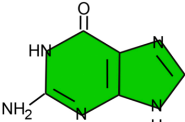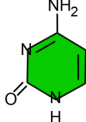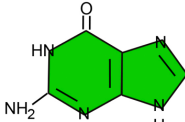• opening the split-frame view
• loading features
These skills can be used to study other biological molecules present in public databases. We suggest a good place to start researching other biomole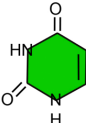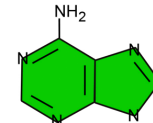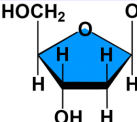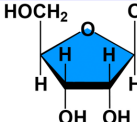cules is the 'Molecule of the Month' section (https://pdb101.rcsb.org/motm/motm-by-category) in the RCSB PDB 101 website.
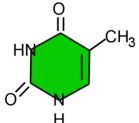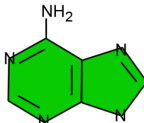
Appendix

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.

TTT, TTC → F	CTT, CTC, CTA, CTG → L	ATT, ATC, ATA → I	GTT, GTC, GTA, GTG → V
TTA, TTG → L		ATG → M start	
TCT, TCC, TCA, TCG → S	CCT, CCC, CCA, CCG → P	ACT, ACC, ACA, ACG → T	GCT, GCC, GCA, GCG → A
TAT, TAC → Y	CAT, CAC → H	AAT, AAC → N	GAT, GAC → D
TAA, TAG → stop	CAA, CAG → Q	AAA, AAG → K	GAA, GAG → E
TGT, TGC → C	CGT, CGC, CGA, CGG → R	AGT, AGC → S	GGT, GGC, GGA, GGG → G
TGA → stop; TGG → W		AGA, AGG → R	

- The DNA is made up of **nucleotide bases**.
- DNA regions that **code** for proteins are called **exons**, whilst **non-coding** regions are called **introns**.
- Intron are transcribed but then removed during the production of **messenger RNA** (mRNA).
- **Genes** in the DNA code for specific **proteins**.

Appendix
Amino Acids

Amino acids are the building blocks (sub-units) of proteins. Attached to the central carbon is a hydrogen, a NH_2 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.
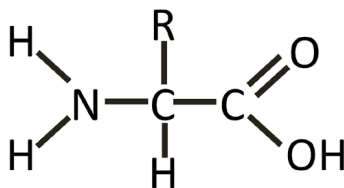
Amino acids link together by forming peptide bonds between the COOH and NH_2 of neighbouring amino acids during a condensation reaction that releases water. Multiple amino acids join together to form a polypeptide. A protein is produced when the amino acid strand contains more than 50 amino acid residues.
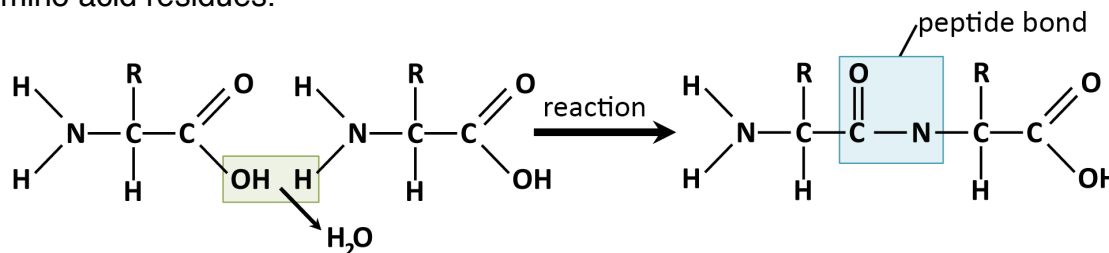
Table 3: Amino Acids

Amino Acid	3 letter code	1 letter code	Sidechain polarity	Linear	Species
Alanine	Ala	A	nonpolar	CH_3-$CH(NH_2)$-COOH	$C_3H_7NO_2$
Arginine	Arg	R	basic polar	$HN=C(NH_2)$-NH-$(CH_2)_3$-$CH(NH_2)$-COOH	$C_6H_{14}N_4O_2$
Asparagine	Asn	N	polar	H_2N-CO-CH_2-$CH(NH_2)$-COOH	$C_4H_8N_2O_3$
Aspartic acid	Asp	D	acidic polar	HOOC-CH_2-$CH(NH_2)$-COOH	$C_4H_7NO_4$
Cysteine	Cys	C	nonpolar	HS-CH_2-$CH(NH_2)$-COOH	$C_3H_7NO_2S$
Glutamic acid	Glu	E	acidic polar	HOOC-$(CH_2)_2$-$CH(NH_2)$-COOH	$C_5H_9NO_4$
Glutamine	Gln	Q	polar	H_2N-CO-$(CH_2)_2$-$CH(NH_2)$-COOH	$C_5H_{10}N_2O_3$
Glycine	Gly	G	nonpolar	H-$CH(NH_2)$-COOH	$C_2H_5NO_2$
Histidine	His	H	basic polar	NH-CH=N-CH=C-CH_2-$CH(NH_2)$-COOH	$C_6H_9N_3O_2$
Isoleucine	Ile	I	nonpolar	CH_3-CH_2-$CH(CH_3)$-$CH(NH_2)$-COOH	$C_6H_{13}NO_2$
Leucine	Leu	L	nonpolar	$(CH_3)_2$-CH-CH_2-$CH(NH_2)$-COOH	$C_6H_{13}NO_2$
Lysine	Lys	K	basic polar	H_2N-$(CH_2)_4$-$CH(NH_2)$-COOH	$C_6H_{14}N_2O_2$
Methionine	Met	M	nonpolar	CH_3-S-$(CH_2)_2$-$CH(NH_2)$-COOH	$C_5H_{11}NO_2S$
Phenylalanine	Phe	F	nonpolar	Ph-CH_2-$CH(NH_2)$-COOH	$C_9H_{11}NO_2$
Proline	Pro	P	nonpolar	-NH-$(CH_2)_3$--*CH-COOH	$C_5H_9NO_2$
Serine	Ser	S	polar	HO-CH_2-$CH(NH_2)$-COOH	$C_3H_7NO_3$
Threonine	Thr	T	polar	CH_3-CH(OH)-$CH(NH_2)$-COOH	$C_4H_9NO_3$
Tryptophan	Trp	W	nonpolar	Ph-NH-CH=C-CH_2-$CH(NH_2)$-COOH	$C_{11}H_{12}N_2O_2$
Tyrosine	Tyr	Y	polar	HO-Ph-CH_2-$CH(NH_2)$-COOH	$C_9H_{11}NO_3$
Valine	Val	V	nonpolar	$(CH_3)_2$-CH-$CH(NH_2)$-COOH	$C_5H_{11}NO_2$

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.

Nucleoside:- nucleotides without a phosphate group.

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.