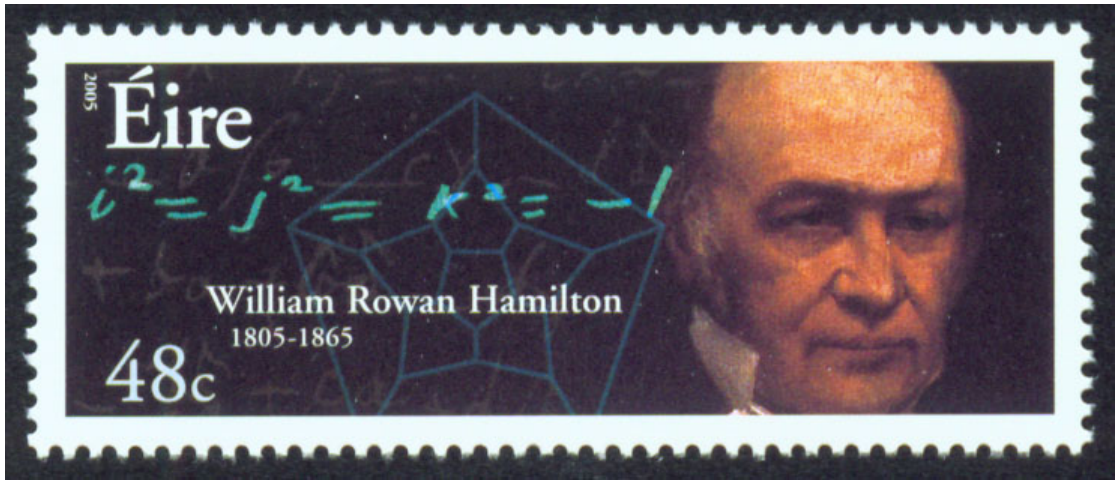
An aide to memory



Students sometimes have trouble remembering the difference between Eulerian and Hamiltonian tours and I'm not unsympathetic: after all, both are named after very famous, long-dead, European mathematicians.

- Almost everything about Eulerian multigraphs begins with 'E': *Eulerian* tours include every *edge* and are *easy* to find when all vertices have *even* degree.
- Some of the words needed to describe Hamiltonian graphs start with 'H': *Hamiltonian* tours include every vertex and are *hard* to find.

These slides are available at `https://bit.ly/3tJnBr3`

The remaining material in these slides is not examinable

Theorem (Directed Eulerian Tours)

Let $G(V, E)$ be a strongly-connected, directed multigraph. Then the following statements are equivalent:

(i) *G has an Eulerian tour.*

(ii) *For all vertices $v \in V$ we have $\deg_{in}(v) = \deg_{out}(v)$.*

(iii) *The edge set of G can be partitioned into cycles.*

The proof of this theorem is very similar to that of the undirected version covered in the vides, slides and lecture notes.

These slides are available at `https://bit.ly/3tJnBr3`

Another problem, not obviously involving graphs . . .

Ribonucleic acid (RNA) is a family of molecules whose members play a great many roles[1] in the chemistry of living cells. RNA is similar to its more famous sister molecule, DNA (deoxyribonucleic acid), in that it consists of long chains composed of 4 subunits called *bases*. For RNA the names of the bases are usually abbreviated `U`, `C`, `G`, and `A` and they can be combined in any order, so a mathematician can, if she likes, think of an RNA molecule as a string of letters such as

`AGUCAGUGAGCA`

These slides are available at `https://bit.ly/3tJnBr3`

[1] If you want to learn more about RNA and its amazingly rich biochemistry, I recommend a book by (Manchester Uni's own) Terry Brown, *Genomes 3*, Garland Scientific (2006). The previous edition, *Genomes 2*, is freely available online at https://bit.ly/Genomes2.

Sequencing RNA in the 1960's

In the early days of RNA research biochemists could sequence accurately only rather short pieces of RNA, but were interested in much longer chains. They addressed the problem by chopping up the longer chains into smaller fragments that they could sequence directly. These collections of smaller fragments are called *digests* of the original molecule and one way to sequence a long chain is to compare and combine the results of two different sorts of digest.

These slides are available at `https://bit.ly/3tJnBr3`

Two different sorts of digest

Our first enzyme cuts the RNA sequence after every G, while the second cuts after both U and C:

$$\texttt{AGUCAGUGAGCA} \xrightarrow{\text{G-digest}} \texttt{AG UCAG UG AG CA}$$

$$\texttt{AGUCAGUGAGCA} \xrightarrow{\text{UC-digest}} \texttt{AGU C AGU GAGC A}$$

Of course, the digests don't come so neatly organized: in a real scientific problem we'd just get two jumbles of short fragments such as

G-digest: AG AG CA UG UCAG
UC-digest: A C AGU AGU GAGC

and have to work out how to rearrange them in a sensible order.

These slides are available at https://bit.ly/3tJnBr3

A thought experiment: digesting the digests

Imagine attacking each fragment from the G-digest with the UC-enzyme. They'd break down as follows:

$$\texttt{AG} \xrightarrow{\text{UC-digest}} \texttt{AG} \qquad \texttt{UCAG} \xrightarrow{\text{UC-digest}} \texttt{U C AG}$$

$$\texttt{UG} \xrightarrow{\text{UC-digest}} \texttt{U G} \qquad \texttt{AG} \xrightarrow{\text{UC-digest}} \texttt{AG} \qquad \texttt{CA} \xrightarrow{\text{UC-digest}} \texttt{C A}$$

The small pieces produced in this way are called *extended bases*.

Note that we'd get the same set of extended bases if we'd started with the fragments of the UC-digest and attacked then with the G-enzyme.

These slides are available at `https://bit.ly/3tJnBr3`

We'll mainly be interested in those cases, such as UCAG $\xrightarrow{\text{UC-digest}}$ U C AG, where a second digest produces two or more extended bases: we'll use them to draw a very helpful graph.

The results of the G-digest were

AG AG CA UG UCAG

so the double-digests that yield two or more pieces are

$$\text{CA} \xrightarrow{\text{UC-digest}} \text{C A}$$

$$\text{UCAG} \xrightarrow{\text{UC-digest}} \text{U C AG}$$

$$\text{UG} \xrightarrow{\text{UC-digest}} \text{U G}$$

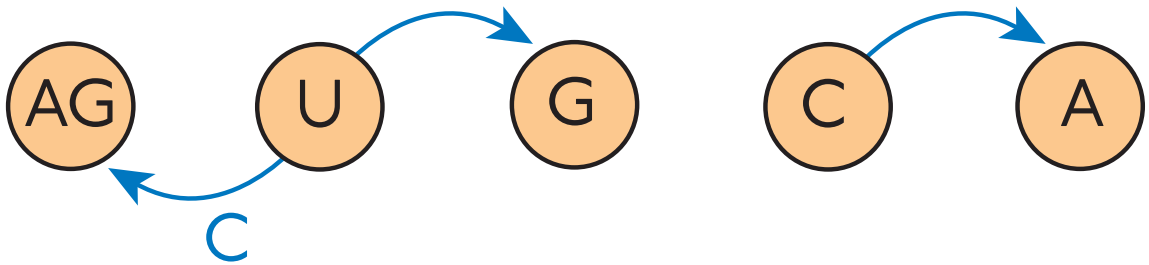
These slides are available at https://bit.ly/3tJnBr3

Starting from the double-digests of the G-fragments

$$\texttt{CA} \xrightarrow{\text{UC-digest}} \texttt{C A} \qquad \texttt{UG} \xrightarrow{\text{UC-digest}} \texttt{U G} \qquad \texttt{UCAG} \xrightarrow{\text{UC-digest}} \texttt{U C AG}$$

we build a graph which has:

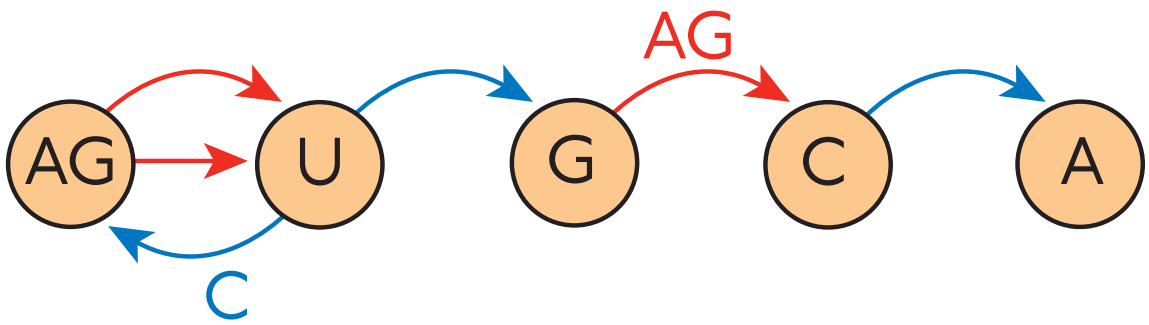
- a vertex for each extended base that appears on the right-hand side of one of the double-digests;
- a directed edge (a bit like a bridge with a one-way street on it) connecting the first and last extended base in each of the double-digests and
- for those double-digests where we get three or more extended bases, an edge label.



We now produce a second set of double-digests by attacking the original UC-fragments—`A C AGU AGU` and `GAGC`—with the G enzyme. The ones that break into two or more extended bases turn out to be

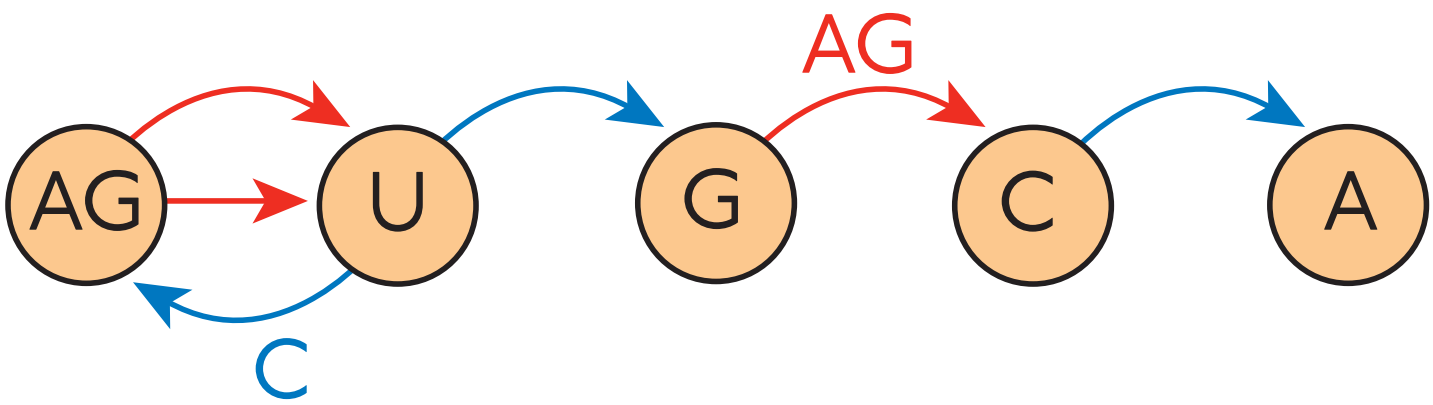
$$\texttt{AGU} \xrightarrow{\text{G-digest}} \texttt{AG U} \text{ (twice)} \qquad \texttt{GAGC} \xrightarrow{\text{G-digest}} \texttt{G AG C}$$

and if we then add the corresponding edges and labels to our graph we end up with:



where the edges and labels arising from the second group of double-digests are shown in red.

These slides are available at `https://bit.ly/3tJnBr3`



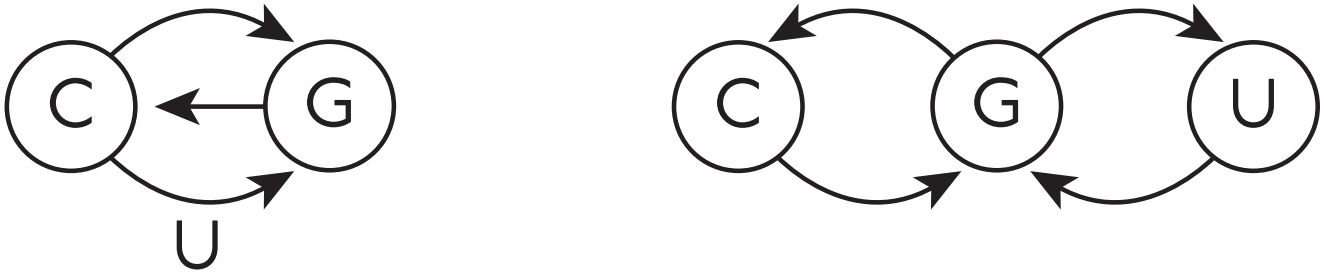
There's exactly one (directed) Eulerian trail through this graph and, if we write down the labels of its vertices and edges in the order we encounter them we get

$$\texttt{AG} \cdot \texttt{U} \cdot \texttt{C} \cdot \texttt{AG} \cdot \texttt{U} \cdot \texttt{G} \cdot \texttt{AG} \cdot \texttt{C} \cdot \texttt{A} = \texttt{AGUCAGUGAGCA}$$

which is the original sequence!

These slides are available at `https://bit.ly/3tJnBr3`

Some graphs have multiple tours or trails



The graph at left above contains two Eulerian trails:

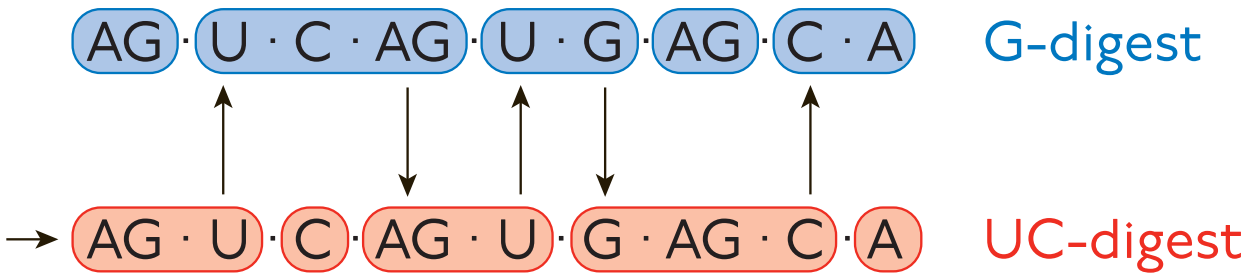
C·G·C·UG and C·UG·C·G.

Both give the same sets of digests. Similarly, the graph on the right has two Eulerian tours

G·C·G·U·G and G·U·G·C·G.

These are the shortest RNA sequences whose digests are ambiguous, in the sense that they produce the same graph as another, distinct sequence.

These slides are available at https://bit.ly/3tJnBr3



The diagram above shows two views of our example sequence, one with G-fragments shown in red rounded boxes and another with the UC-fragments in blue boxes: in both cases, the extended bases are shown separated by dots.

- It's only the larger fragments—those containing two or more extended bases—that contribute to the graph.
- These large fragments overlap in a very particular way: the extended base that marks the end of a large blue fragment also marks the beginning of the next large red one.
- The endpoints of large fragments are the vertices of our graph, while the graph's directed edges, along with their labels, come from the large fragments themselves.

These slides are available at `https://bit.ly/3tJnBr3`

Further Reading

I first learned about this application from a textbook

> Edgar G. Goodaire and Michael M. Paramenter (2006). *Discrete Mathematics with Graph Theory*, 3rd Edition, Prentice Hall. ISBN 0-13-167995-3.

whose authors give lots of related exercises. Early research papers about the Eulerian path approach to RNA sequencing include

> George Hutchinson (1969), Evaluation of Polymer Sequence Fragment Data Using Graph Theory, *Bulletin of Mathematical Biophysics*, **31**:541–562. DOI: 10.1007/BF02476636

> Robert W. Holley *et al.* (1965), Structure of a Ribonucleic Acid, *Science*, **147**:1462–1465. DOI: 10.1126/science.147.3664.1462

while the following article presents a more recent graph-theoretic approach to problems arising in whole-genome sequencing.

> P. E. C. Compeau,*et al.* (2011), How to apply de Bruijn graphs to genome assembly, *Nature Biotechnology*, **29**:987–991. DOI: 10.1038/nbt.2023