COLLÈGE
DE FRANCE
1530

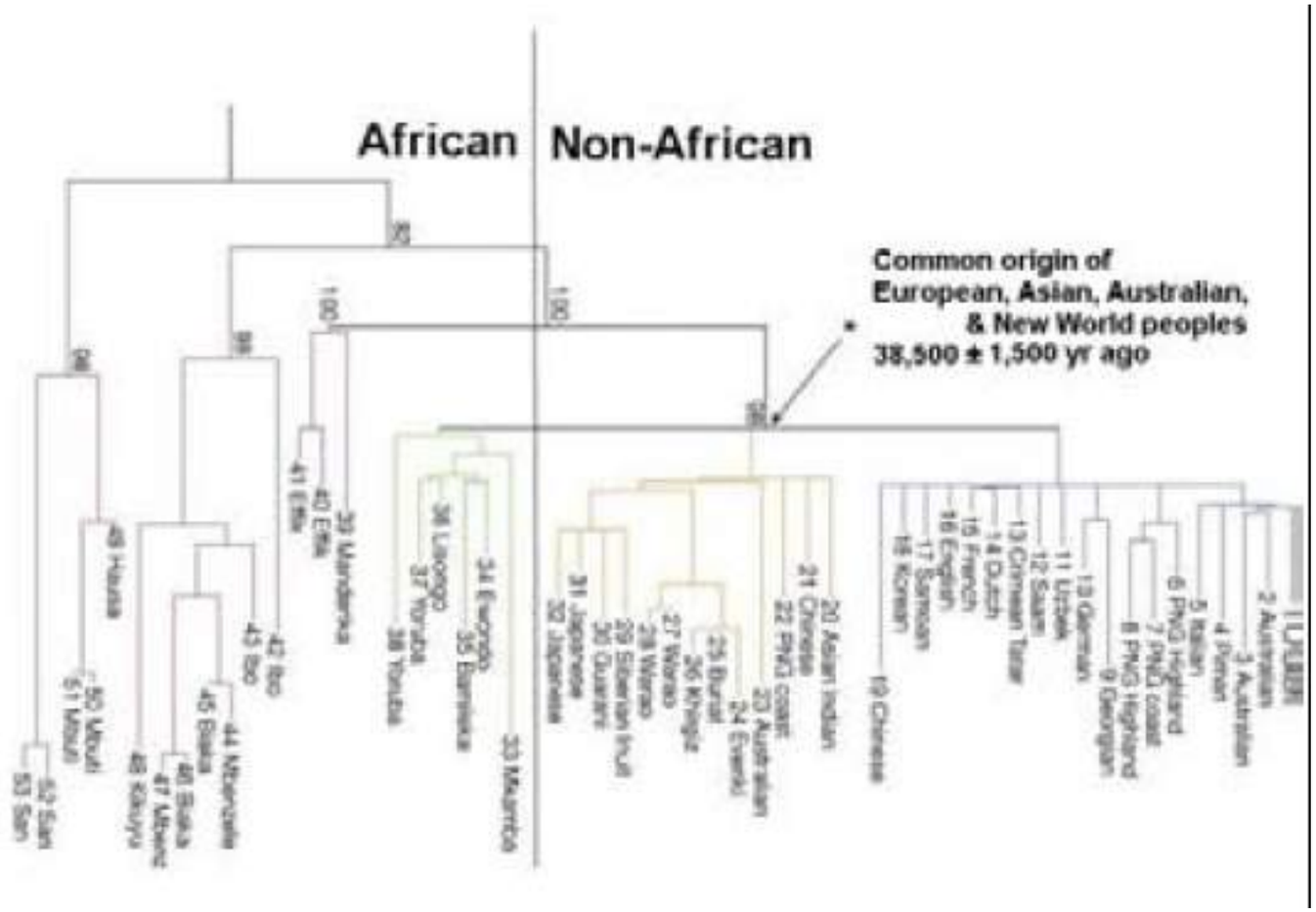
Algorithmes : conclusion

Claire Mathieu

Reconstruction d’arbre à partir d’ADN
Arbre de Steiner et méthode primale-dual
Conclusion

Reconstruction d’arbre à partir d’ADN

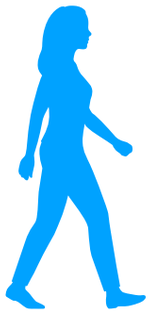
Arbre phylogénétique



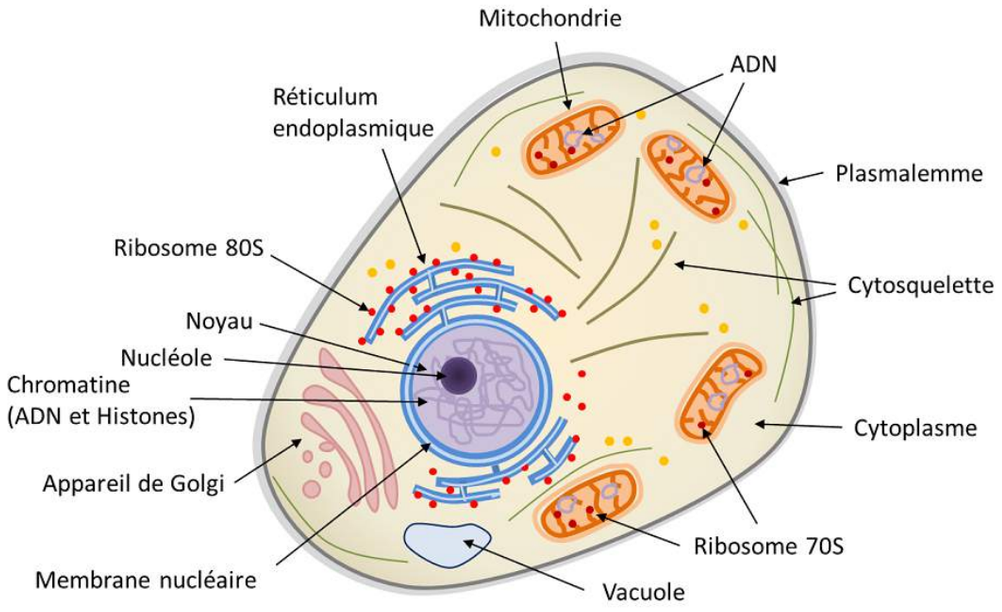
Reconstruction basée sur l'ADN :
l'ancêtre commun à tous les hommes modernes
a vécu en Afrique il y a environ 200 000 ans

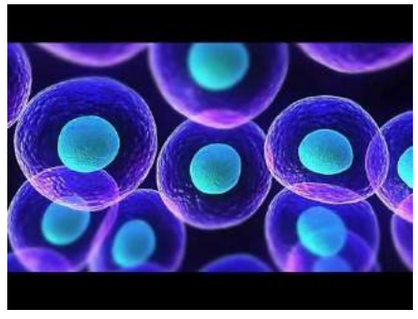
L’homme vivant dont l’ADN est probablement le plus ancien de l'humanité

L’ADN pour reconstruire l’évolution



Personne





Cellules

noyau

22 paires de chromosomes autosomes

ADN

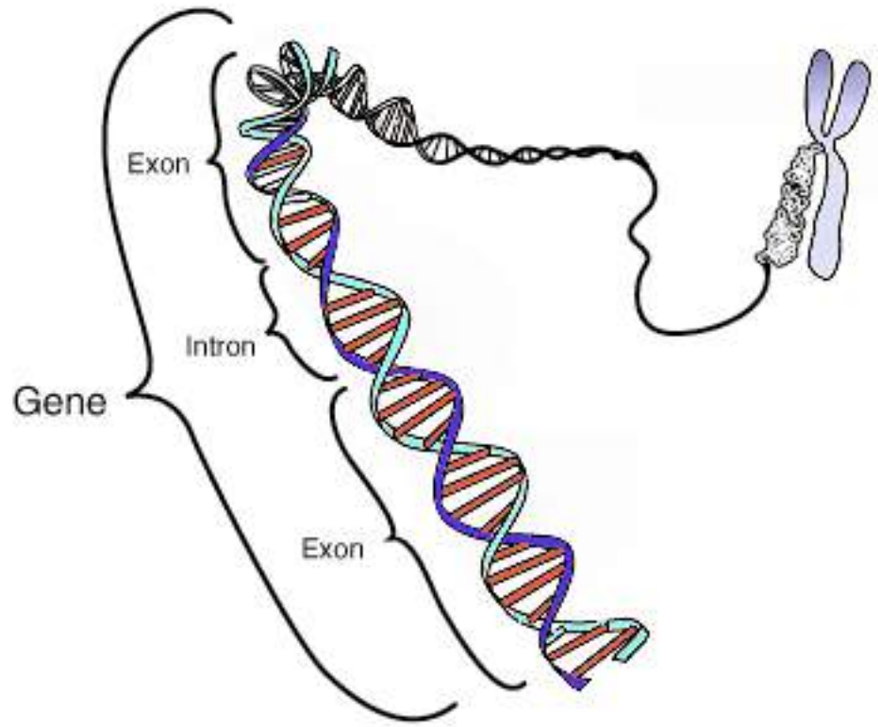
1 paire de chromosomes XY (homme) ou XX (femme)

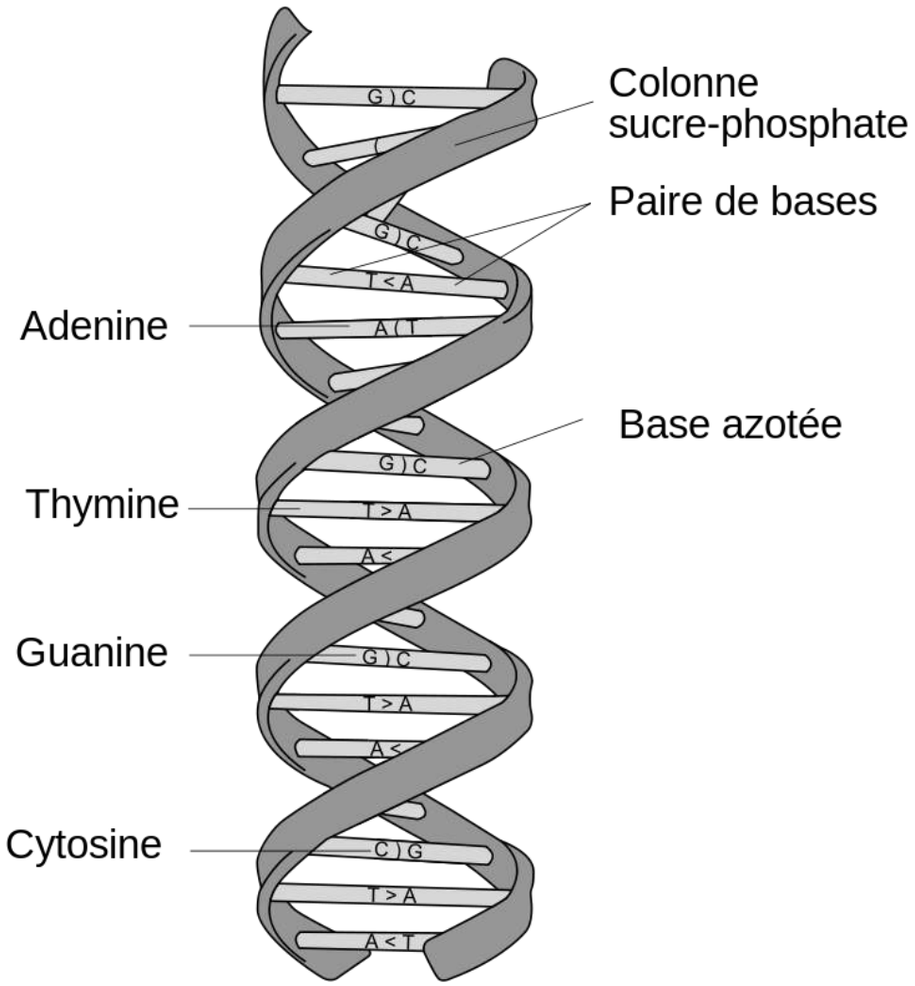
ADN

mitochondries

ADN

Le chromosome a 2 brins d’ADN.
La mitochondrie aussi.
Gènes : fragments d’ADN

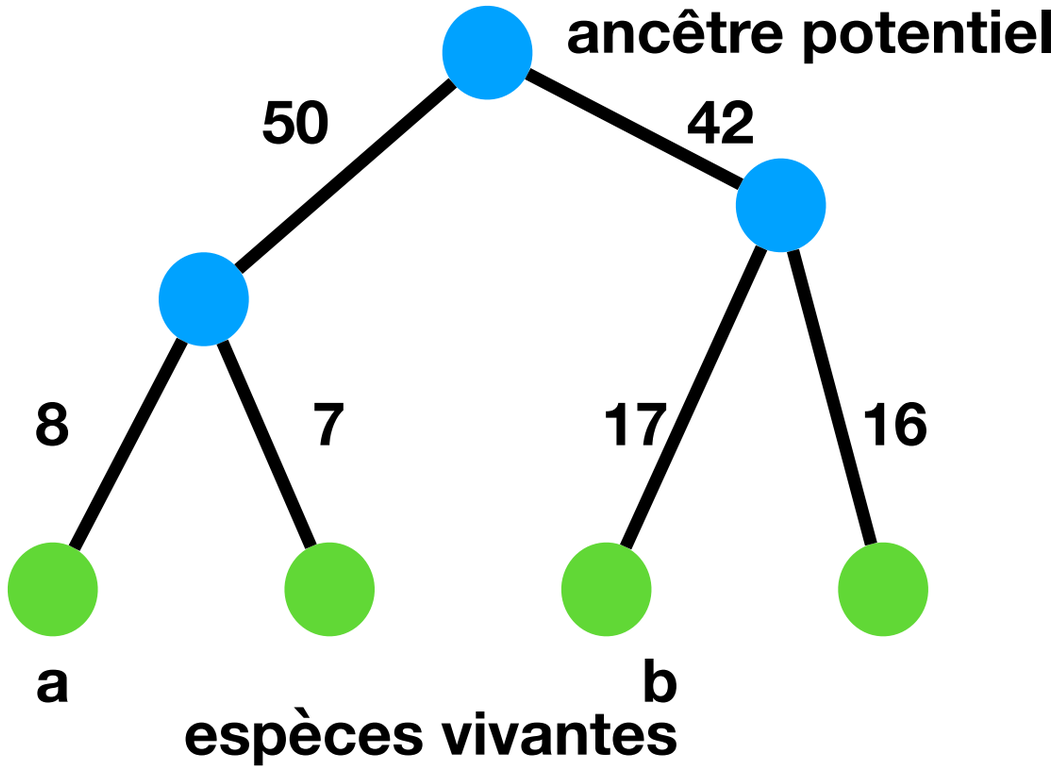


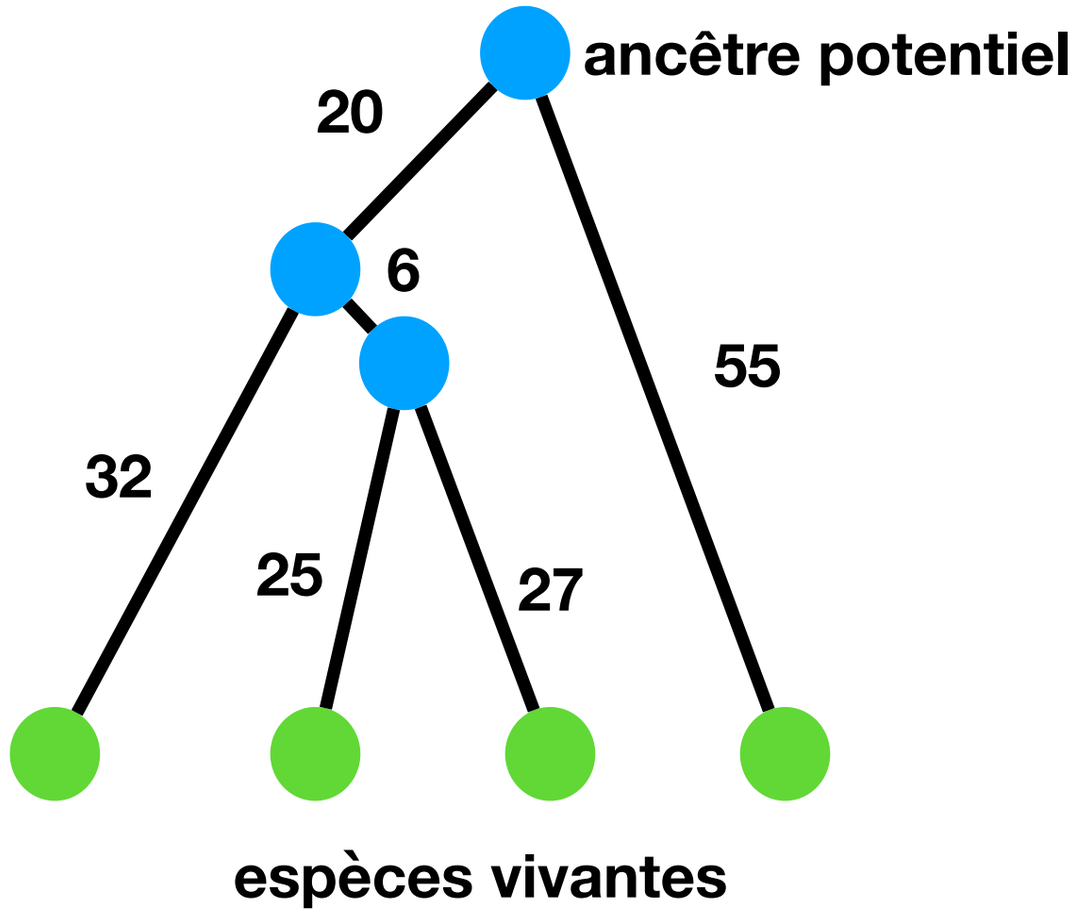


Mot écrit avec les lettres A,C,G,T
(lettre = base azotée du nucléotide)

Évolution : mutations
Modifications du mot

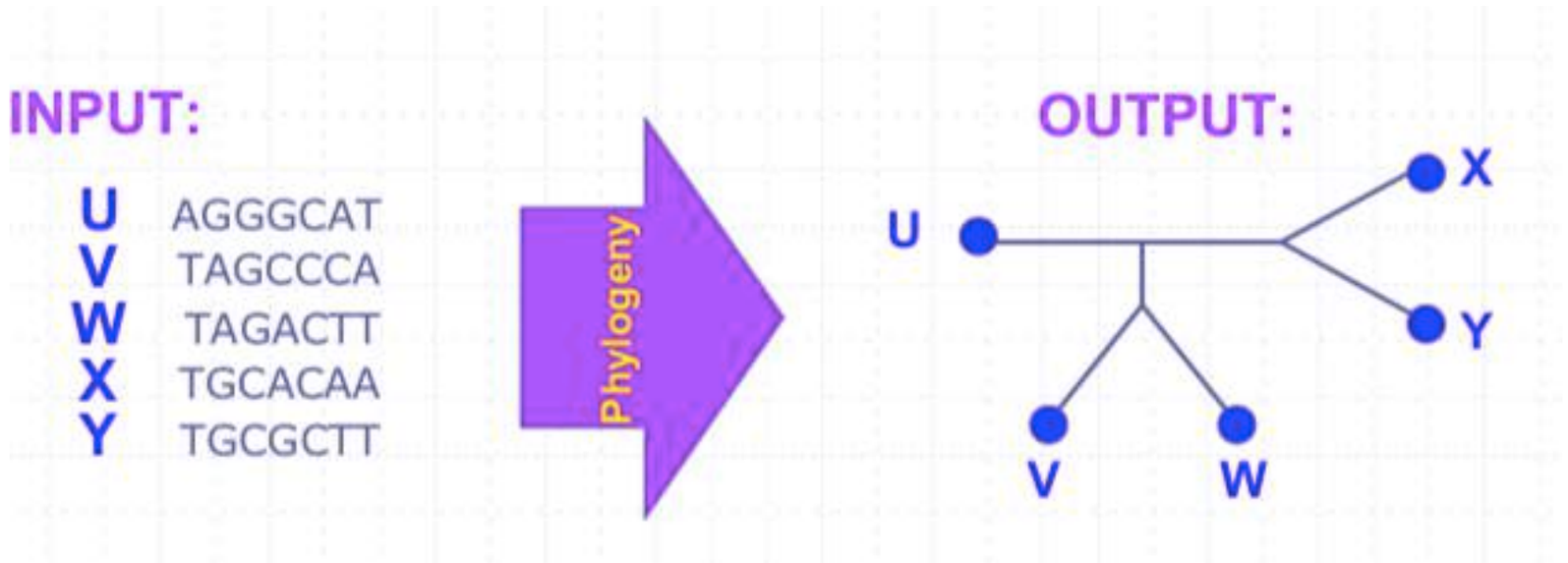
D(i,j) : distance entre le mot de l'individu i et le mot de l'individu j
nombre de mutations
temps écoulé depuis différentiation de leur ancêtre commun le plus récent





D(a,b)=8+50+42+17=117

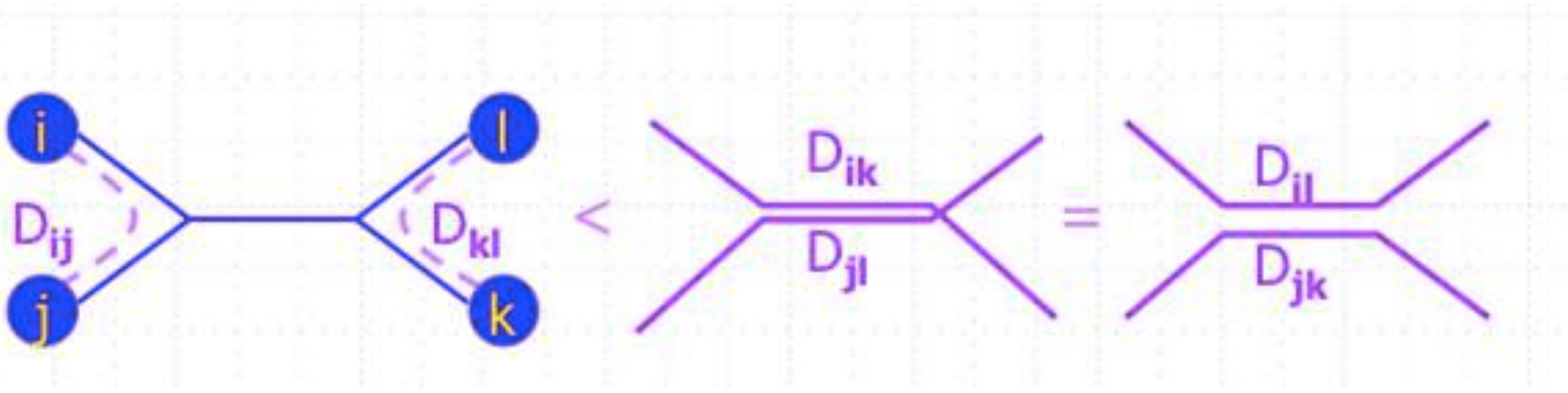
Étant donné les distances, reconstruire l'arbre.



Définition

Une matrice D est additive si pour tout quadruplet i,j,k,l:

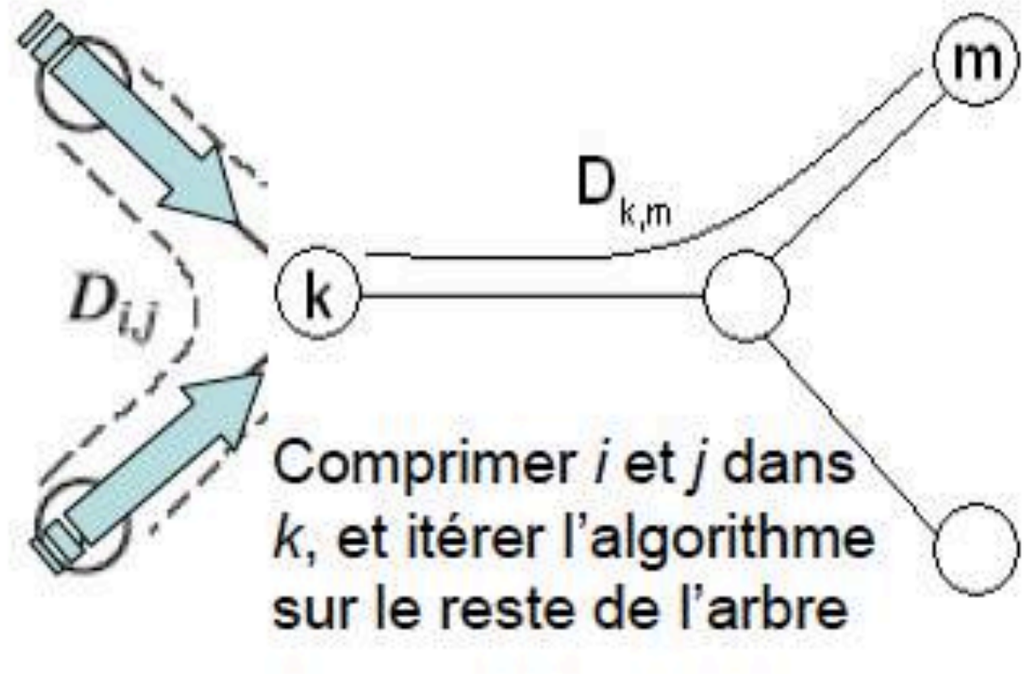
$$D_{ij} + D_{kl} \leq D_{ik} + D_{jl} = D_{il} + D_{jk}$$



Algorithme pour distances additives

- Trouver des feuilles « voisines » *i* et *j* avec père *k*
- Effacer lignes et colonnes de *i* et j de la matrice
- Ajouter une nouvelle ligne et colonne correspondant a k, ou la distance entre k et toute autre feuille m peut être calculée comme suit :

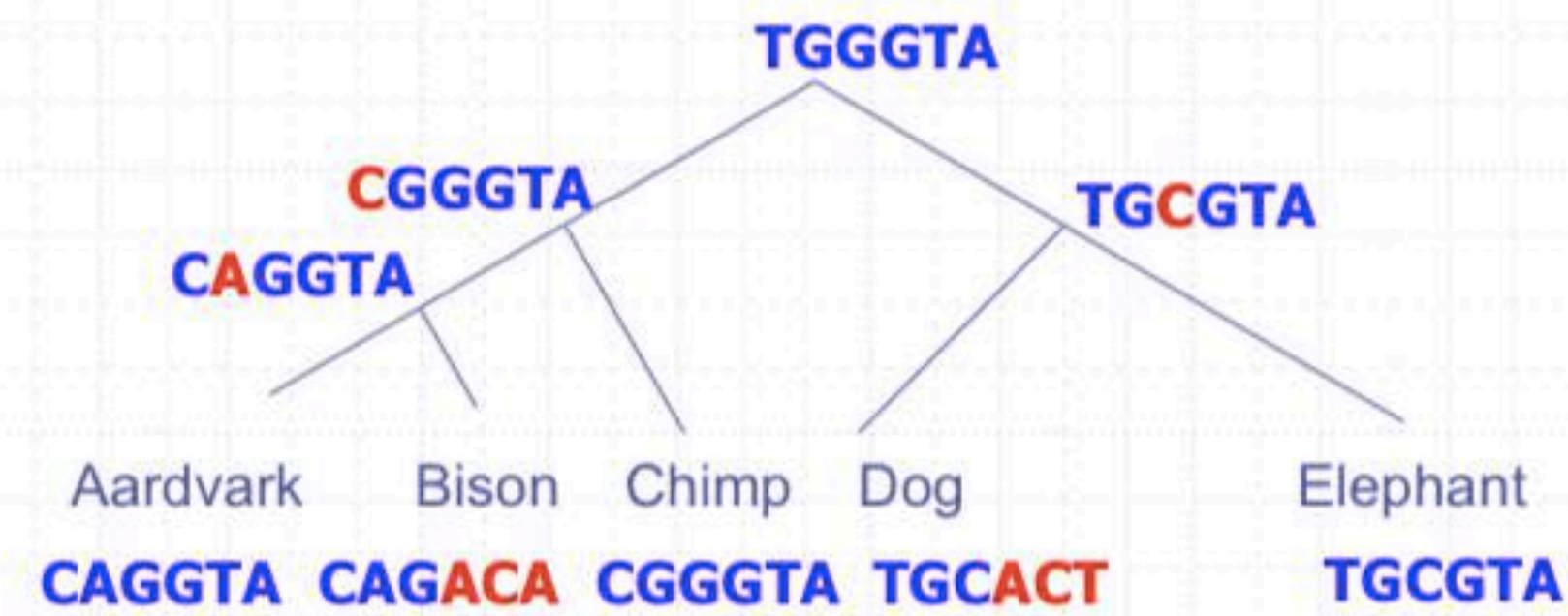
$$D_{km} = (D_{im} + D_{jm} - D_{ij})/2$$



(HT A Carbone)

Trouver les mots qui “expliquent” au mieux le présent

```
Aardvark: CAGGTA
Bison:    CAGACA
Chimp:    CGGGTA
Dog:      TGCACT
Elephant: TGCGTA
```



- **Au niveau des espèces : arbre d'évolution**
- **Au niveau des populations : ancêtre commun**
- **Au niveau des langages : évolution des langues indo-européennes**
- **Au niveau des individus : arbres généalogiques ?**

Marqueur : Segment d'ADN DYS 393.
Tous les hommes, quelle que soit leur origine,
ont la même séquence AGAT sur ce même marqueur.
Ils ne diffèrent que par le nombre de répétitions.
Celles ci peuvent aller de 9 à 17 pour ce marqueur.

Plus le nombre de marqueurs identiques est élevé,
plus la parenté entre 2 hommes est proche.
Les résultats des différents marqueurs donne
notre identité génétique appelée HAPLOTYPE.
C'est avec cet haplotype que l'on peut
rechercher ses cousins dans une base de données.

Locus 1 2 3 4 5 6 7 8 9 10
DYS 393 390 19 391 385a 385b 426 388 439 389
Allèles 13 24 14 11 11 14 12 12 13 13

Table 1
Y-STR profile for the putative Louis XVI sample, independently replicated in the Bologna and Barcelona laboratories.

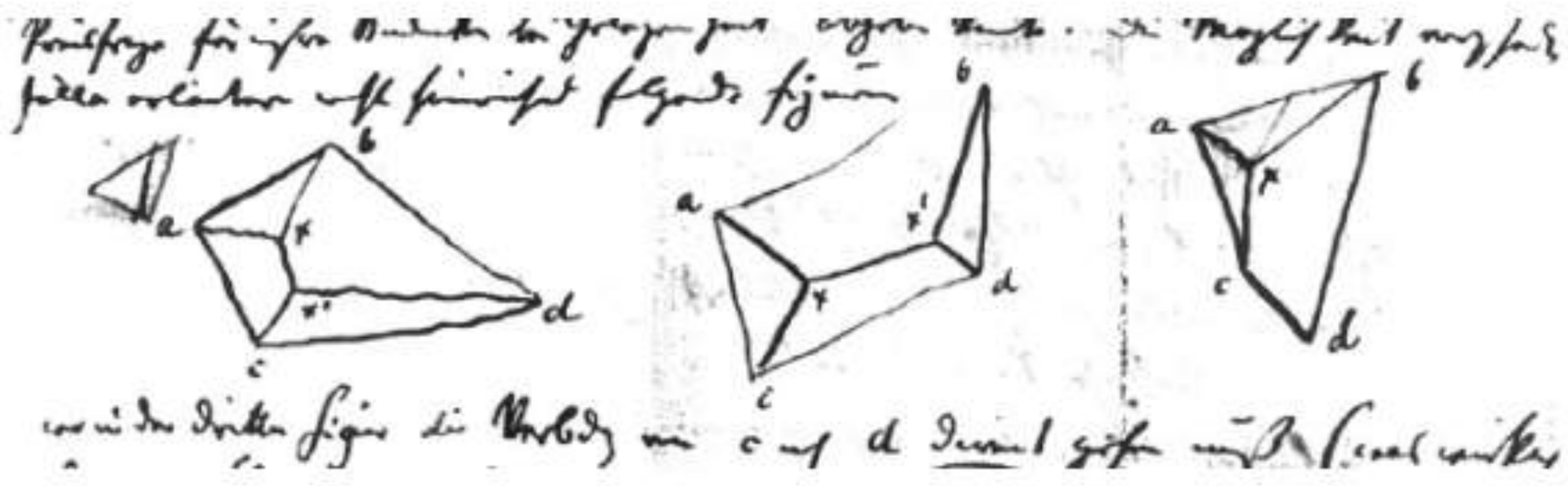
Marker	Alleles (Bologna)	Alleles (Barcelona)
DYS389I	12	12
DYS389II	30	30
DYS390	22	22
DYS456	15	15
DYS19	15	15
DYS385	13, 18	13, 18
DYS458	21	21
DYS437	15	15
DYS438	10	10
DYS448	21	21
YGATAH4	12	12
DYS391	10	10
DYS392	11	11
DYS393	14	14
DYS439	12	12
DYS635	21	21

L’anonymat du don d’ADN est un leurre…

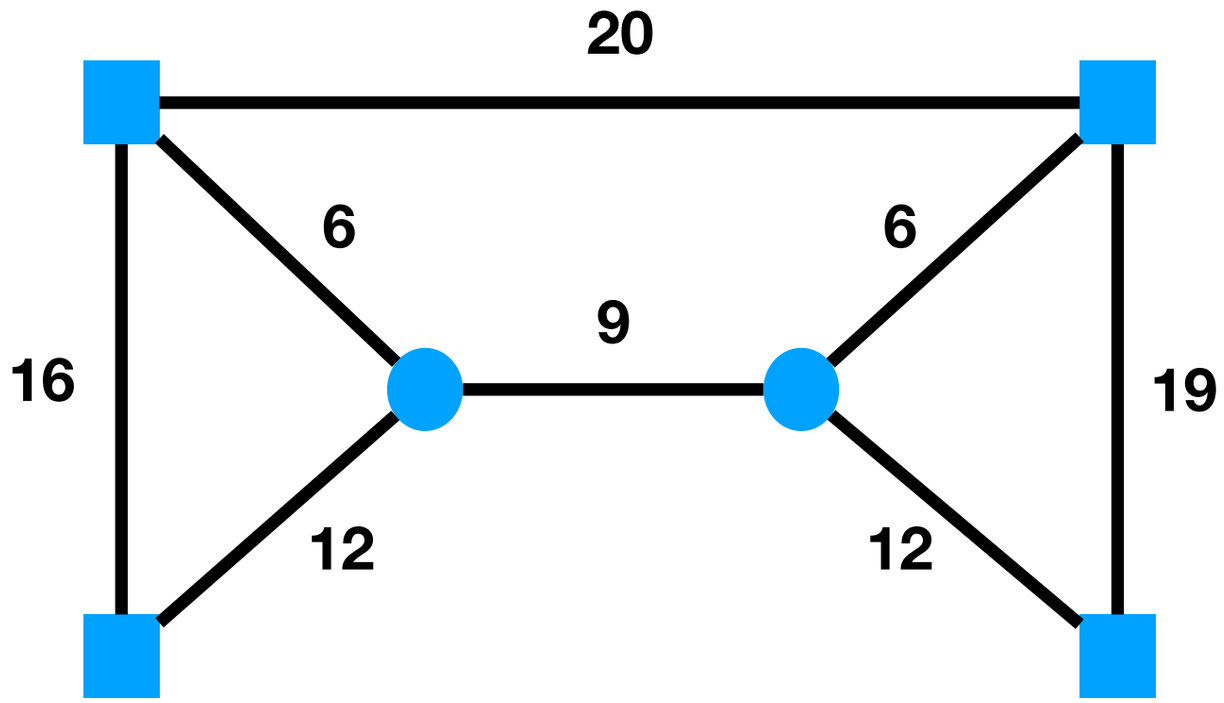
Arbre de Steiner et méthode primal-dual

arbre de Steiner

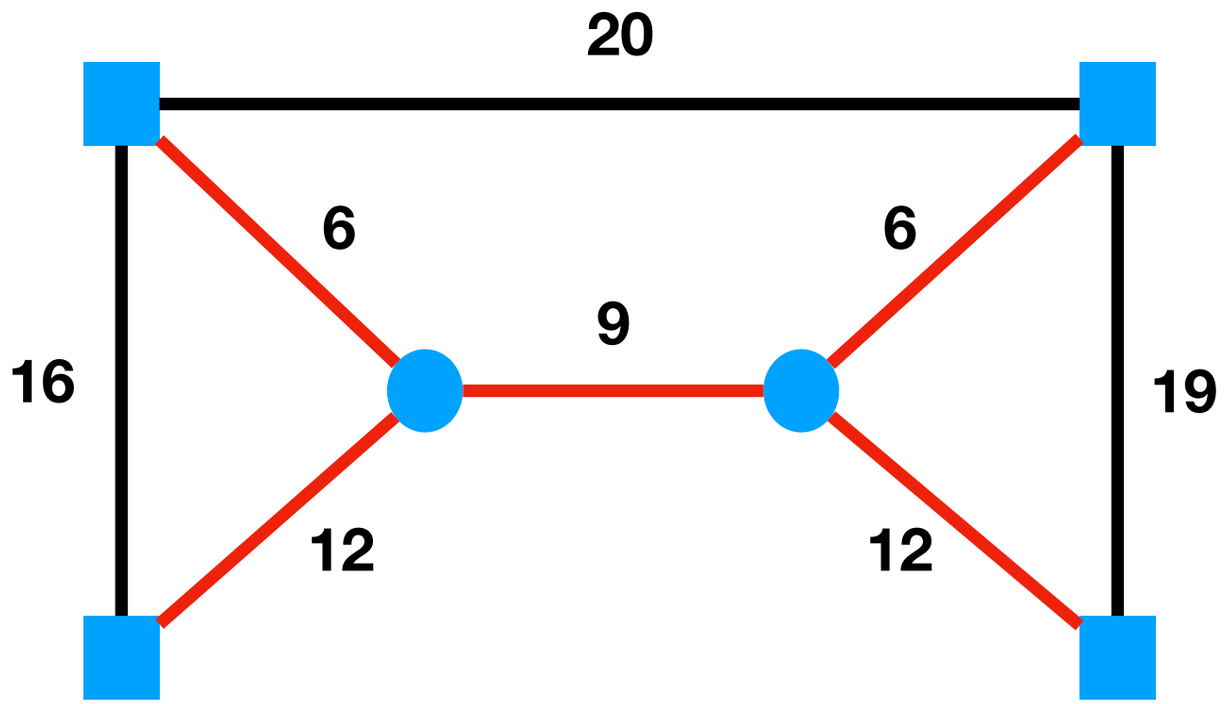
Quelle est la façon la plus courte de connecter quatre points a,b,c,d ?



Mentionné en 1836 dans une lettre de Gauss à Schumacher



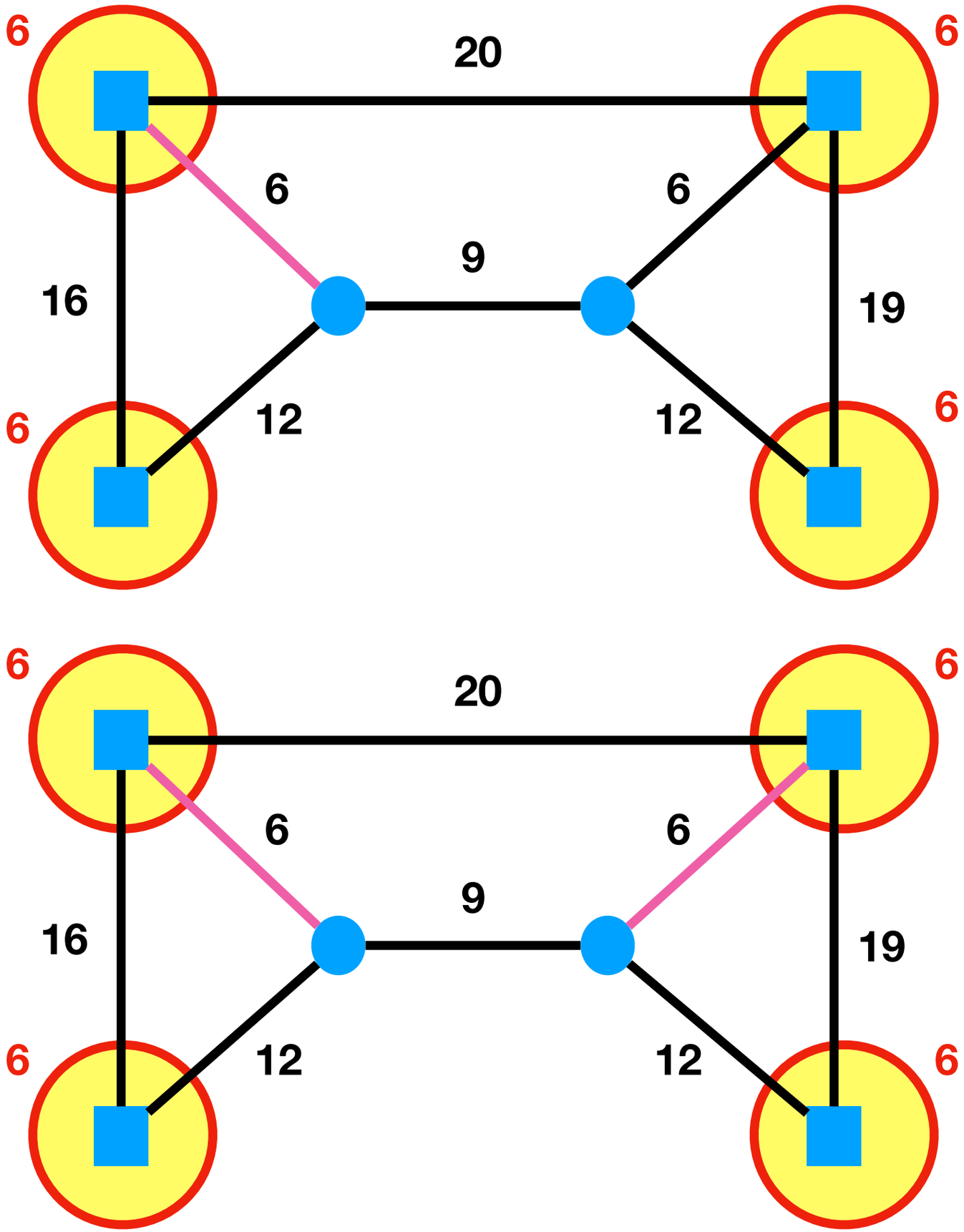
Graphes avec terminaux à connecter



OPT

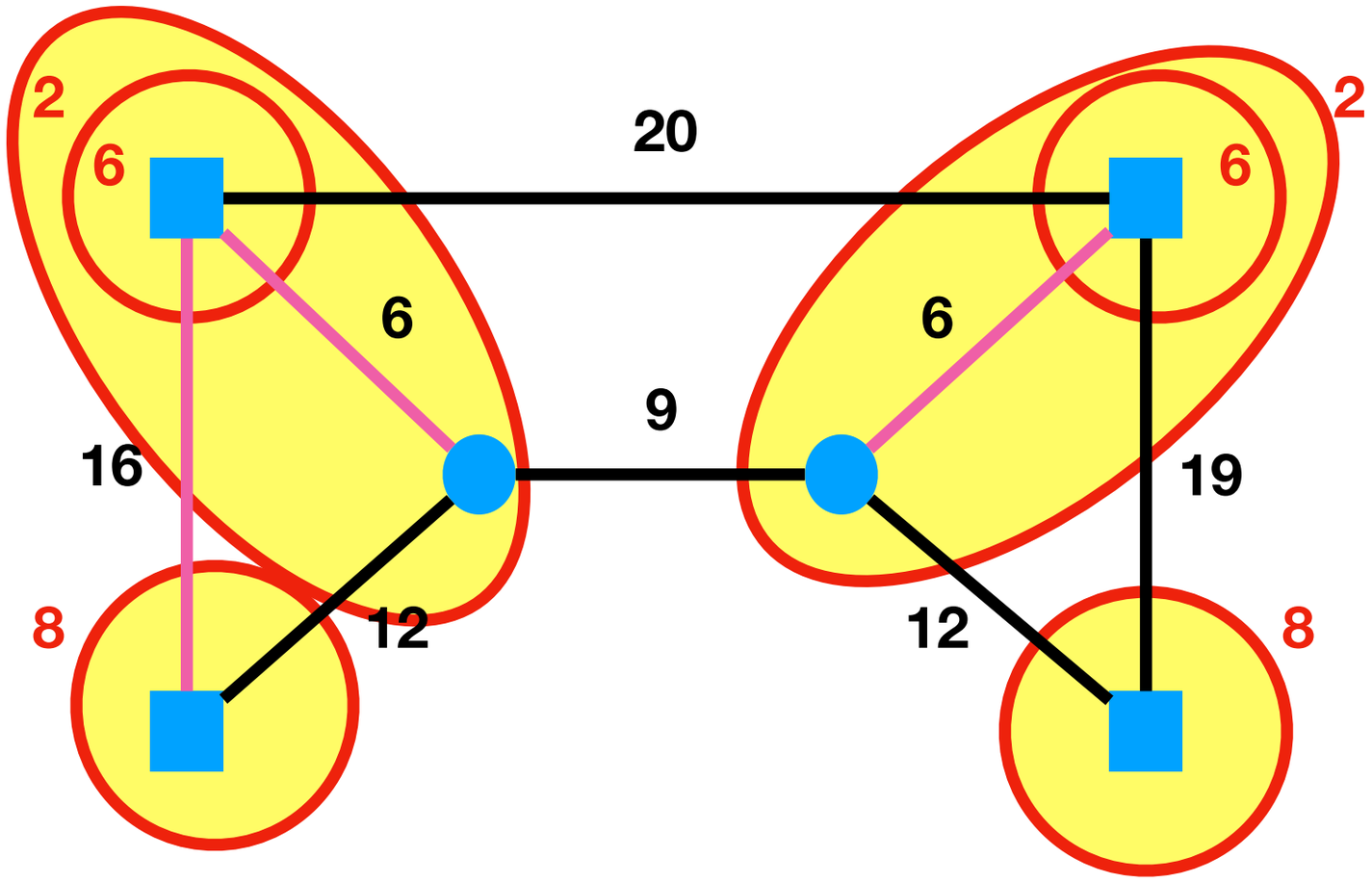
6+12+9+6+12=45

Algorithme

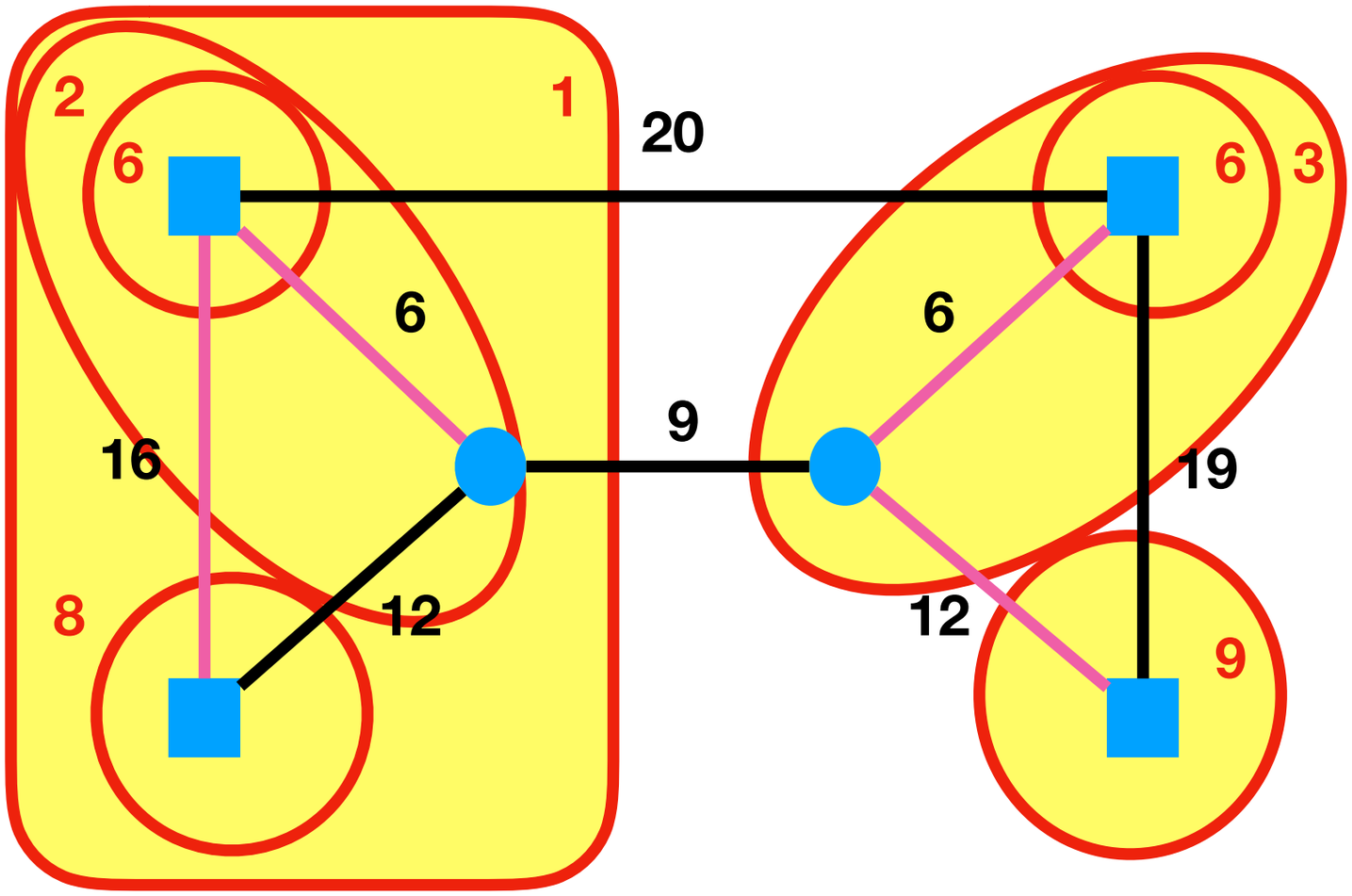


t=6

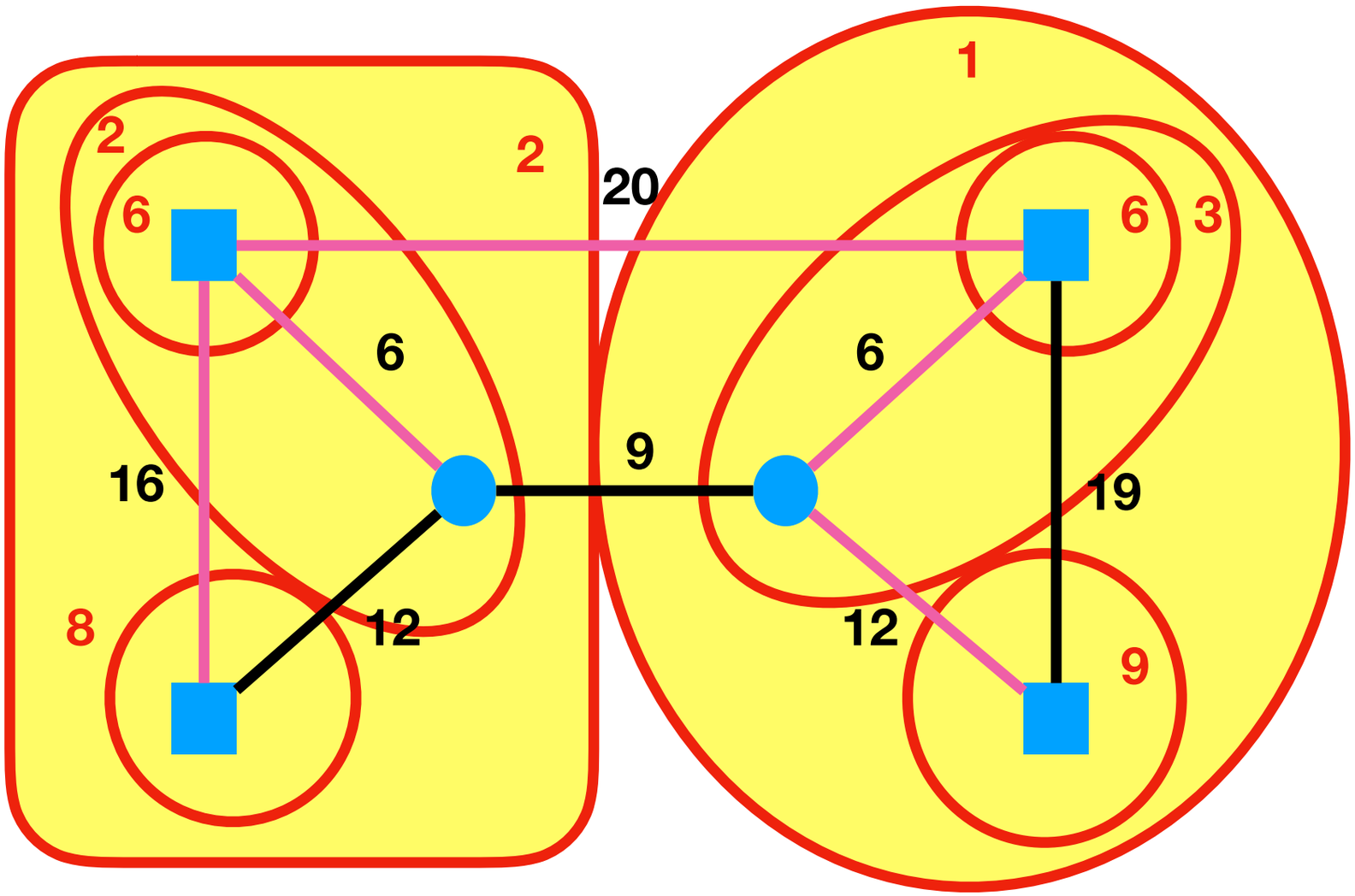
t=6



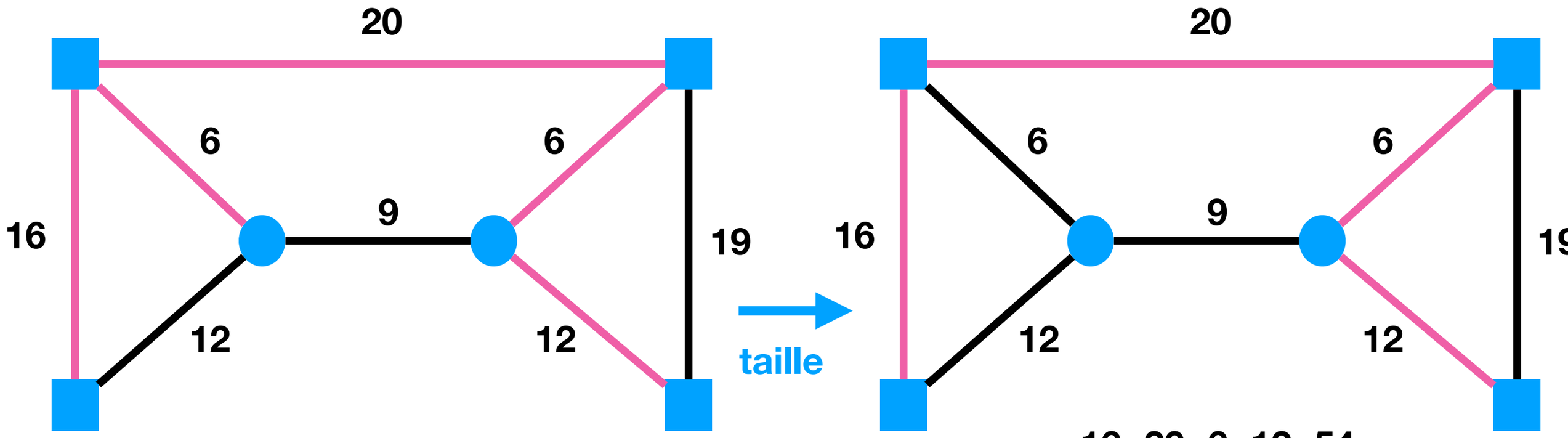
t=8



t=9

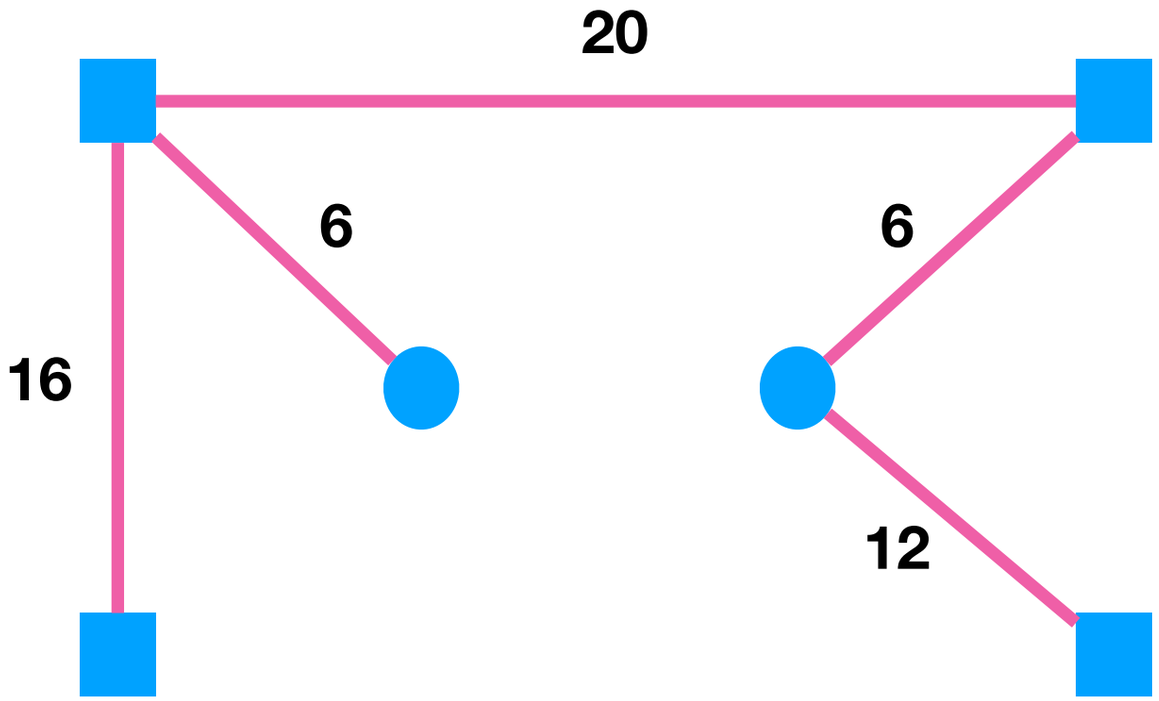


t=10



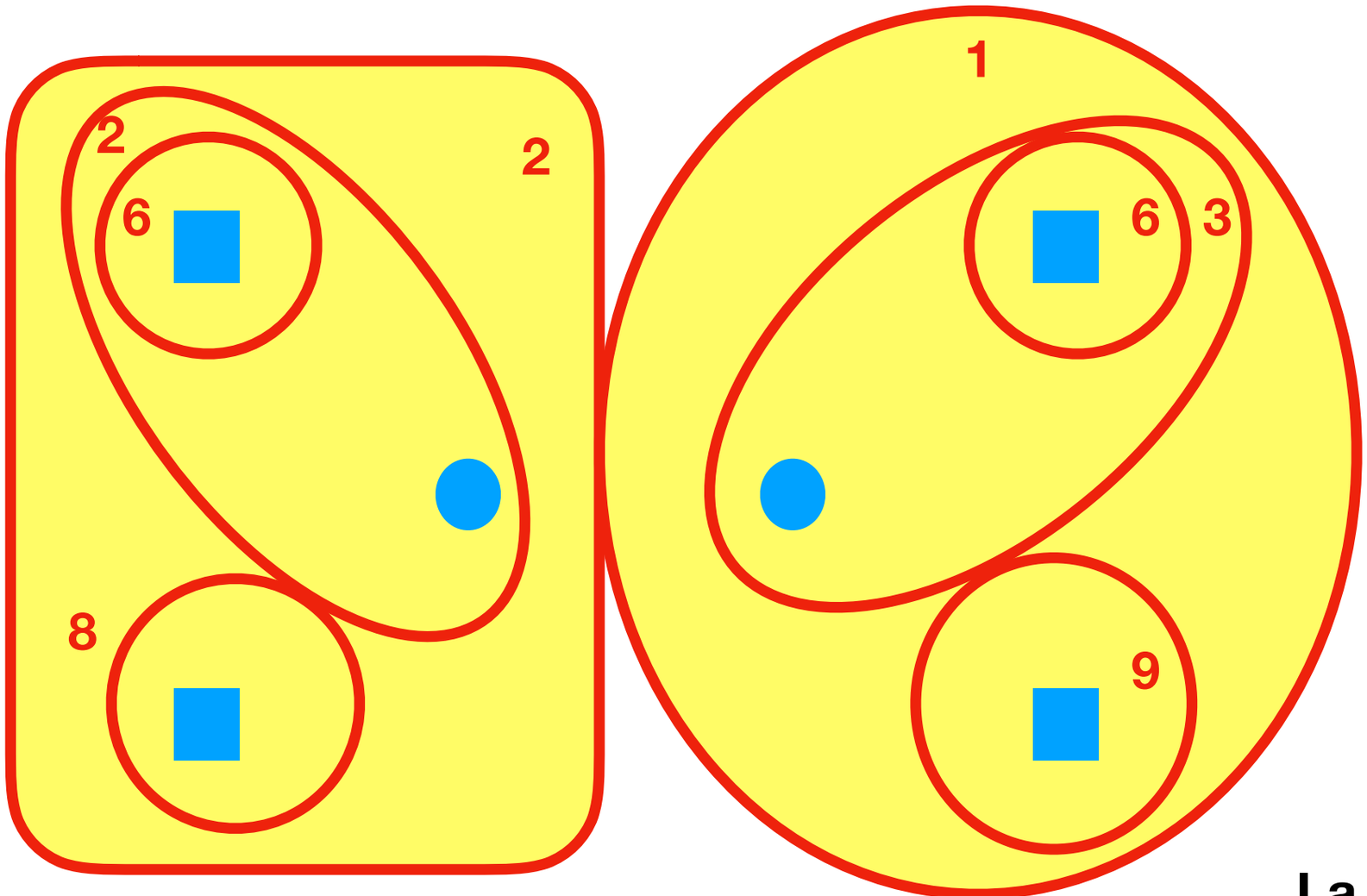
16+20+6+12=54

Ce que l'algorithme fait :
deux structures construites en simultané



Un arbre
connectant les terminaux
(structure primale)

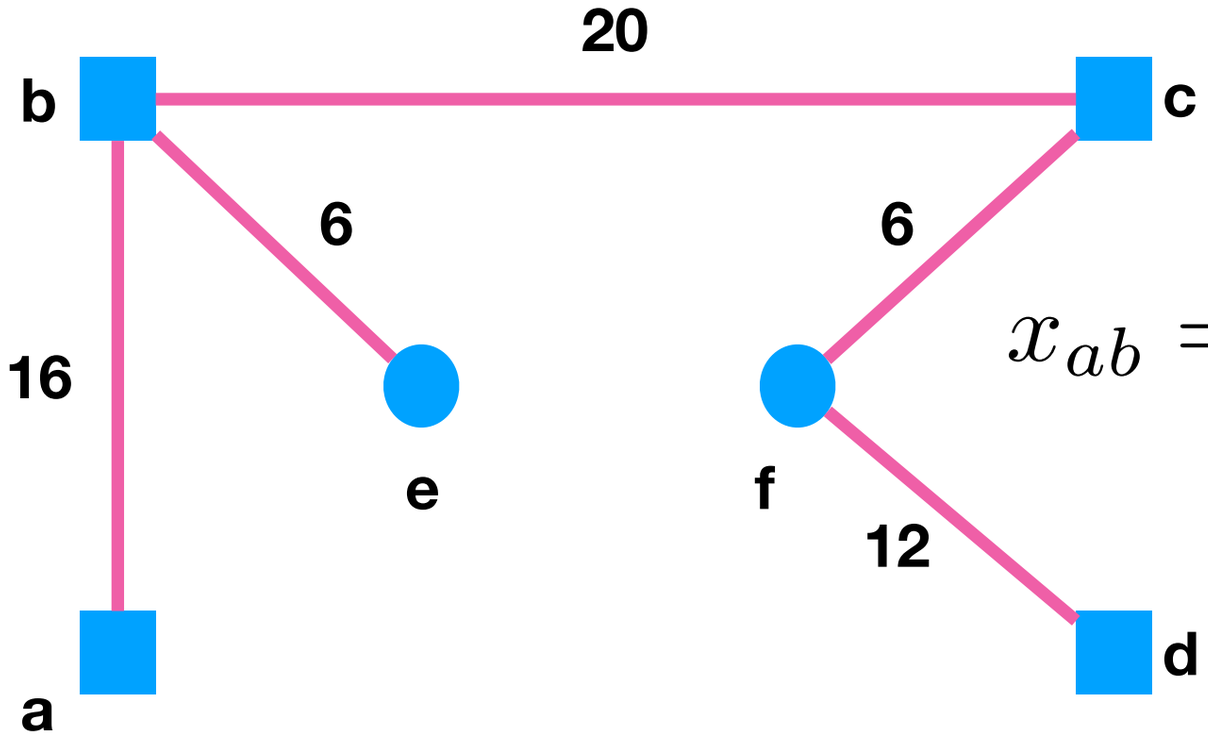
La structure définie par
le problème



Un ensemble de
coupes imbriquées
(structure duale)

Une structure "cachée"
révélée par l'analyse
du problème

La 2e structure guide l'algorithme

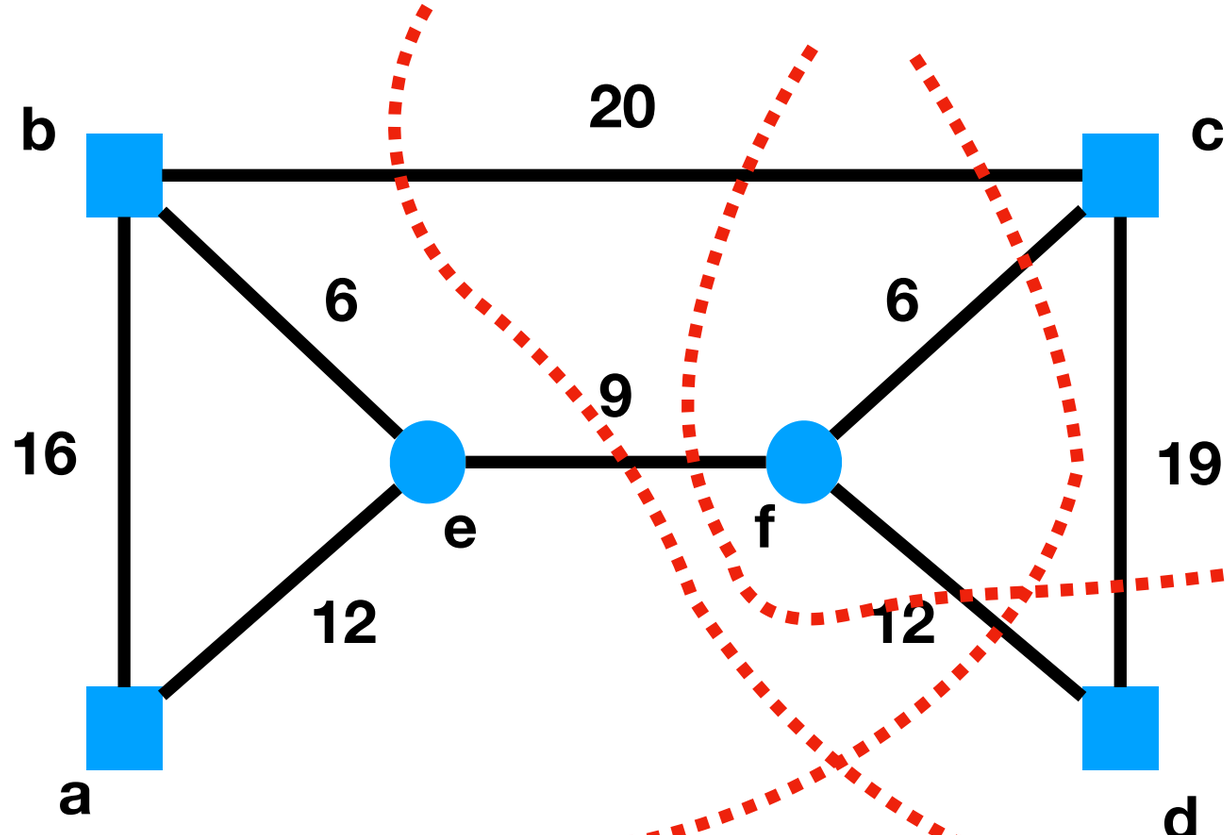


Un arbre connectant les terminaux (structure primale)

$$x_{ab} = x_{be} = x_{bc} = x_{cf} = x_{fd} = 1$$

$$x_{ae} = x_{ef} = x_{ad} = x_{cd} = 0$$

$$\min 16x_{ab} + 6x_{be} + 12x_{ae} + 20x_{bc} + 9x_{ef} + 6x_{fc} + 12x_{fd} + 19x_{cd}$$



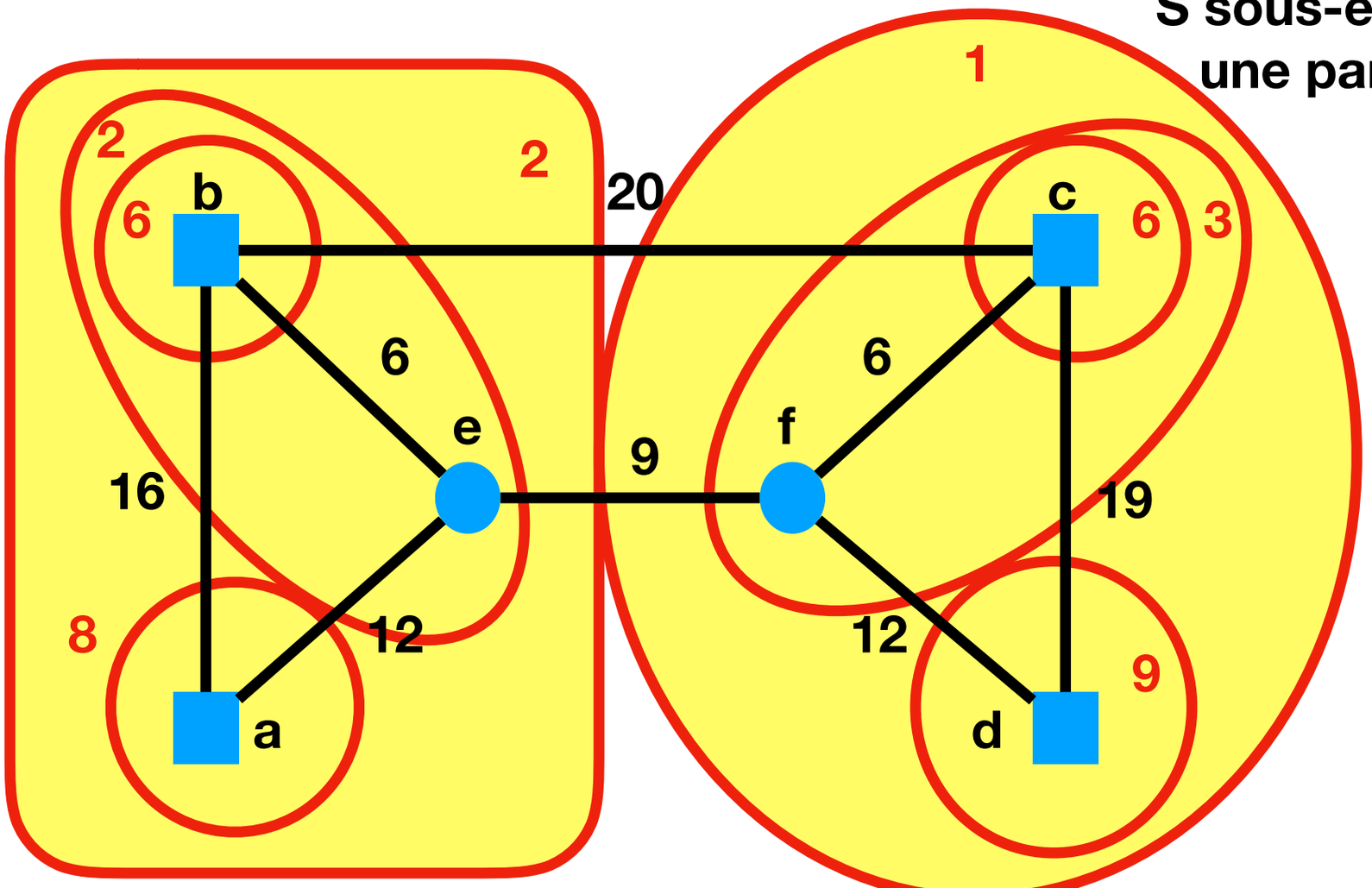
$$0 \leq x_e \leq 1$$

Toute coupe S séparant les terminaux doit être traversée par au moins une arête de l'arbre

$$x_{bc} + x_{cf} + x_{fd} \geq 1$$

$$x_{bc} + x_{cf} + x_{fd} \geq 1$$

$$x_{bc} + x_{ef} \geq 1$$



S sous-ensemble des sommets contenant une partie non-triviale des terminaux = coupe séparant les terminaux

$$0 \leq y_S \leq 1$$

$$y_b + y_{be} + y_{bea} + y_{cfd} + y_{cf} + y_c + \cdots \leq 20$$

$$y_b + y_{be} + y_a + \cdots \leq 16$$

$$\sum_{S : e \in S \times (V \setminus S)} y_S \leq (\text{poids de } e)$$

$$\max y_{bea} + y_{be} + y_b + y_e + y_a + y_{fcd} + y_{fc} + y_f + y_c + y_d + \cdots$$

Toute coupe S séparant les terminaux doit être traversée par au moins une arête e de l'arbre

Variables x(e)

Minimiser le poids total des arêtes sélectionnées

Toute arête e doit être coupée par au plus poids(e) coupes S séparant les terminaux

Variables y(S)

Maximiser le nombre total de coupes sélectionnées

Dualité

$$\begin{array}{l} \min \sum_e w(e) x_e : \\ \begin{cases} \sum_{e \text{ coupé par } S} x_e & \geq 1 \quad (\forall S) \\ 0 \leq x_e & \leq 1 \quad (\forall e) \end{cases} \end{array} \longleftrightarrow \begin{array}{l} \max \sum_S y_S : \\ \begin{cases} \sum_{S \text{ coupant } e} y_S & \leq w_e \quad (\forall e) \\ 0 \leq y_S & \leq 1 \quad (\forall S) \end{cases} \end{array}$$

Théorème

$$\sum_e x_e^* = \min \sum_e w(e) x_e = \max \sum_S y_S = \sum_S y_S^*$$

On peut montrer que l'analyse construit (x_e) **et** (y_S) **tels que** $\sum_e x_e \leq 2 \sum_S y_S$

Donc poids de l'arbre résultat

$$= \sum_e x_e \leq 2 \sum_S y_S \leq 2 \sum_S y_S^* = 2 \sum_e x_e^* \leq 2\ OPT$$

✓

Comprendre la dualité (1/4)

$$\begin{array}{l} \max 6x_1 + 14x_2 + 13x_3 : \\ \left\{ \begin{array}{rcl} x_1 + 4x_2 + 2x_3 & \leq & 48 \\ x_1 + 2x_2 + 4x_3 & \leq & 60 \\ 0 \leq x_1, x_2, x_3 & & \end{array} \right. \end{array}$$

Est-il possible que la valeur optimale vaille 10000?

$$6x_1 + 14x_2 + 13x_3 = 10000?$$

Non, car

$$6x_1 + 14x_2 + 13x_3 \leq 6.5(x_1 + 4x_2 + 2x_3) \leq 312$$

Comprendre la dualité (2/4)

$$\begin{array}{l} \max 6x_1 + 14x_2 + 13x_3 : \\ \begin{cases} x_1 + 4x_2 + 2x_3 & \leq \quad 48 \\ x_1 + 2x_2 + 4x_3 & \leq \quad 60 \\ 0 \leq x_1, x_2, x_3 \end{cases} \end{array}$$

Est-il possible que la valeur optimale vaille 312?

$$6x_1 + 14x_2 + 13x_3 = 312?$$

Non, car

$$6x_1 + 14x_2 + 13x_3 \leq 5(x_1 + 4x_2 + 2x_3) + (x_1 + 2x_2 + 4x_3) \leq 300$$

Comprendre la dualité (3/4)

$$\max 6x_1 + 14x_2 + 13x_3 :$$
$$\begin{cases} x_1 + 4x_2 + 2x_3 \leq 48 \\ x_1 + 2x_2 + 4x_3 \leq 60 \\ 0 \leq x_1, x_2, x_3 \end{cases}$$

Quelle est la meilleure borne qu'on puisse obtenir ainsi ?
On multiplie la première inégalité par un coefficient positif y_1
On multiplie la deuxième inégalité par un coefficient positif y_2
tels que

$$6 \leq y_1 + y_2$$
$$14 \leq 4y_1 + 2y_2$$
$$13 \leq 2y_1 + 4y_2$$

Alors on a la borne $6x_1 + 14x_2 + 13x_3 \leq 48y_1 + 60y_2$
Donc la meilleure borne qu'on puisse obtenir ainsi est :

$$\min(48y_1 + 60y_2)$$

En résumé (4/4)

Primal

$$\max 6x_1 + 14x_2 + 13x_3 : \begin{cases} x_1 + 4x_2 + 2x_3 & \leq & 48 \\ x_1 + 2x_2 + 4x_3 & \leq & 60 \\ 0 \leq x_1, x_2, x_3 \end{cases}$$

Dual

$$\min 48y_1 + 60y_2 : \begin{cases} y_1 + y_2 & \geq & 6 \\ 4y_1 + 2y_2 & \geq & 14 \\ 2y_1 + 4y_2 & \geq & 13 \\ 0 \leq y_1, y_2 \end{cases}$$

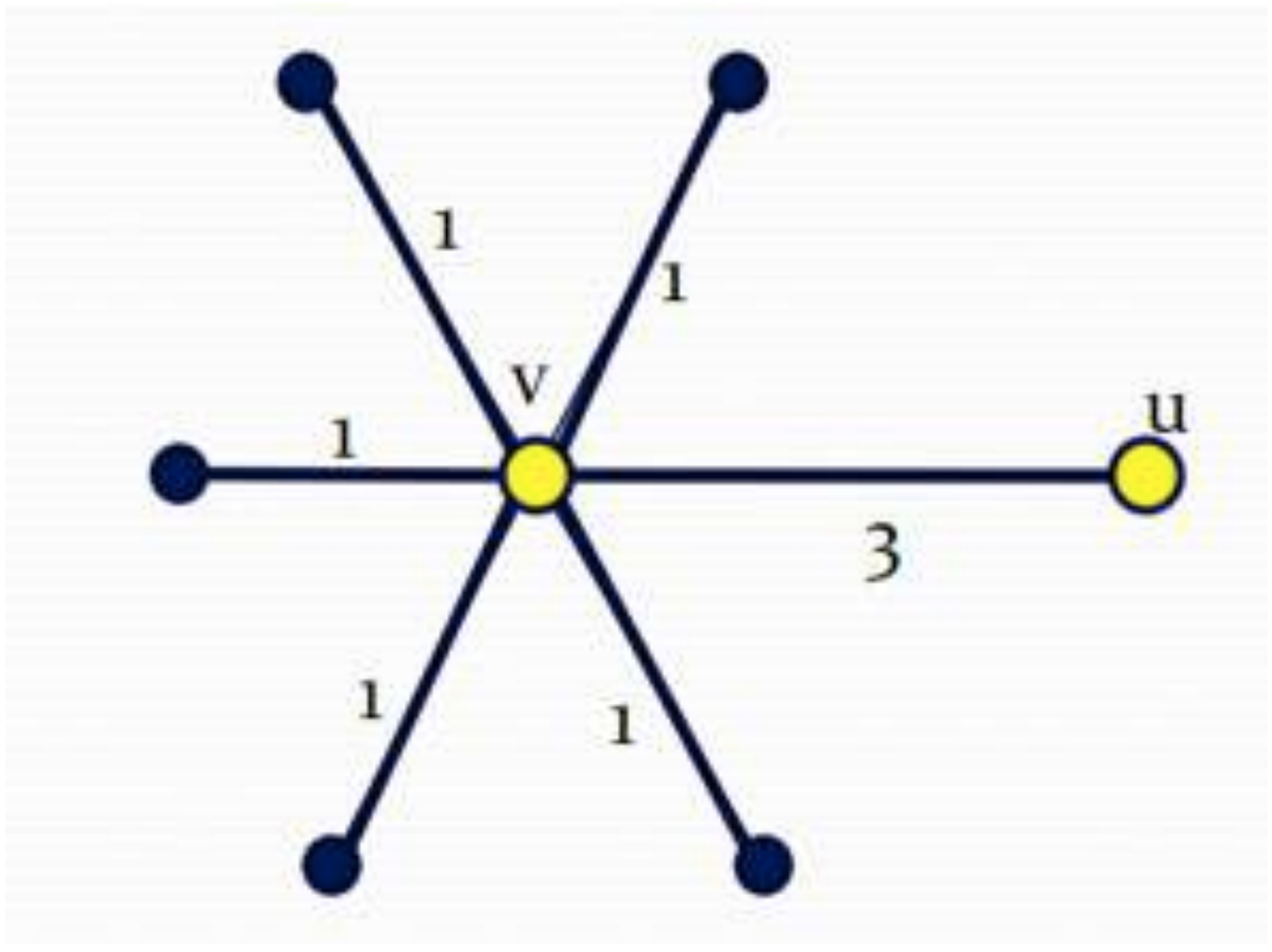
On a argumenté que :

$$\max 6x_1 + 14x_2 + 13x_3 \leq \min 48y_1 + 60y_2$$

Théorème

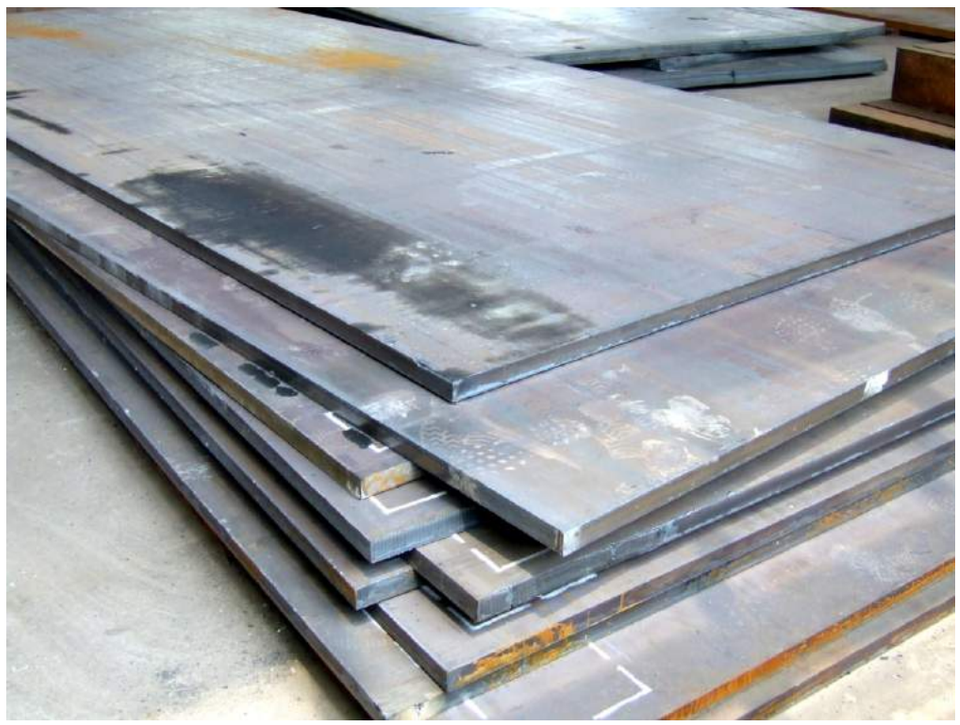
$$\sum_e x_e^* = \min \sum_e w(e)x_e = \max \sum_S y_S = \sum_S y_S^*$$

Remarque : utilité de la taille finale

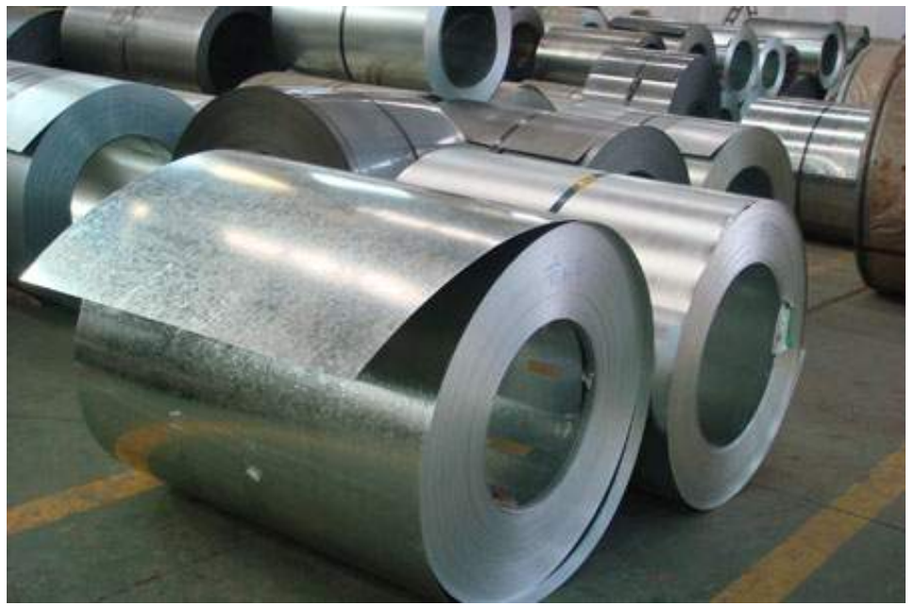


Application : sidérurgie

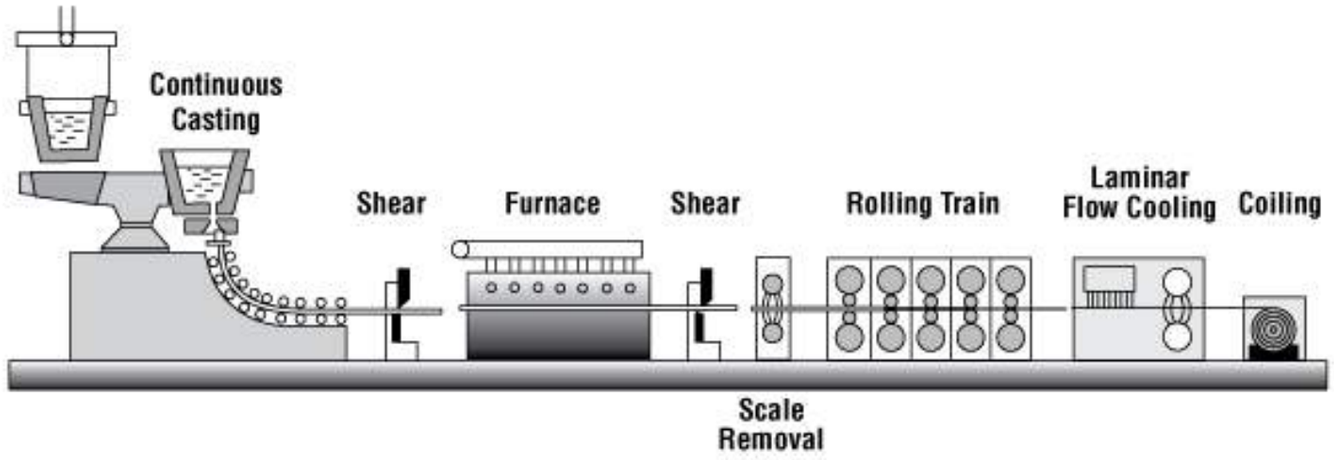
Usine de laminage d’acier



Brames d’acier



Une commande : rouleau d’acier de largeur, longueur, épaisseurs fixées

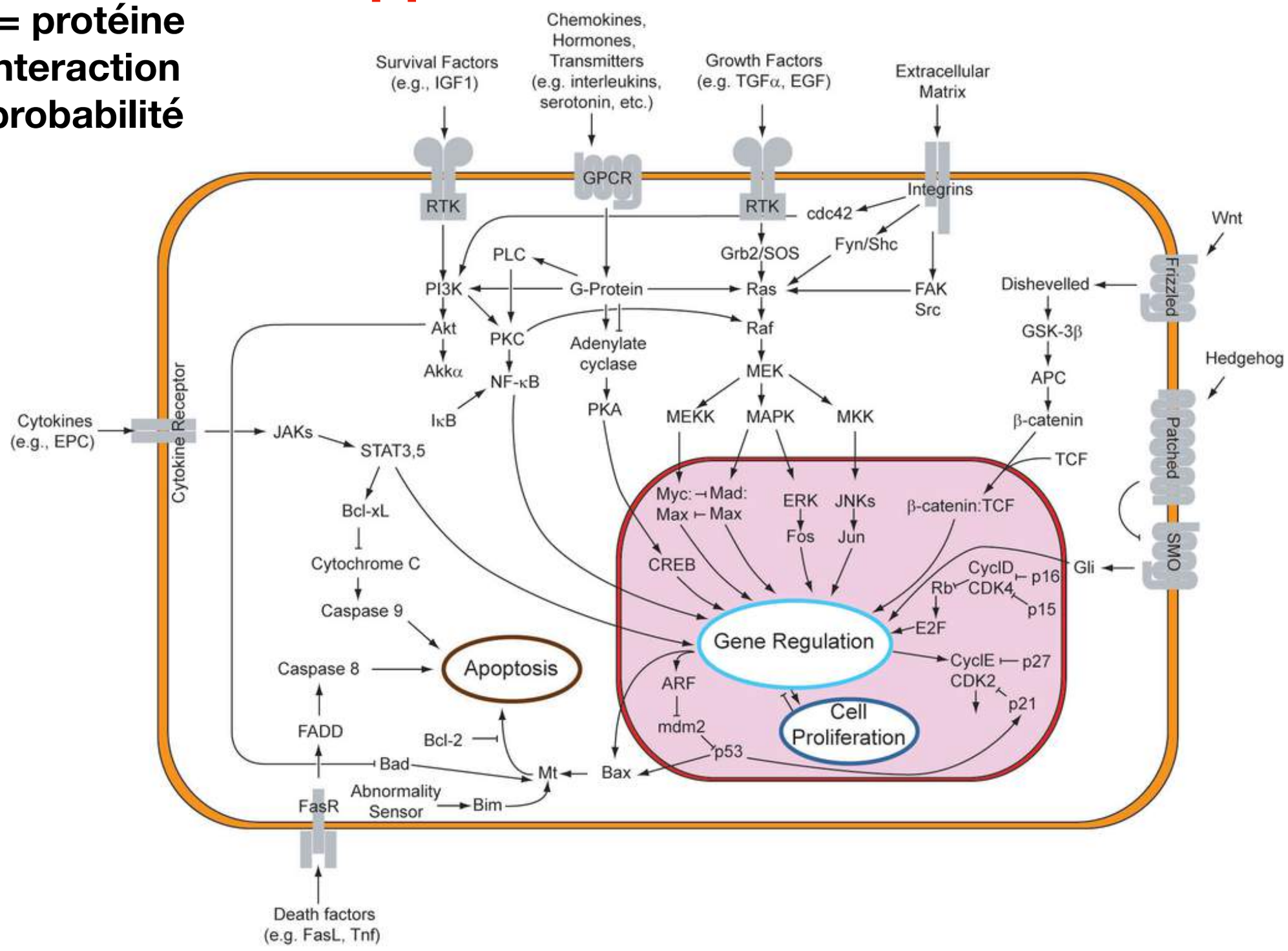


Coût de transition d’une commande à une autre

Voyageur de commerce avec récompenses

Réseau biologique
sommet = protéine
arête = interaction
poids = probabilité

Application : médecine



Flux d'information par interactions entre protéines
Identification de voies de signalisation cellulaire

Arbre de Steiner
avec récompenses

"We identify the role of the COS8 protein, a member of a gene family of previously unknown function,"

De l’utilité de l’étude théorique des algorithmes

- Méthode générale
- Algorithmes simples
- Quasi-impossibles à trouver sans connaître la dualité

Conclusion

Problèmes, algorithmes, analyse

Flux de données

Modèles

Graphes dynamiques

Questions

Modélisation et réseaux sociaux

Distribué

Théorie des jeux

Aléa

Techniques

Marches aléatoires

Schémas d’approximation

Programmation dynamique

Applications

Programmation linéaire

Biologie

Primal-dual

Réseaux

Médecine

Allocation de ressources dans la société

Usines

Problèmes ouverts