Advanced Integer Linear Programming for Cancer Phylogenetics

Mohammed El-Kebir
UIUC

April 2, 2024

Combinatorial Optimization in Computational Biology

Many processes in biology are discrete and combinatorial in nature!

Combinatorial biological process

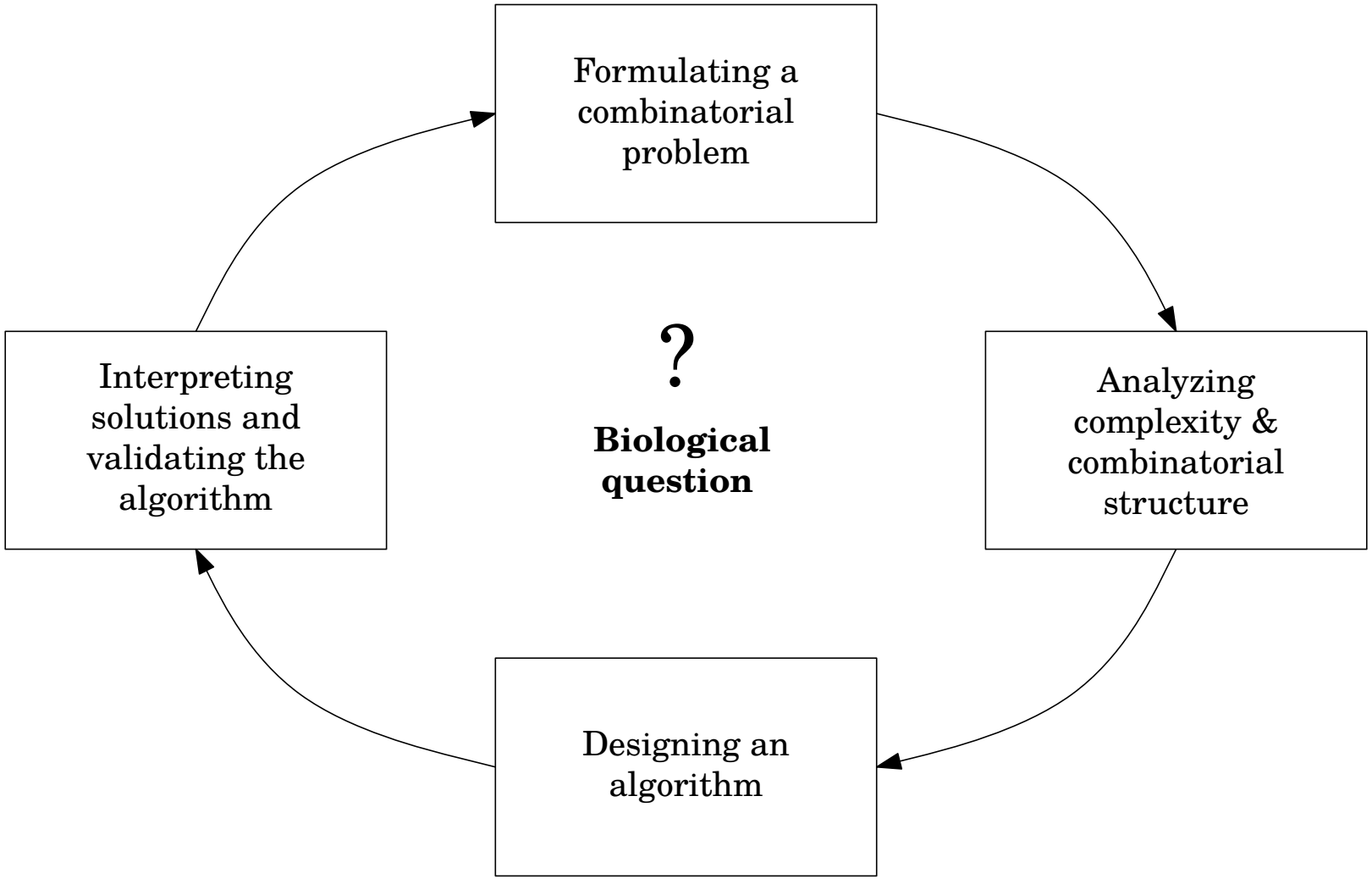
- Evolution: mutations accumulating in biological populations
- Different combinations of genetic elements result in varied gene expression
- Proteins may interact and form protein complexes
- …

Computational tasks

- Reconstructing evolutionary trees from measurements at present time
- Inferring a gene-expression network from RNA-seq data
- Comparing protein-protein interaction networks
- Assembling a genome from reads
- …

Goal: Inferring combinatorial objects from data subject to biological constraint --- finding the right problem is non-trivial!

Combinatorial Optimization in Computational Biology



Problem ≠ Algorithm

Different Types of Problems

Problem Π with instance/input X and feasible solution set $\Pi(X)$:

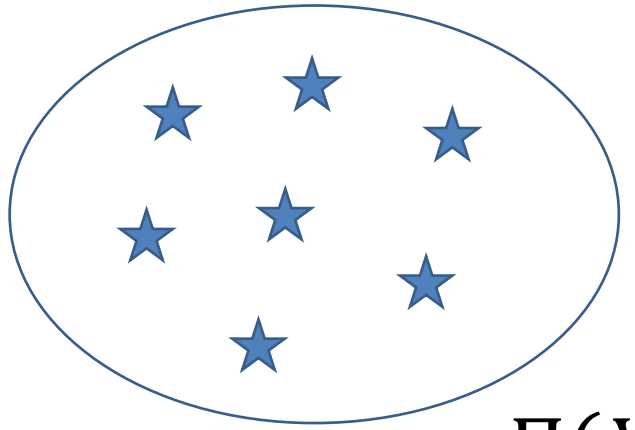
- Decision problem:
  - Is $\Pi(X) = \emptyset$?
- Optimization problem:
  - Find $y^* \in \Pi(X)$ s.t. $f(y^*)$ is optimum.
- Counting problem:
  - Compute $|\Pi(X)|$.
- Sampling problem:
  - Sample uniformly from $\Pi(X)$.
- Enumeration problem:
  - Exhaustively enumerate solutions $\Pi(X)$.

Algorithm:

Set of instructions for solving problem.

- Exact
- Heuristic

Running time: How does the number of steps scale as a function of $|X|$?

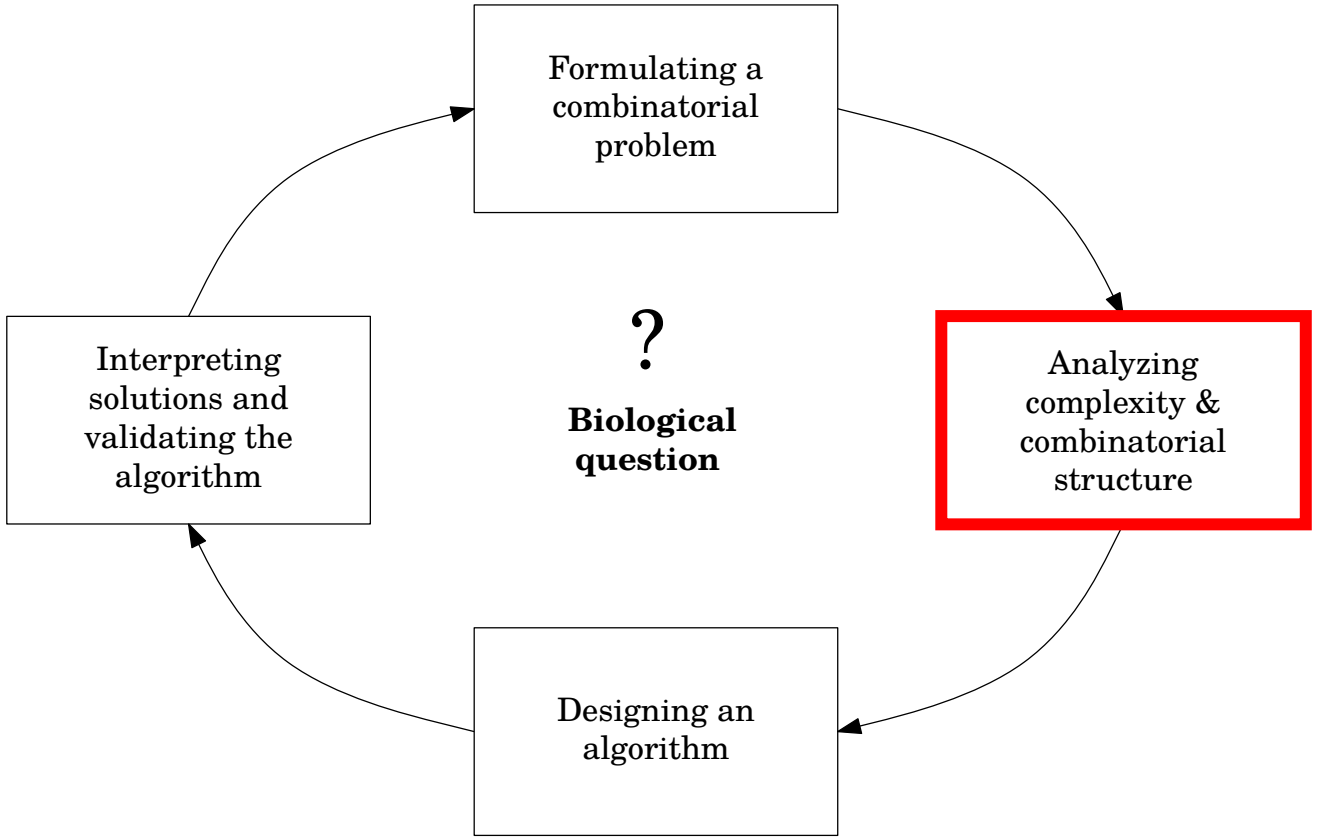


$\Pi(X)$

Many problems do not admit efficient algorithms

Hard Optimization Problems

Many problems do not admit efficient algorithms



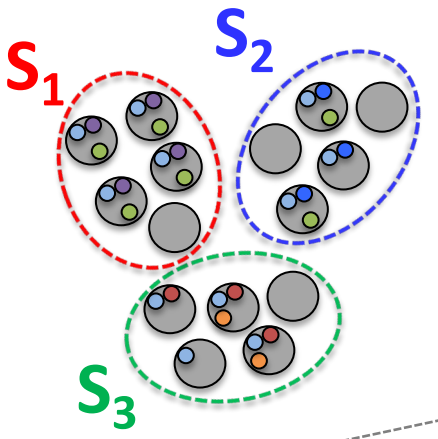
Informally:
NP-hard problems are optimization problems that are really hard, and unlikely solvable in polynomial time.

But often we can still exactly solve practical problem instances!

Outline – Advanced Integer Linear Programming Techniques

- Enumerating the solution space
- Piecewise linear approximation
- Cutting planes
- Column generation

Problem 1: Deconvolving Bulk DNA-seq Data



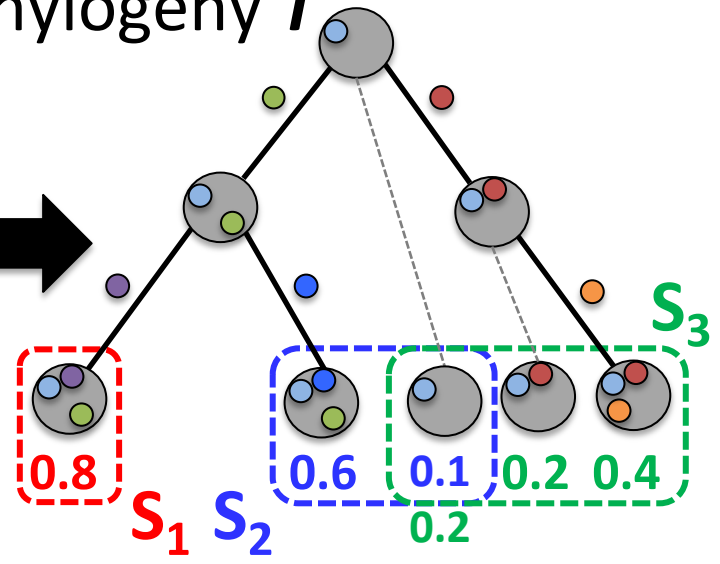
n mutations

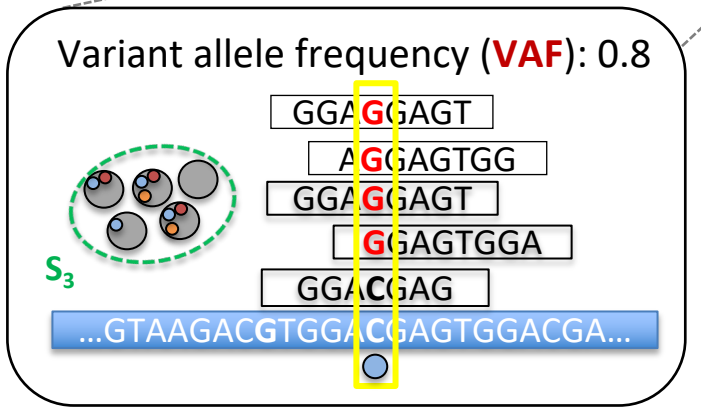
m samples

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ \boxed{0.8} & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

Frequency Matrix ***F***

Phylogeny ***T***





Input: Mixtures of unknown leaves ***L(T)***
of an unknown tree ***T*** in unknown proportions ***U***

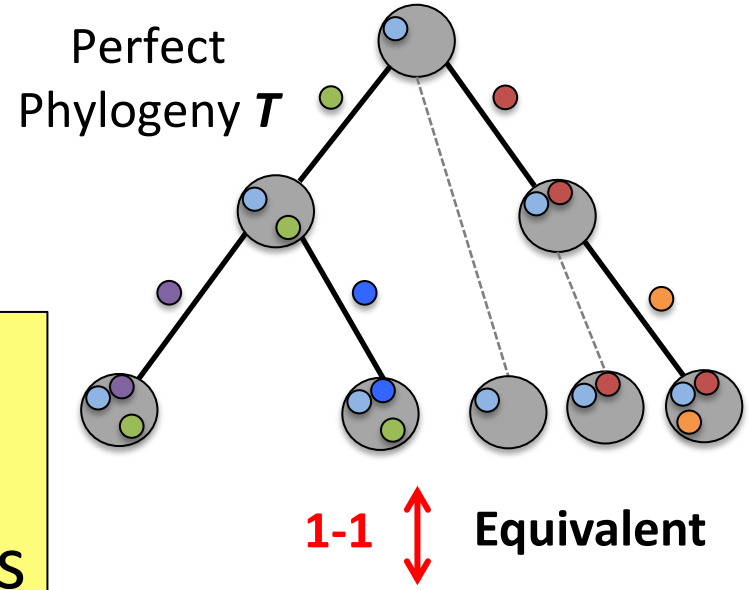
Biological question: How to infer an evolutionary tree ***T*** from frequencies ***F***

Problem 1: Perfect Phylogeny Mixture Deconvolution (PPMD)

Variant of PPMD:

TrAp [Strino *et al.*, 2013], PhyloSub [Jiao *et al.*, 2014],
CITUP [Malikic *et al.*, 2015], BitPhylogeny [Yuan *et al.*, 2015]
LICHeE [Popic *et al.*, 2015], AncesTree [El-Kebir, Oesper *et al.*, 2015], …

PPMD: [El-Kebir*, Oesper* et al., 2015]
Given ***F***, find ***U*** and ***B*** such that (i) ***F* = *U B***, (ii) ***B*** is a perfect phylogeny and (iii) ***U*** has nonnegative entries



n mutations

m samples

S_1	0.8	0.8	0.8	0.0	0.0	0.0
S_2	0.7	0.6	0.0	0.6	0.0	0.0
S_3	0.8	0.0	0.0	0.0	0.6	0.4

Frequency Matrix ***F***

=

clones

m samples

S_1	0.0	0.0	0.8	0.0	0.0	0.0
S_2	0.1	0.0	0.0	0.6	0.0	0.0
S_3	0.2	0.0	0.0	0.0	0.2	0.4

Mixture Matrix ***U***

1	0	0	0	0	0	
1	1	0	0	0	0	
1	1	1	0	0	0	
1	1	0	1	0	0	
1	0	0	0	1	0	
1	0	0	0	1	1	

clones

Perfect Phylogeny Matrix ***B***

Combinatorial Characterization of PPMD

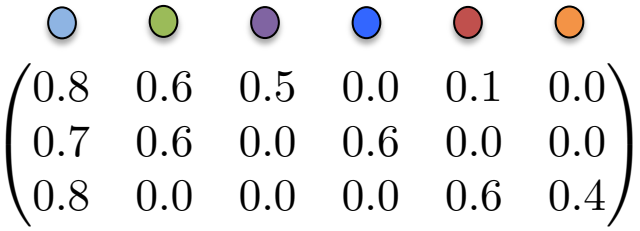
Given ***F***, find ***U*** and ***B*** such that (i) ***F = U B***, (ii) ***B*** is a PP tree and (iii) $\boldsymbol{U} \geqslant 0$

Lemma (Sum Condition):

Given ***F*** and ***T***, for all samples *p* and mutations *j*,

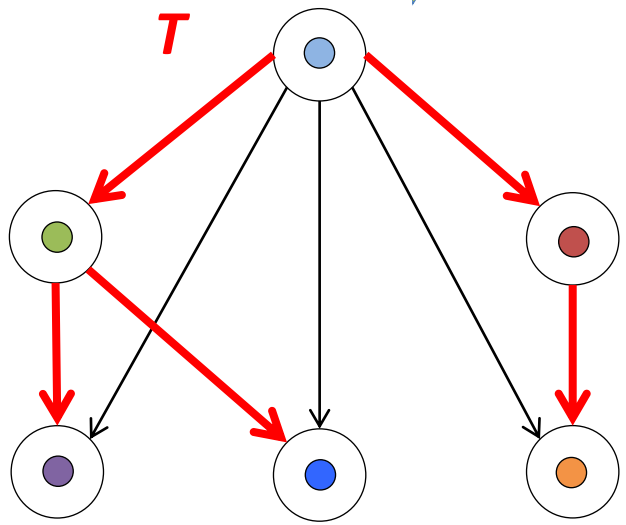
$$f_{pj} \geq \sum_{k \text{ child of } j} f_{pk}$$

necessary sufficient



$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F



Ancestry Graph ***G* = (*V*, *A*)**

Lemma (Ancestry Condition):

Given ***F*** and ***T***, for all samples *p* and mutations *k* child of *j*, $f_{pj} \geq f_{pk}$

necessary

Ancestry graph *G* = (*V*, *A*); given ***F***

- Vertex for every mutation
- Edge $(j, k) \in A$ if $f_{pj} \geq f_{pk}$ for all samples *p*

Theorem 1:

T is a solution to the PPM if and only if ***T*** is a spanning tree of ***G*** satisfying the Sum Condition

Theorem 2:

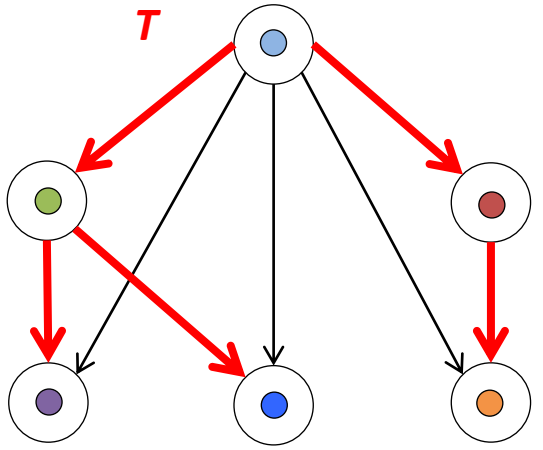
PPM is NP-complete even if *m* = 2

M. El-Kebir*, L. Oesper*, H. Acheson-Field and B. J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics (Special Issue: Proceedings of ISMB), 31(12):i62-i70, 2015.

ILP for PPMD

$$\begin{pmatrix} 0.8 & 0.6 & 0.5 & 0.0 & 0.1 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

F



Ancestry Graph ***G***

$$\max \sum_{(i,j)\in E(G)} x_{i,j}$$

Maximize the number of edges in the tree

$$\text{s.t.} \sum_{i=1}^{n} r_i = 1$$

Only a single mutation is the root

$$r_j + \sum_{(i,j)\in E(G)} x_{i,j} = 1 \qquad \forall j \in [n]$$

Every mutation has a parental mutation or is the root

$$\sum_{(i,j)\in E(G)} f_{p,j} \cdot x_{i,j} \leq f_{p,i} \qquad \forall p \in [m], i \in [n]$$

Sum condition

$$r_i \in \{0,1\} \qquad \forall i \in [n]$$

$$x_{i,j} \in \{0,1\} \qquad \forall (i,j) \in E(G)$$

Integrality

$O(|E(G)|)$ binary variables, $O(mn + |E(G)|)$ constraints

Maximum Likelihood for PPMD

n mutations

m samples

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{pmatrix} 2 & 4 & 4 & 0 & 0 & 0 \\ 7 & 6 & 0 & 3 & 0 & 0 \\ 8 & 0 & 0 & 0 & 3 & 2 \end{pmatrix}$$

Variant Read Counts ***A***

n mutations

m samples

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix}$$

Frequency Matrix ***F***

n mutations

m samples

$$\begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{pmatrix} 5 & 5 & 5 & 10 & 10 & 5 \\ 10 & 10 & 5 & 5 & 5 & 5 \\ 10 & 5 & 5 & 2 & 5 & 5 \end{pmatrix}$$

Total Read Counts ***D***

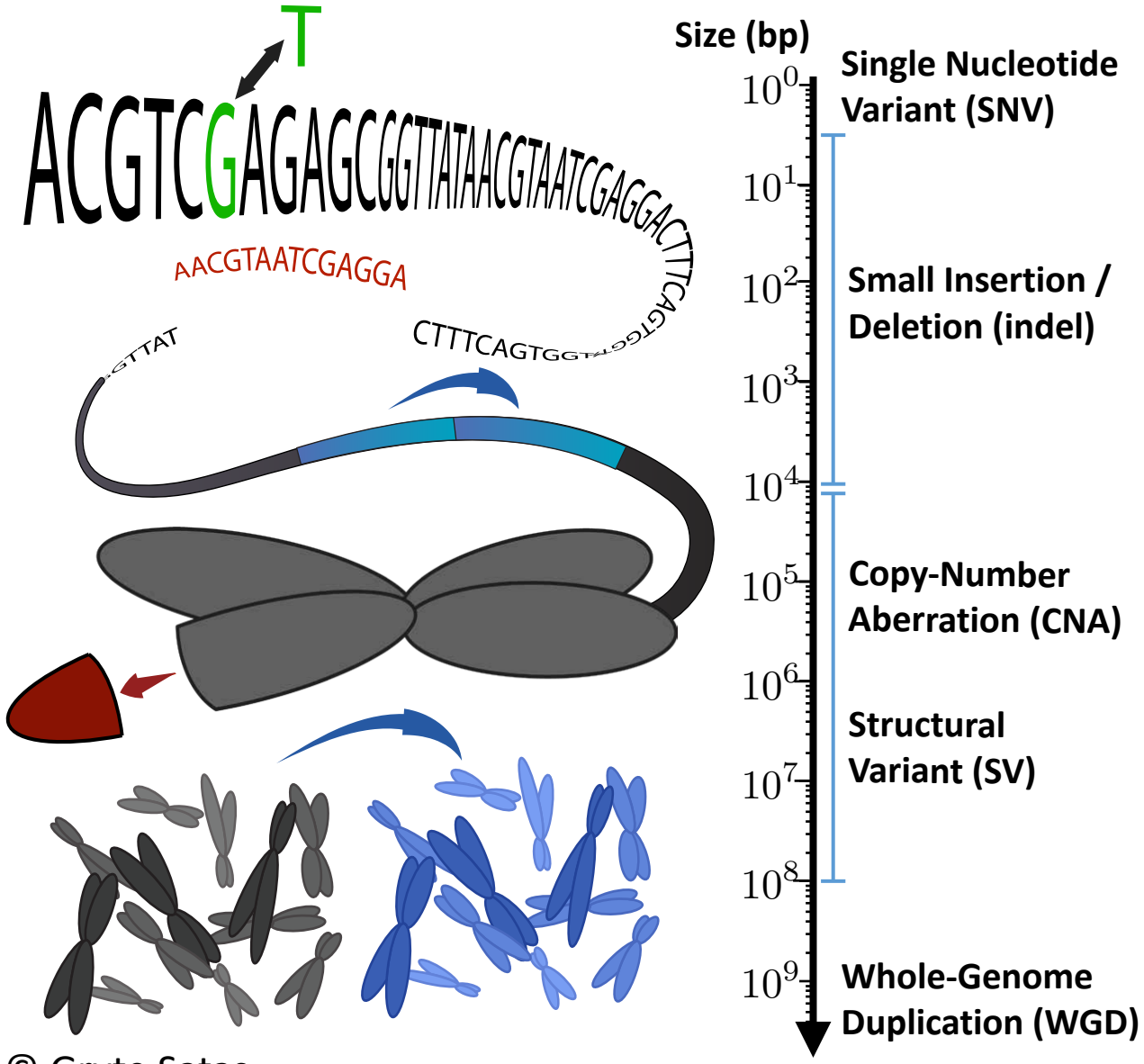
Given ***A, D***, find ***F, U*** and ***B*** such that (i) ***F = U B***, (ii) ***B*** is a PP tree, (iii) $\boldsymbol{U} \geqslant 0$ and (iv) $\Pr(\boldsymbol{A} \mid \boldsymbol{D}, \boldsymbol{F})$ is maximum

$$\Pr(A \mid D, F) = \prod_{p=1}^{m} \prod_{i=1}^{n} \text{binom}(a_{p,i} \mid d_{p,i}, f_{p,i}) = \prod_{p=1}^{m} \prod_{i=1}^{n} \binom{d_{p,i}}{a_{p,i}} (f_{p,i})^{a_{p,i}} (1 - f_{p,i})^{d_{p,i} - a_{p,i}}.$$

Outline – Advanced Integer Linear Programming Techniques

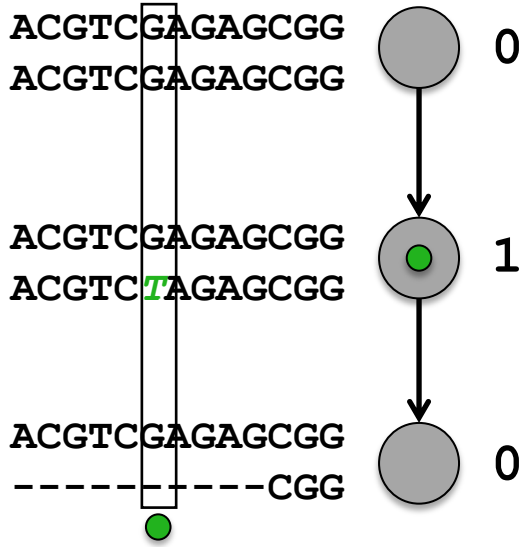
- Enumerating the solution space
- Piecewise linear approximation
- Cutting planes
- Column generation

Infinite Sites Assumption is too Restrictive for SNVs



© Gryte Satas

SNVs can be **lost** due to CNAs



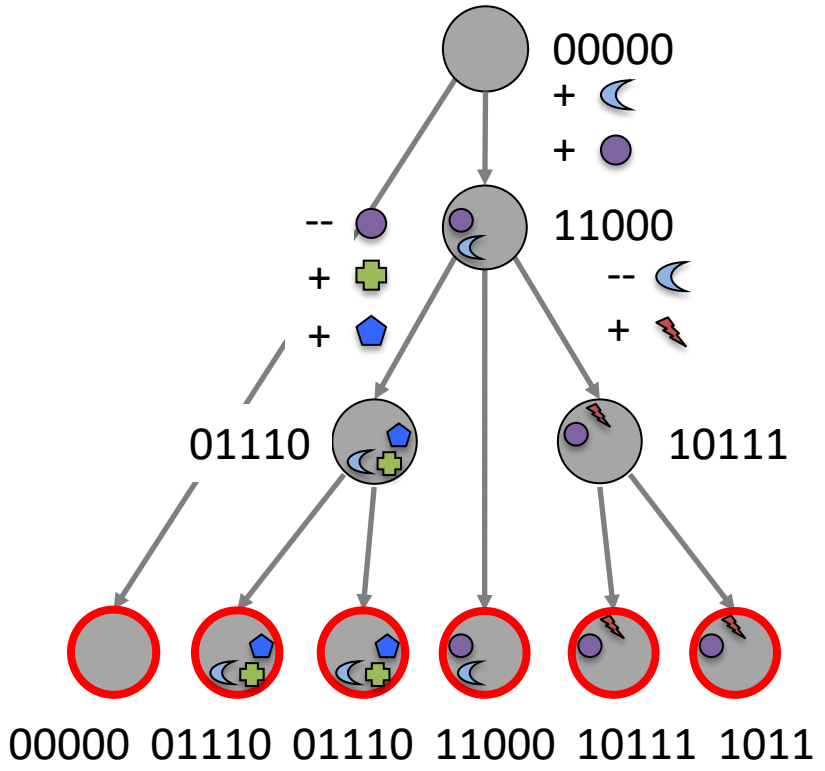
Infinite sites assumption:

- No parallel evolution of SNVs
- No loss of SNVs
- SCITE [Jahn et al. 2016]
- OncoNEM [Ross and Markowetz, 2016]

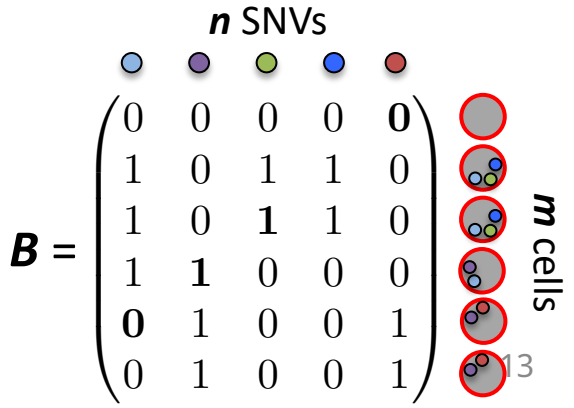
k-Dollo Phylogeny (*k*-DP) Problem

Definition 1. A *k-Dollo phylogeny* T is a rooted, node-labeled tree subject to the following conditions.

1. Each node v of T is labeled by a vector $\mathbf{b}_v \in \{0,1\}^n$.
2. The root r of T is labeled by vector $\mathbf{b}_r = [0,\ldots,0]^T$.
3. For each character $c \in [n]$, there is exactly one *gain edge* (v,w) in T such that $b_{v,c} = 0$ and $b_{w,c} = 1$.
4. For each character $c \in [n]$, there are at most k *loss edges* (v,w) in T such that $b_{v,c} = 1$ and $b_{w,c} = 0$.



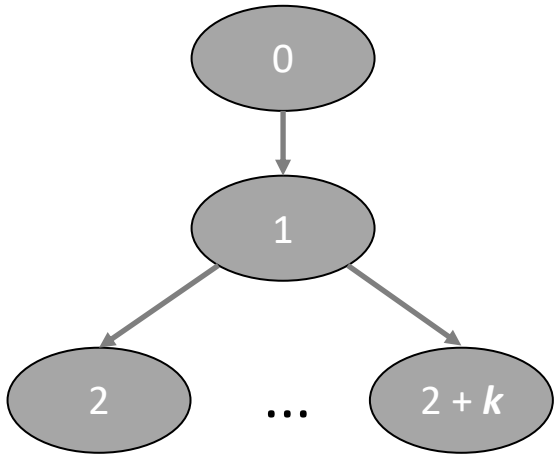
***k*-Dollo Phylogeny problem** (*k*-DP). Given a binary matrix $B \in \{0,1\}^{m\times n}$ and parameter $k \in \mathbb{N}$, determine whether there exists a k-Dollo phylogeny for B, and if so construct one.



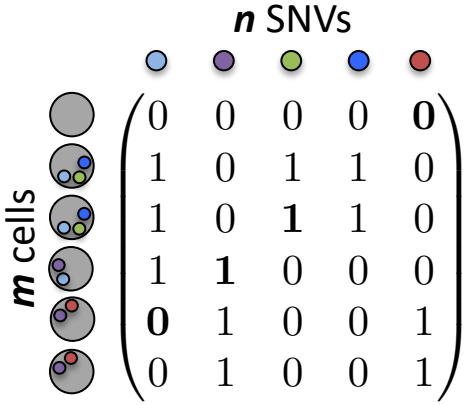
Combinatorial Characterization of *k*-DP

Theorem 3. Let $B \in \{0,1\}^{m \times n}$. The following statements are equivalent.

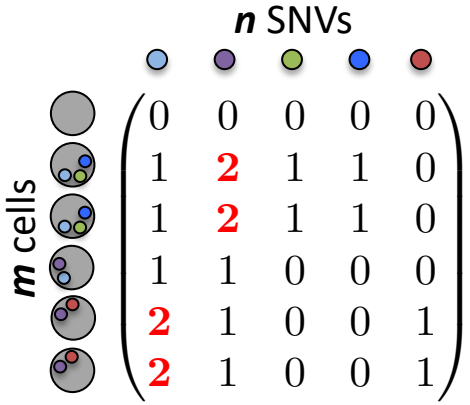
1. There exists a k-Dollo phylogeny T for B.
2. There exists a k-Dollo completion A of B.
3. There exists a k-completion A of B, and perfect phylogeny T for A whose characters are consistent with $S[k]$.



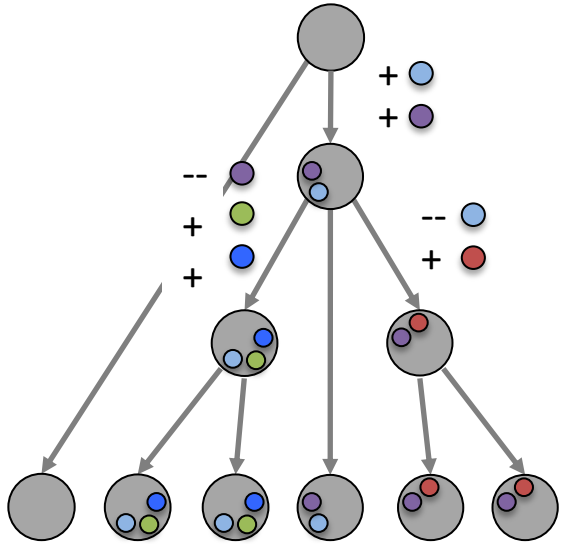
***k*-Dollo State Tree *S*[*k*]**



Input Matrix ***B***



k*-Dollo Completion *A



k*-Dollo Phylogeny *T

Forbidden Submatrices in Solutions A to k-DP

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

$k = 0$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 2 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & 2 \\ 1 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & 2 \\ 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 2 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & 2 \\ 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 0 & 2 \\ 2 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 \\ 0 & 2 \\ 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ 0 & 2 \\ 2 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 0 & 2 \\ 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 0 & 2 \\ 2 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 1 & 2 \\ 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ 1 & 2 \\ 2 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \\ 2 & 2 \end{pmatrix}$$

$k = 1$

Naïve ILP

$$\min \quad \sum_{p=1}^{m}\sum_{c=1}^{n} a_{p,c,2}$$

Minimize the number of losses

$$\text{s.t.} \quad a_{p,c,1} = b_{p,c} \qquad \forall p \in [m], c \in [n]$$

1-state in input <-> 1-state in output

$$a_{p,c,1} + a_{p,d,0} + a_{q,c,0} + a_{q,d,1} + a_{r,c,1} + a_{r,d,1} \leq 5 \qquad \forall p, q, r \in [m], c, d \in [n]$$

1st forbidden submatrix [[1,0], [0,1] , [1,1]]

$$\vdots$$

25th forbidden submatrix [[2,1], [1,2] , [2,2]]

$$a_{p,c,2} + a_{p,d,1} + a_{q,c,1} + a_{q,d,2} + a_{r,c,2} + a_{r,d,2} \leq 5 \qquad \forall p, q, r \in [m], c, d \in [n]$$

$$a_{p,c,i} \in \{0, 1\} \qquad \forall p \in [m], c \in [n], i \in \{0, \ldots, 2\}$$

$O(mn)$ binary variables, $O(m^3n^2)$ constraints

Callbacks

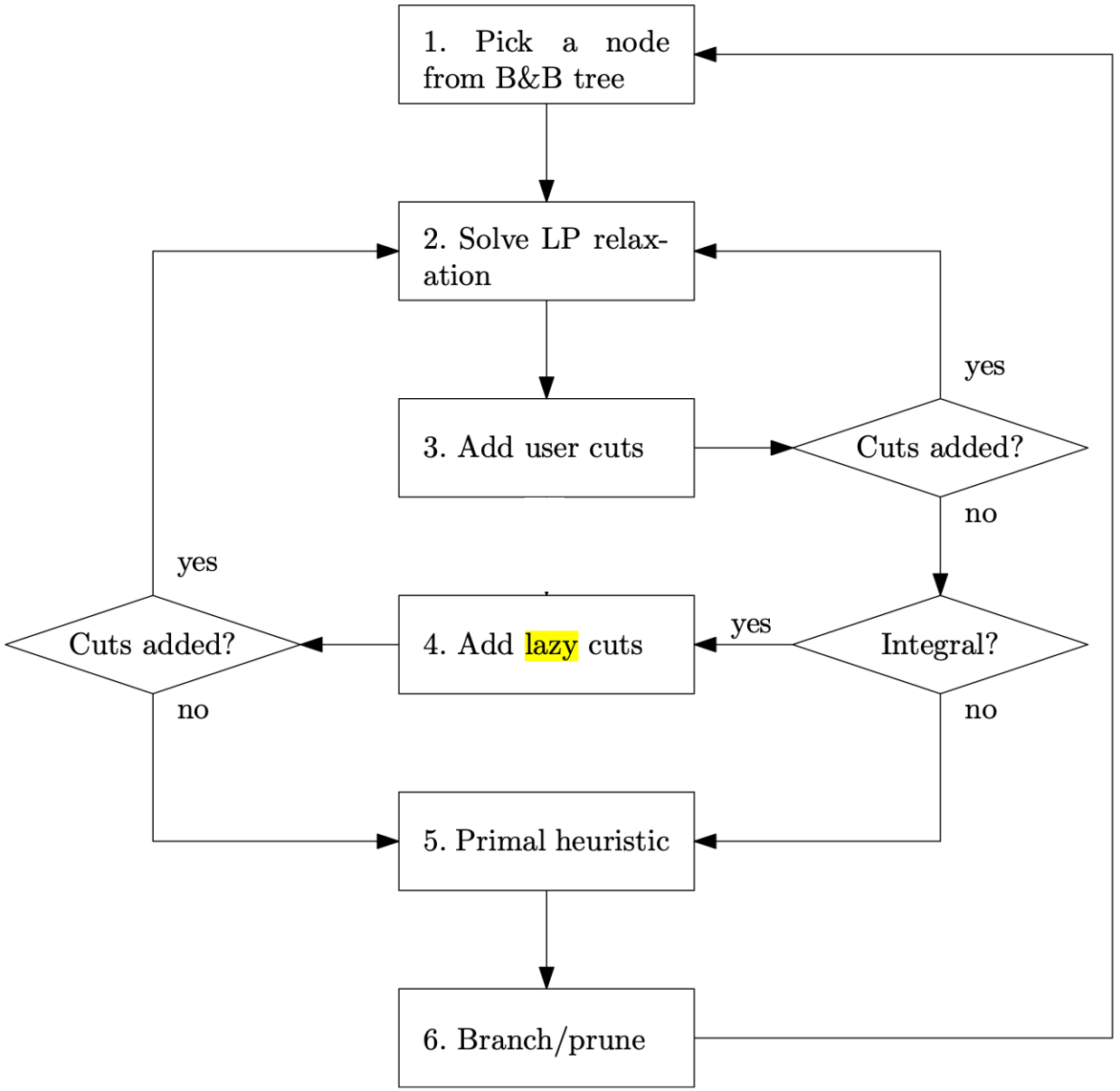
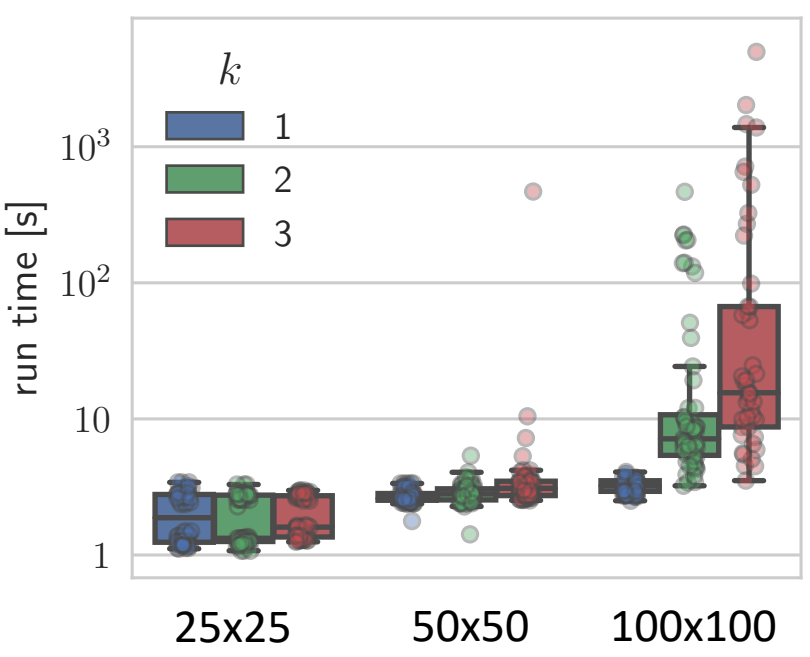
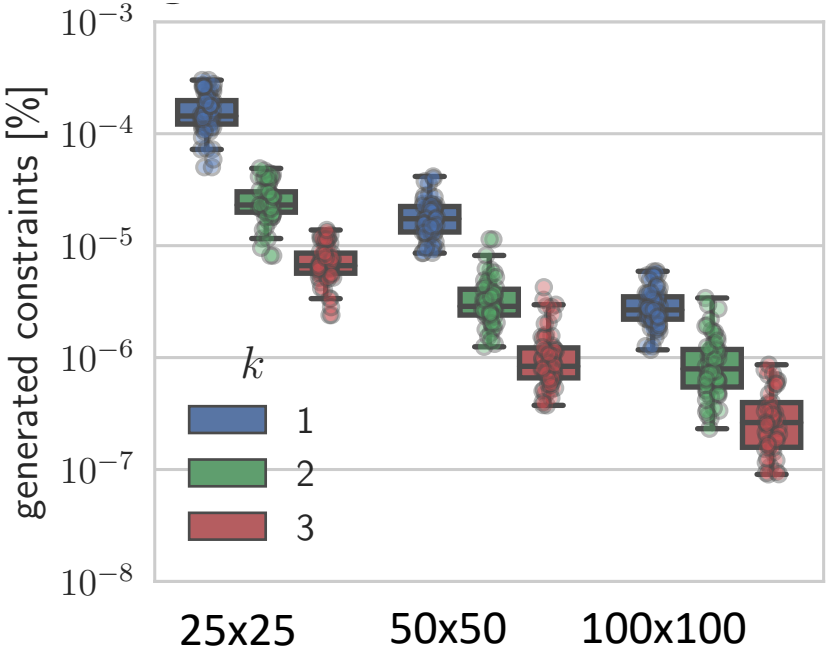


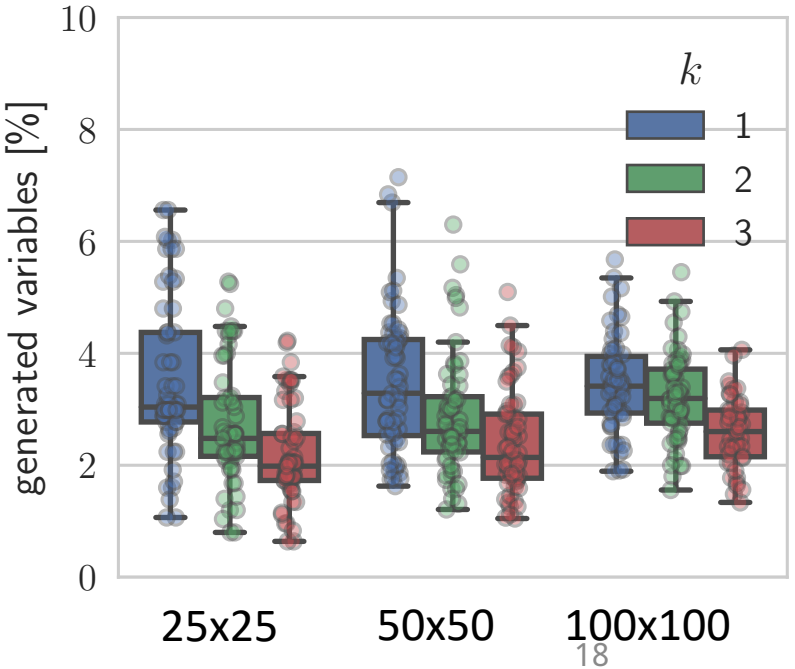
Figure 5.5: CPLEX flow diagram.

Results for *k*-DP

- Naive ILP does not scale and has O(*mnk*) variables and O($m^3n^2k^4$) constraints
- Column and cutting plane generation
  - Introduce variables and constraints only when needed
- Simulations with 60 instances for each each m, n and k







Advanced Integer Linear Programming Techniques

- Enumerating the solution space
- Piecewise linear approximation
- Cutting planes
- Column generation