La distance de Levensthein

Problème

Etant donnés deux textes $x = x_1 ... x_m$ et $y = y_1 ... y_n$, quel est le nombre minimal d'opérations de remplacement d'une lettre par une autre (mismatch) et d'insertion/délétion de lettres permettant de passer de x à y ?

Exemples

```
groan
|||:|
grown
```

```
vermiform
::||:::::
formation
```

```
colo-r
|||| |
colour
```

```
vermiform-----
     ||||
-----formation
```

```
theatre
|||||::
theater
```

```
disestablishment
|||       |  |||
dis-------s--ent
```

```
elephant
|||: |||
eleg-ant
```

```
disestablishment
|||         :|||
dis---------sent
```

tiré de https://www.cs.umd.edu/class/fall2011/cmsc858s/Alignment.pdf

Calcul de la distance de Levensthein

$$d(x_{1..i}, y_{1..j}) = \min \begin{cases} s(x_i, y_j) + d(x_{1..i-1}, y_{1..j-1}) \\ 1 + d(x_{1..i-1}, y_{1..j}) \\ 1 + d(x_{1..i}, y_{1..j-1}) \end{cases}$$

avec $s(x_i, y_j) = 1$ si $x_i \neq y_j, 0$ sinon

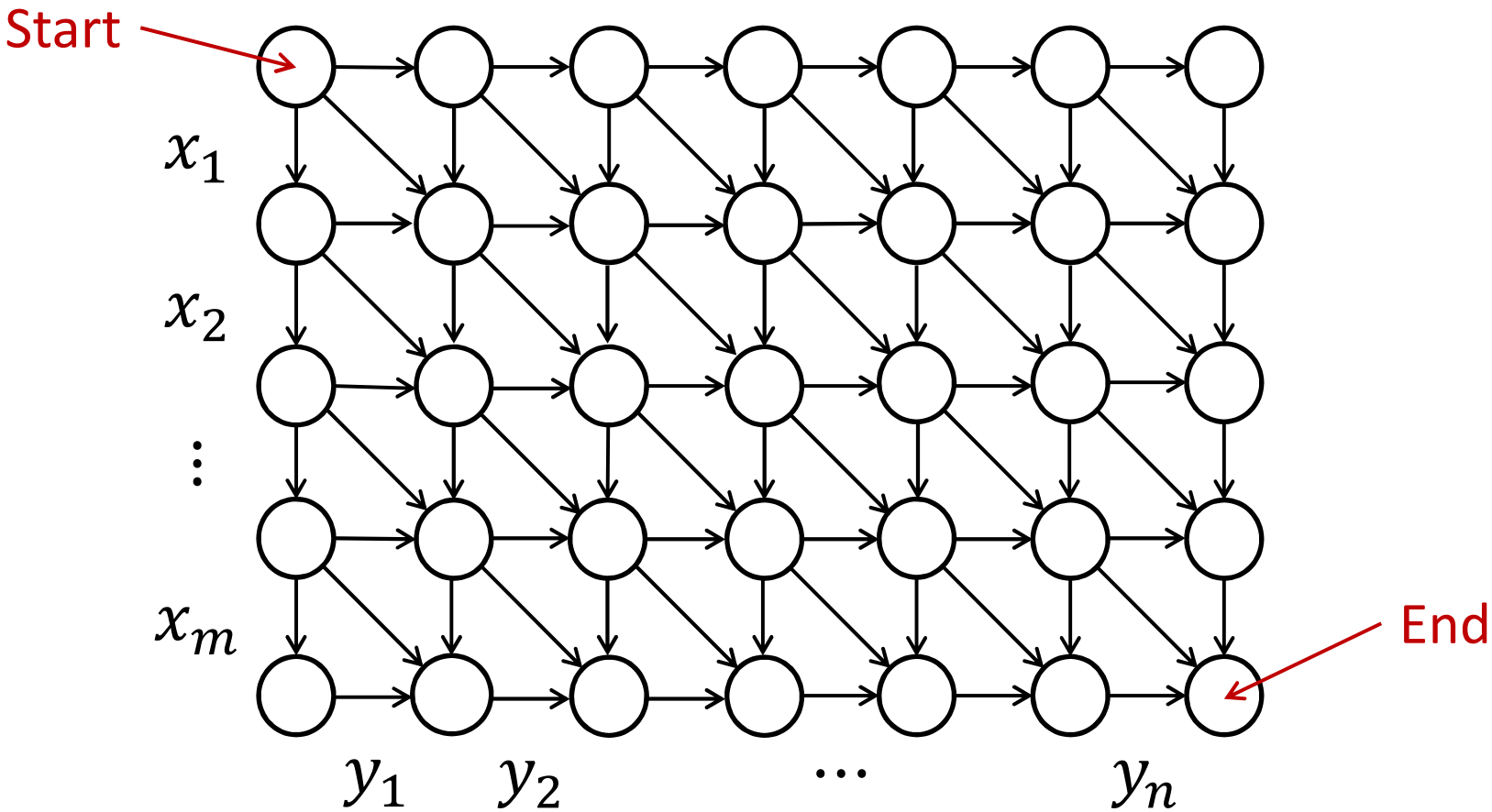
Calcul de la distance de Levensthein

$$d(x_{1..i}, y_{1..j}) = \min \begin{cases} s(x_i, y_j) + d(x_{1..i-1}, y_{1..j-1}) \\ 1 + d(x_{1..i-1}, y_{1..j}) \\ 1 + d(x_{1..i}, y_{1..j-1}) \end{cases}$$

avec $s(x_i, y_j) = 1$ si $x_i \neq y_j, 0$ sinon

et par convention $d(x_{1..0}, y_{1..j}) = j, d(x_{1..i}, y_{1..0}) = i$

Graphe et programmation dynamique



Alignement

Passage d'un calcul de distance à une mesure de similarité :

- $s(x_i, y_j)$ devient un score entre deux lettres de l'alphabet
- on introduit le coût g d'un indel

Problème

Etant donnés deux textes $x = x_1 ... x_n$ et $y = y_1 ... y_m$, quel est la similarité maximale entre x et y avec les opérations de conservation d'une lettre (match), de remplacement d'une lettre par une autre (mismatch) et d'insertion/délétion de lettres (indel) permettant de passer de x à y ?

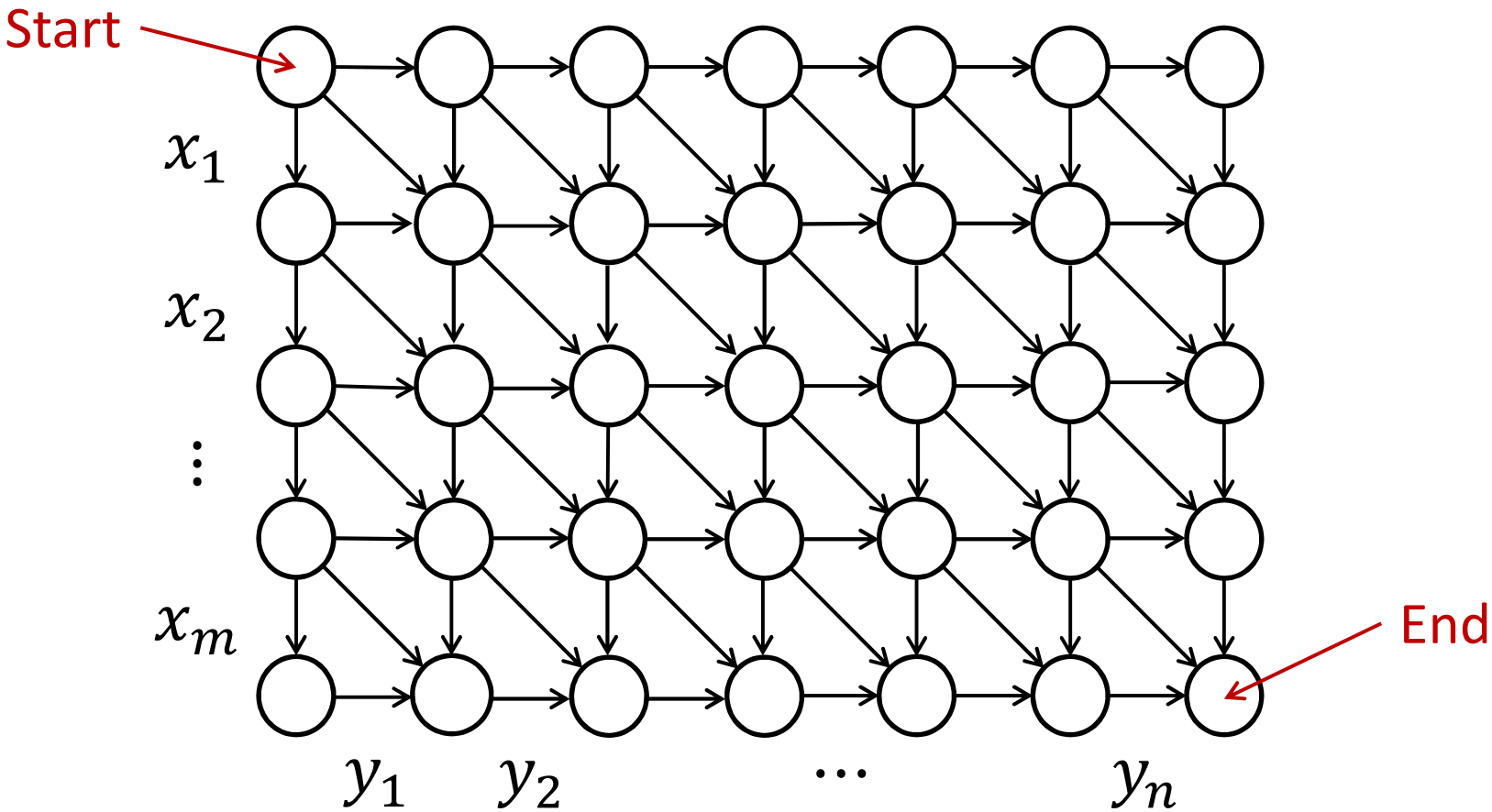
Calcul du score d'alignement (global) maximal

$$\mathcal{S}(x_{1..i}, y_{1..j}) = \max \begin{cases} s(x_i, y_j) + \mathcal{S}(x_{1..i-1}, y_{1..j-1}) \\ g + \mathcal{S}(x_{1..i-1}, y_{1..j}) \\ g + \mathcal{S}(x_{1..i}, y_{1..j-1}) \end{cases}$$

avec $s(x_i, y_j)$ une matrice de scores

et par convention $\mathcal{S}(x_{1..0}, y_{1..j}) = j \times g, \mathcal{S}(x_{1..i}, y_{1..0}) = i \times g$

Graphe et programmation dynamique



Des exemples insatisfaisants (1/2)

```
vermiform-----
     ||||
-----formation
```

Ce qui nous intéresse, c'est juste détecter que le préfixe ressemble au suffixe.

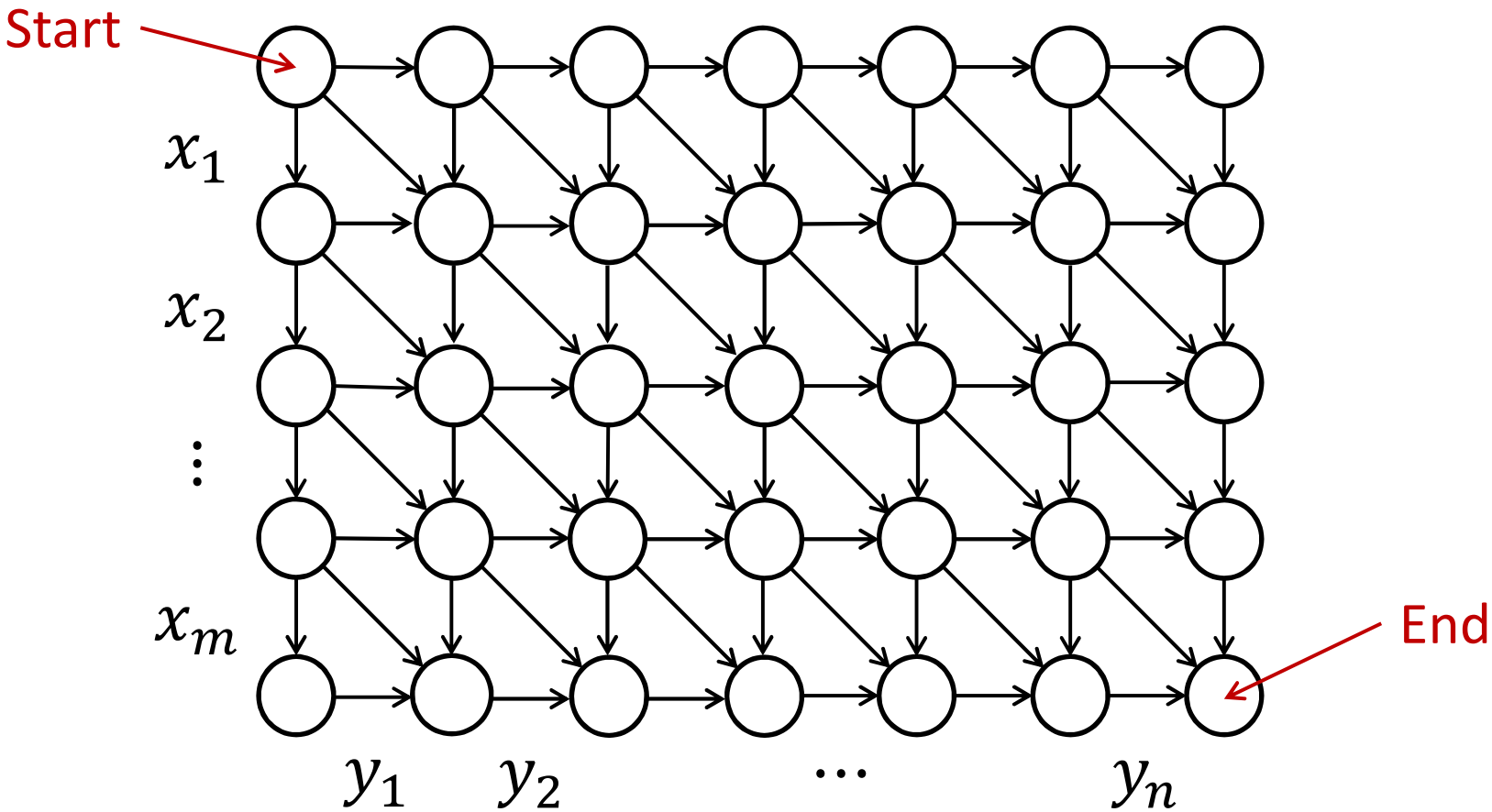
Calcul du score d'alignement (semi-global) maximal

$$\mathcal{S}(x_{1..i}, y_{1..j}) = \max \begin{cases} s(x_i, y_j) + \mathcal{S}(x_{1..i-1}, y_{1..j-1}) \\ g + \mathcal{S}(x_{1..i-1}, y_{1..j}) \\ g + \mathcal{S}(x_{1..i}, y_{1..j-1}) \end{cases}$$

et par convention $\mathcal{S}(x_{1..0}, y_{1..j}) = 0, \mathcal{S}(x_{1..i}, y_{1..0}) = 0$

Le score d'alignement est le max des $\mathcal{S}(\cdot, y_{1..n}), \mathcal{S}(x_{1..0}, \cdot)$

Graphe et programmation dynamique



Des exemples insatisfaisants (2/2)

```
mind the gap -- ----   score al global       : -10
:::: ||| |||
keep the gap in mind
```

Des exemples insatisfaisants (2/2)

```
mind the gap -- ----
:::: ||| |||
keep the gap in mind
```

```
score al global       : -10
score al semi-global : 2
```

Des exemples insatisfaisants (2/2)

```
mind the gap -- ----    score al global       : -10
:::: ||| |||            score al semi-global  : 2
keep the gap in mind    score al. local       : 6
```

Calcul du score d'alignement (local) maximal

$$\mathcal{S}(x_{1..i}, y_{1..j}) = \max \begin{cases} 0 \\ s(x_i, y_j) + \mathcal{S}(x_{1..i-1}, y_{1..j-1}) \\ g + \mathcal{S}(x_{1..i-1}, y_{1..j}) \\ g + \mathcal{S}(x_{1..i}, y_{1..j-1}) \end{cases}$$

et par convention $\mathcal{S}(x_{1..0}, y_{1..j}) = 0, \mathcal{S}(x_{1..i}, y_{1..0}) = 0$

Le score d'alignement est le max des $\mathcal{S}(\dot{;})$

Pour aller plus loin

```
mind the gap -- ----
:::: ||| |||
keep the gap in mind
```

```
score al. global       : -10
score al. semi-global  : 2
score al. local        : 6
score des al. locaux   : 10
```

On aimerait que la ressemblance soit mesurée en totalisant les scores des meilleurs alignements locaux (non chevauchants).