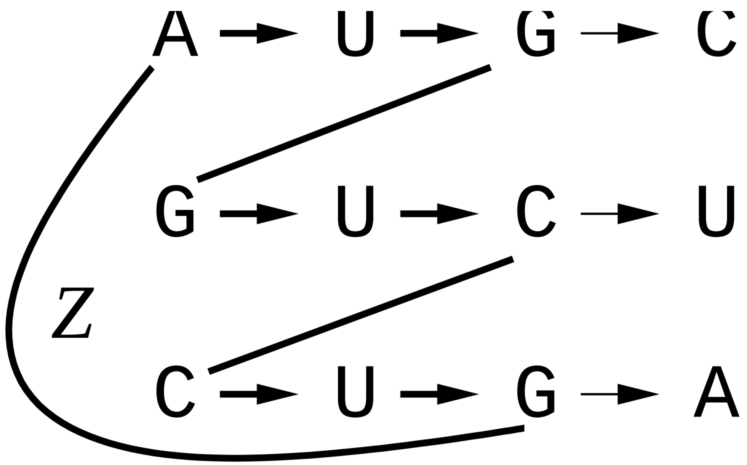
Mixed cycles

For a given choice of alignment edges we can efficiently check whether the connected components allow such a partial order by searching for a *mixed cycle* Z, which is a cycle in the extended alignment graph $G = (V, E, H)$:



A mixed cycle contains at least one arc $a \in H$ and hence at least two alignment edges $e, f \in E$.

Mixed cycles (2)

A mixed cycle Z is called *critical*, if all nodes in $Z \cap a^p$ occur consecutively in Z, for all sequences $a^p \in A$. That is, the cycle enters and leaves each sequence at most once.

We have the following result:

Lemma. A subset $T \subseteq E$ is a trace, if and only if $G' = (V, T, H)$ does not contain a critical mixed cycle.

Proof. Exercise.

Given edge weights for the alignment edges, we can reformulate the Maximum Weight Trace problem as follows:

Problem. Given an extended alignment graph $G = (V, E, H)$, find a subset $T \subseteq E$ with maximal weight such that $G = (V, T, H)$ does not contain a mixed cycle.

Integer LP for the MWT problem

How to encode the Maximum Weight Trace Problem problem as an integer LP?

Assume we are given an extended alignment graph $G = (V, E, H)$, with $E = \{e_1, e_2, \dots, e_n\}$.

Each edge $e_i \in E$ is represented by a variable x_i, that will take on value 1, if e_i belongs to the best scoring trace, and 0, if not.

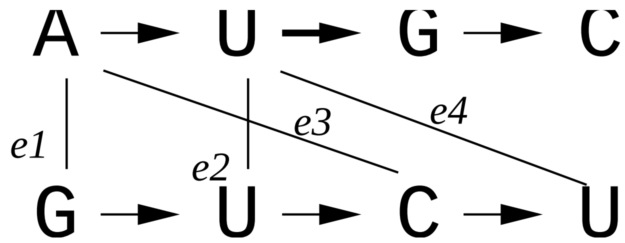
Hence, our variables are $x_1, x_2, \dots, x_n$.

To ensure that the variables are *binary*, we add constraints $x_i \leq 1$ and $x_i \geq 0$ and require the x_i to be integer.

Additional inequalities must be added to prevent mixed cycles.

ILP for the MWT problem (2)

For example, consider:



There are three possible critical mixed cycles in the graph, one using e_1 and e_3, one using e_2 and e_3, and one using e_2 and e_4. We add the constraints

$$x_1 + x_3 \leq 1,$$

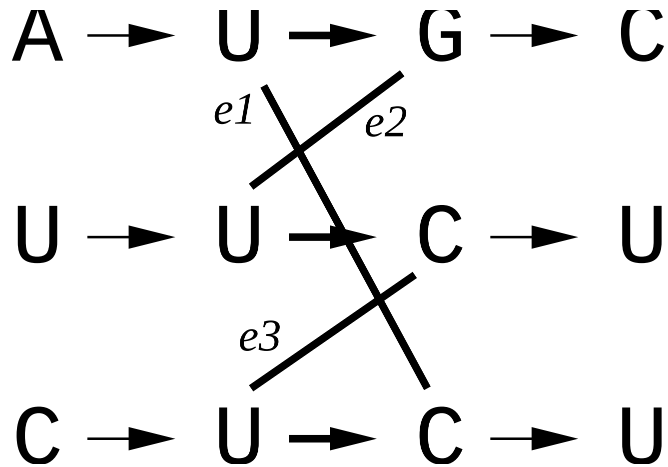
$$x_2 + x_3 \leq 1,$$

$$x_2 + x_4 \leq 1.$$

to ensure that none of the critical mixed cycles is realized.

ILP for the MWT problem (3)

For example, consider:



with three edges e_1, e_2, and e_3 that all participate in a critical mixed cycle. The constraint

$$x_1 + x_2 + x_3 \leq 2$$

prevents them from being realized simutaneously.

ILP for the MWT problem (4)

In summary, given an extended alignment graph $G = (V, E, H)$ with $E = \{e_1, e_2, \dots, e_n\}$, and a score ω_i defined for every edge edge $e_i \in E$.

We can obtain a solution to the MWT problem by solving the following ILP:

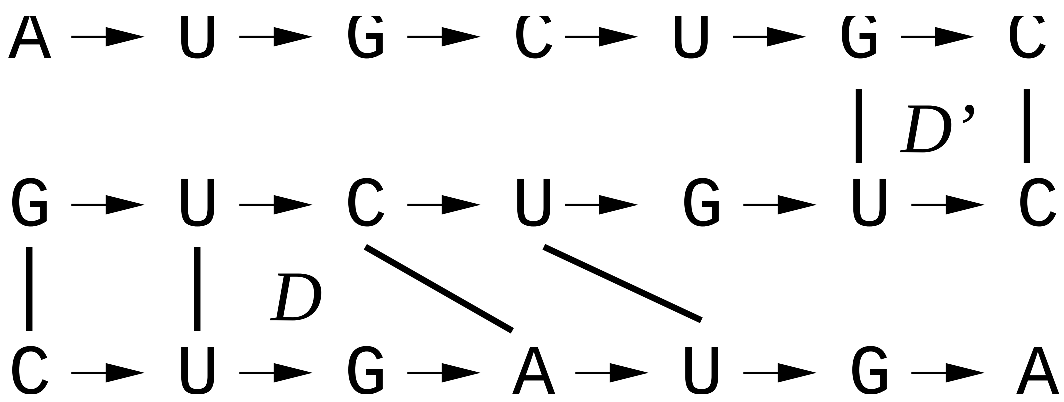
$$\begin{aligned}
\max \quad & \sum_{e_i \in E} \omega_i x_i & \\
\text{subject to} \quad & \sum_{e_i \in C \cap E} x_i \leq |C \cap E| - 1 & \text{for all critical mixed cycles } C \\
& x_i \in \{0, 1\} & \text{for all } i = 1, \dots, n
\end{aligned}$$

Now we discuss an extension of this ILP formulation. [Remember: one advantage of ILPs is that problem variants/modifications can often be expressed quite easily.]

Block partition

Given a set of sequences $A = \{a^1, a^2, \ldots, a^r\}$. The complete alignment graph is usually too big to be useful.

Often, we are given a set of *block matches* between pairs of the sequences a^p and a^q, where a match relates a substring of a^p and a substring of a^q via a run of *non-crossing* edges (called a *block*), as shown here for two blocks D and D':



In the following, we will assume that the edges of the alignment graph $G = (V, E)$ were obtained from a set of matches, and we are given a partition of E into blocks. [Note that overlapping matches lead to a *multigraph*. Fortunately, this does not cause problems in our formulation.]

Generalized maximum trace problem (2)

Given a partition $\mathcal{D}$ of the edges of $G = (V, E)$ obtained from a set of matches. Then we require that for any given block $D \in \mathcal{D}$, either all edges in D are realized, or none. Each block D is assigned a positive weight $\omega(D)$.

Problem. Given an extended alignment graph $G = (V, E, H)$ and a partition $\mathcal{D}$ of E into blocks with weights $\omega(D)$ for all $D \in \mathcal{D}$. The *generalized maximum trace problem (GMT)* is to determine a set $M \subseteq \mathcal{D}$ of maximum total weight such that the edges in $\bigcup_{D \in M} D$ do not induce a mixed cycle on G.

Instead of having a variable for every edge in an extended alignment graph, we now have a variable for every set of the partition $\mathcal{D}$.

Otherwise the ILP remains the same.

ILP for the GMT

We define a surjective function $v : E \longrightarrow \mathcal{D}$, which maps each edge $e \in E$ to the block $d \in \mathcal{D}$ in which e is contained and define

$$v(X) = \bigcup_{e \in X} v(e) \qquad \text{for } X \subseteq E \ .$$

It is now easy to formulate GMT as an integer linear program. For every $d \in \mathcal{D}$ we have a binary variable $x_d \in \{0, 1\}$ indicating whether d is in the solution or not. Then the GMT-problem can be written as:

$$\begin{aligned} \max \quad & \sum_{d \in \mathcal{D}} \omega_d \cdot x_d & \\ \text{s.t.} \quad & \sum_{d \in v(C \cap E)} x_d \le |v(C \cap E)| - 1 & \forall \text{ critical mixed cycles } C \text{ in } G \\ & x_d \in \{0, 1\} & \forall d \in \mathcal{D} \end{aligned}$$