

## Funzione python

- Scrivere una funzione `align(s1, s2)` che:
  - prenda in ingresso due stringhe `s1` e `s2`
  - stampi:
    - \* un allineamento completo della seconda stringa sulla prima che massimizzi il numero di match di caratteri.
    - \* il numero di match ottenuti.
- E.g. (`align.py`):

```
>>> from align import *
>>> align("FDHAIRHEED", "HAID")
HAIR
HAID
Alignment score: 3
```

- L'output non deve essere per forza in questo formato, basta che contenga le stesse informazioni

## Funzione python: suggerimento

1. l'idea e' far scorrere la seconda sequenza sulla prima, per tutte le possibili posizioni iniziali della prima
2. ad ogni possibile posizione iniziale calcolare la somma di match come punteggio del corrispondente allineamento.
3. salvarsi punteggio massimo ottenuto e sua posizione
4. alla fine stampare l'allineamento con massimo punteggio.

## Funzione python: versione semplificata

- Per chi non riesca a creare la funzione, provare a creare una funzione `frequency(s, chars)` che:
  - prenda in ingresso una due stringhe `s` e `chars`
  - stampi la frequenza con cui ciascun carattere in `chars` compare in `s`

```
>>> from align import *
>>> frequency("abcdsf sdf", "dsfc")
{'c': 1, 's': 2, 'd': 2, 'f': 2}
```

### Esercizi da linea di comando

- Selezionare da un file FASTA (`seq.fasta`, allegato) tutte le righe che contengano:
  - due sequenze di tipo  $CX_3C$  ossia costituite da una cisteina (C), una sequenza qualsiasi di tre aminoacidi e una seconda cisteina (esempio: CDFGC e CDGHC).
  - le due sequenze possono essere separate solo da una sequenza di almeno due aminoacidi tra istidine (H), cisteine (C), serine (S) e alanine (A).
- e.g. CAKDCCHCCAHC

### Esercizi da linea di comando

- Dato un file (`CSA.dat` allegato) con una serie di righe contenenti:
  - identificativo della proteina
  - nome dell'aminoacido o altra molecola
  - catena
  - numero di sequenza

separati da virgole.

1. rimuovere tutte le righe duplicate
2. quindi contare il numero delle istidine (HIS), cisteine (CYS), aspartici (ASP) e glutammici (GLU) che vi compaiono
3. e riportare l'istogramma delle frequenze come mostrato di seguito:

- Output:

```
26041 ASP
6463 CYS
18810 GLU
22340 HIS
```