Ordered Colourings

JAN VAN DEN HEUVEL

joint work with ARVIND GUPTA, JÁN MAŇUCH

LADISLAV STACHO & XIAOHONG ZHAO

(Simon Fraser University, Canada)

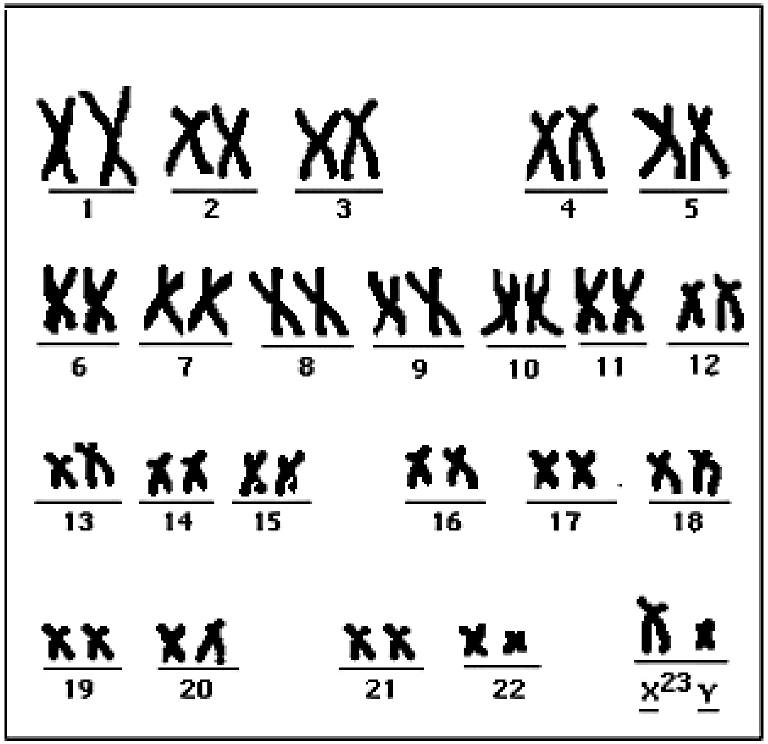
Centre for Discrete and Applicable Mathematics

Department of Mathematics

London School of Economics and Political Science

Inspired by a problem in mathematical biology

- **genotype** : the genetic information of one individual
  - stored on **chromosomes**
- most organisms have **two copies** of each chromosome



Inspired by a problem in mathematical biology

- **genotype** : the genetic information of one individual
  - stored on **chromosomes**
- most organisms have **two copies** of each chromosome
  - one from each parent
  - these two copies need **not be identical**
  - the genetic make-up of one chromosome is called a **haplotype**
- the genotype has the **combined information** of the two haplotypes in some kind of "**mixed format**"
- determining the haplotypes in a laboratory is **much harder** than determining the genotype

Genotype versus haplotypes

-
  - information in the genotype can be one of **0**, **1**, **2**
  - for the haplotype it can be one of **0**, **1**
- interpretation :
  - — **0** — $\longrightarrow$ — **0** — / — **0** —
  - — **1** — $\longrightarrow$ — **1** — / — **1** —
  - — **2** — $\longrightarrow$ — **0** — / — **1** —

Main question

- — **2** — **2** — $\longrightarrow$ — **0** — **0** — / — **1** — **1** — **or** — **0** — **1** — / — **1** — **0** — **?**

Haplotype inferring

- **Input** : the different **variants of genotypes** appearing in some population

  Output : for each genotype : two corresponding **haplotypes**

- satisfying some **rules** (usually biologically inspired)

  - **e.g.**: the fewer variants of haplotypes required, the better

    $\{$ **02**, **20**, **22** $\}$ could be $\left\{ \begin{matrix} \mathbf{00} \\ \mathbf{01} \end{matrix}, \begin{matrix} \mathbf{00} \\ \mathbf{10} \end{matrix}, \begin{matrix} \mathbf{00} \\ \mathbf{11} \end{matrix} \right\}$

    or $\left\{ \begin{matrix} \mathbf{00} \\ \mathbf{01} \end{matrix}, \begin{matrix} \mathbf{00} \\ \mathbf{10} \end{matrix}, \begin{matrix} \mathbf{01} \\ \mathbf{10} \end{matrix} \right\}$

    2nd choice preferable (only three different haplotypes)

Rules from genetic assumptions

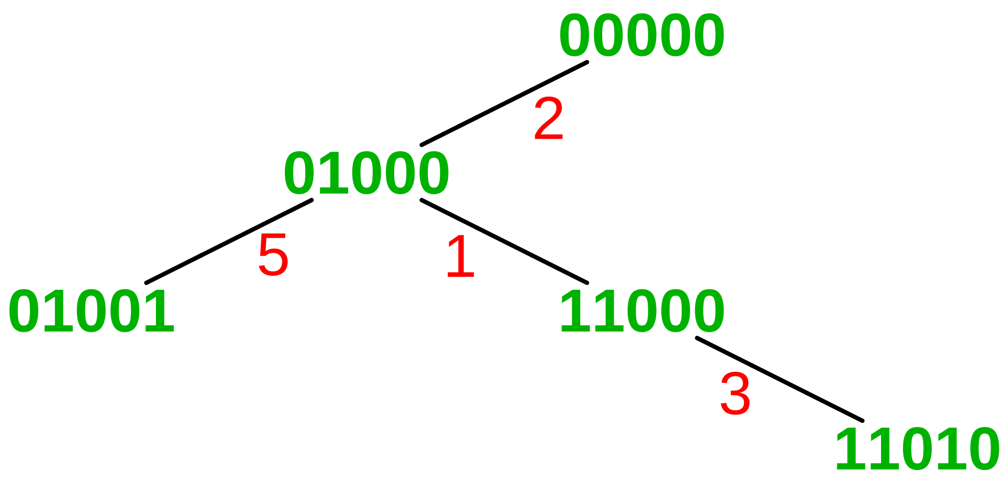
- initial assumptions :
  - **all** haplotypes originate from **one original** one
  - **variants** appeared because of **mutations**

    but only **small chance** that a mutation occurs at a particular place on a chromosome

- so try to obtain a collection of haplotypes requiring **minimum number of mutations**

An even stronger genetic requirement

- the proposed collection of haplotypes should should fit in a **Perfect Phylogenetic Tree** :
  - rooted tree
  - - each node labelled with a proposed haplotype
    - each haplotype appears as the label of one node
  - - each edge labelled by sites in which the ends differ
    - each site occurs at most once as an edge label



Perfect Phylogeny Haplotype Problem

- **Input** : the different **variants of genotypes** appearing in some population

  Question : is there a collection of haplotypes such that

  - each genotype has two corresponding haplotypes,
  - the haplotypes fit in a Perfect Phylogenetic Tree ?

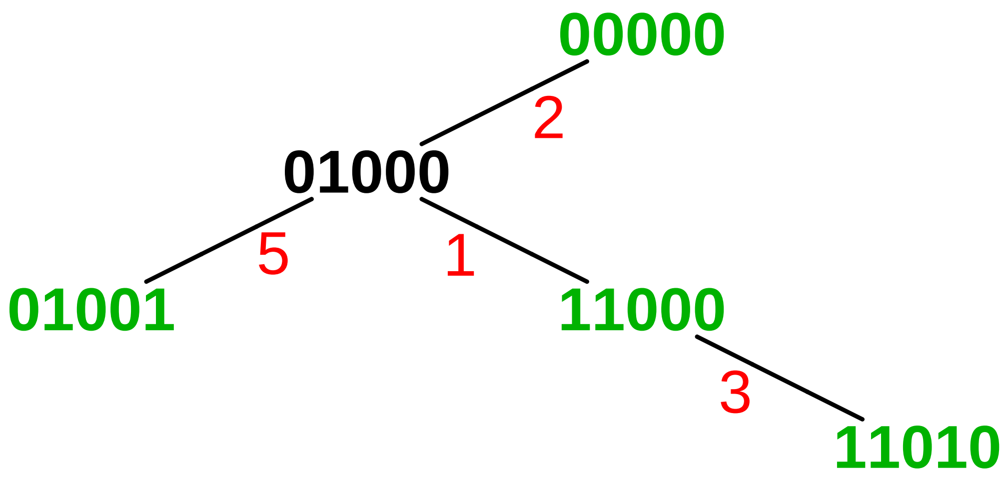
- **Theorem** (Gusfield, 2002)

  this can be solved in **polynomial time**

  (in number and length of the genotypes)

A more difficult problem

- **Imperfect Phylogenetic Tree** :
  some nodes do not appear in the original data
- example : suppose proposed haplotypes are
  00000, **01001**, **11000**, **11010**
  - to explain as a phylogenetic tree, the data needs an extra (”non-observed”) node :



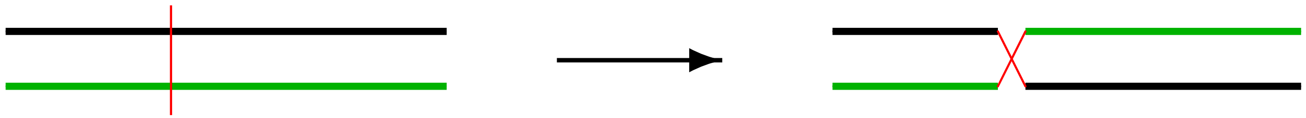
A more difficult problem

- **Imperfect Phylogenetic Tree** :
  some nodes do not appear in the original data

- <u>**Theorem** (Kimmel & Shamir, 2005)</u>
  the **Imperfect Phylogeny Haplotype Problem** is **NP-complete**

Further complications

- new haplotypes can occur because of **recombination** :



  - can be modelled with **cycles** in the phylogenetic tree
  - is assumed to happen very rarely

- leads to more and more
  combinatorial (decision) problems

From now on, mathematics only

an **ordered colouring** problem

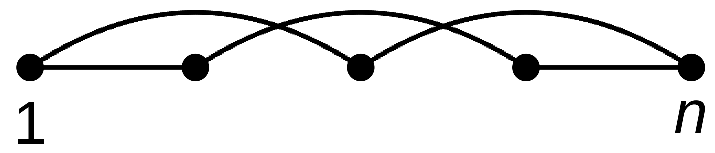
- **given**: graph $G = (V, E)$ with n vertices
  vertices are numbered from $1, \ldots, n$
- **required**: a vertex colouring with colours $\{A, B, C\}$
  i.e., a function $\varphi : V \to \{A, B, C\}$

such that:

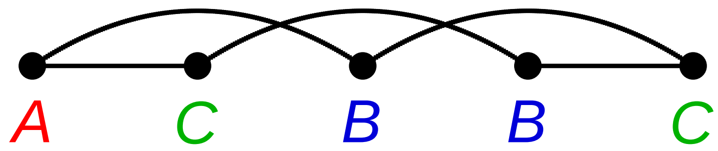
- $\forall uv \in E \ : \ \varphi(u) \neq \varphi(v)$ (i.e., a proper colouring)
- vertices coloured A have a **smaller number** than vertices coloured B

An example

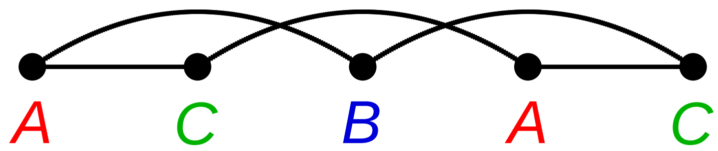
- suppose the graph is :



  - allowed colouring :



  - not allowed :



- **decision problem** :
  given a graph G with vertices $1, \ldots, n$,
  how easy is it to decide if it allows such a 3-colouring ?
  - **note** : normal 3-colouring is NP-complete

Deciding ordered 3-colouring

Theorem

deciding if a given graph allows this kind of 3-colouring can be done in **polynomial time**

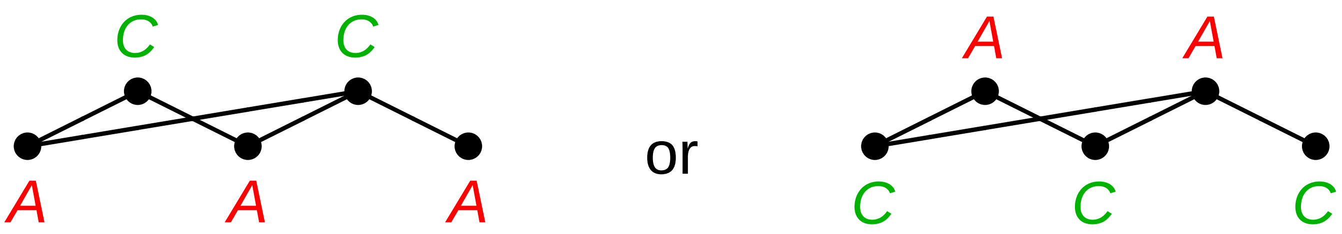
sketch of proof

- choose a vertex m from $1, \ldots, n$ (n choices)
- assume m forms the boundary between
  the vertices that can have colour A ($\leq m$)
  and the vertices that can have colour B ($> m$)
- then vertices $1, \ldots, m$ can only be coloured A or C
  and vertices $m+1, \ldots, n$ can only be coloured B or C
  - easy to check if such colourings exist (for chosen m)

Deciding ordered 3-colouring

sketch of proof (cont.)

- suppose the chosen m allows the required 2-colourings
- so vertices $1, \ldots, m$ form 2-colourable **components**
- similar for the vertices $m + 1, \ldots, n$
- each component has 2 different choices for a 2-colouring

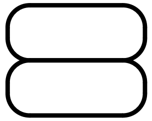


- but edges between vertices in $1, \ldots, m$ and vertices in $m + 1, \ldots, n$ restrict the options
  - not both ends can have colour C

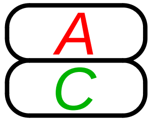
Deciding ordered 3-colouring

sketch of proof (cont.)

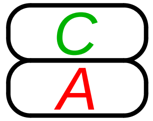
- for each component X_i in $1, \dots, m$
  - choose a **top** and a **bottom** part :



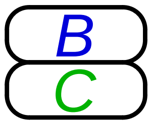
  - introduce a **Boolean variable** x_i with the meaning
    - x_i = TRUE :



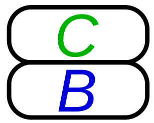
x_i = FALSE :



- for each component Y_j in $m + 1, \dots, n$
  - choose a **top** and a **bottom** part
  - introduce a **Boolean variable** y_j with the meaning
    - y_j = TRUE :

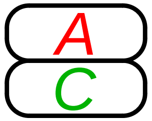


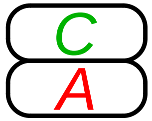
y_j = FALSE :

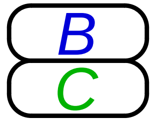


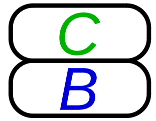
Deciding ordered 3-colouring

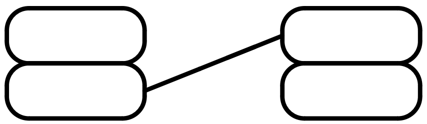
sketch of proof (cont.)

- x_i : 

$\overline{x_i}$: 

y_j : 

$\overline{y_j}$: 

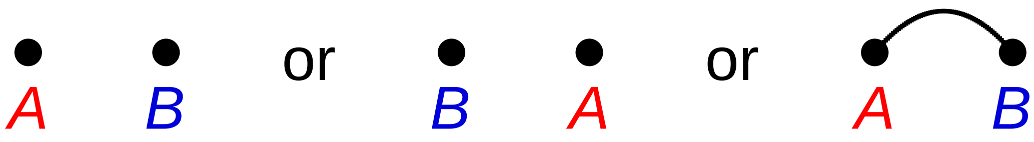
- each edge between a component X_i and a component Y_j gives rise to a **2-literal disjunction**
  - **e.g.** :  $\longleftrightarrow$ $\overline{x_i} \vee y_j$

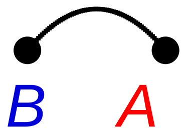
- combining the clauses for **all edges** between components X_i and Y_j gives a **conjunction of 2-literal disjunctions**
  - i.e., an instance of the **2-SATISFIABILITY** problem
  - well known to be decidable in **polynomial time**

A different ordered 3-colouring problem

- **given**: graph $G = (V, E)$ with n vertices
  vertices are numbered from $1, \ldots, n$
- **required**: a vertex colouring with colours $\{A, B, C\}$
  i.e., a function $\varphi : V \to \{A, B, C\}$

such that:

- $\forall uv \in E \, : \, \varphi(u) \neq \varphi(v)$ (i.e., a proper colouring)
- condition that vertices coloured A must have a **smaller number** than vertices coloured B only **has to hold for adjacent vertices**
  - allowed:
    

  - not allowed:
    

A different ordered 3-colouring problem

- **given**: graph $G = (V, E)$ with n vertices
  vertices are numbered from $1, \ldots, n$
- **required**: a vertex colouring with colours $\{A, B, C\}$
  i.e., a function $\varphi : V \to \{A, B, C\}$

  such that:
  - $\forall uv \in E \; : \; \varphi(u) \neq \varphi(v)$ (i.e., a proper colouring)
  - condition that vertices coloured A must have a **smaller number** than vertices coloured B only **has to hold for adjacent vertices**
- **Theorem**
  deciding if for a given graph such a colouring is possible is **NP-complete**

Generalising the problem (even less applied)

- a way to look at these colouring problems is as if there is a **order relation** on the colours $\{ A, B, C \}$

$$\begin{array}{ccc} & & B \\ C & & \uparrow \\ & & A \end{array} \qquad \longleftrightarrow$$

"A before B"
"A - C and B - C are indifferent"

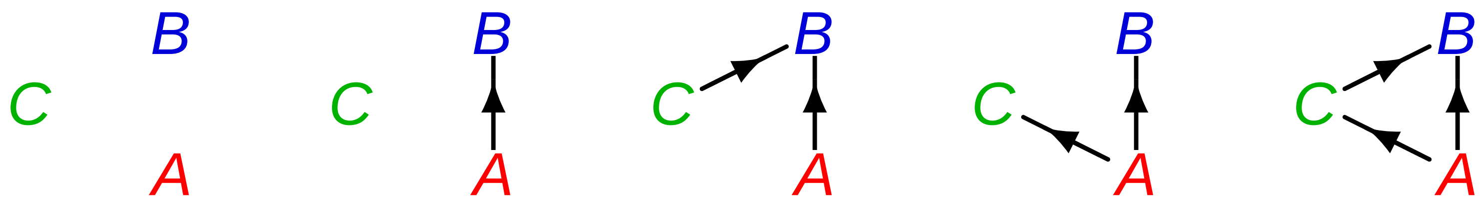
- what would happen if we consider other order relations on the colours ?

More general ordered 3-colourings

- **first variant** :

  $A \rightarrow B$ means "any A before any B"

- possible posets with 3 elements :



- **deciding** if a certain graph with vertices $1, \ldots, n$ has a colouring according to these posets is
  - **NP-complete** for the first poset (just 3-colouring)
  - **polynomial** for the others

More general ordered 3-colourings with more colours

- $A \longrightarrow B$ means "any A before any B"
- suppose we allow **any number** of colours and **any poset** on the set of colours
- **<u>Theorem</u>**

  deciding if a given graph with vertices $1, \ldots, n$ has a colouring according to the fixed colour poset is

  - **NP-complete** if the poset has an anti-chain of length 3
  - **polynomial** otherwise
- similar (more complicated) **dichotomy** result also known for the other variant

  ($A \longrightarrow B$ means "A before B if adjacent")