Alignement plus rapide

1. méthodes heuristiques : hachage, arbres de suffixe, PD limitée (taille totale de trous bornée)

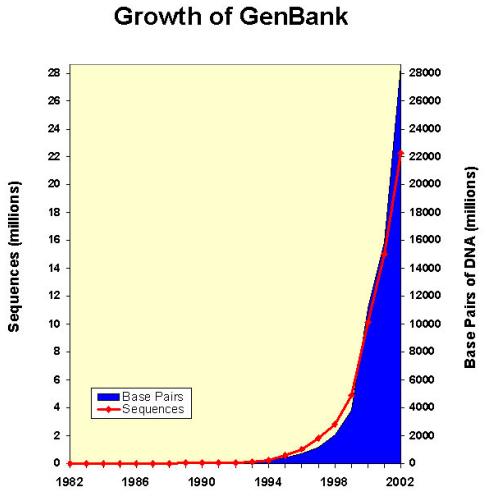
2. PD éparse (pour sous-séquence commune ou *chaînage* en alignement global heuristique)

Banques de séquences

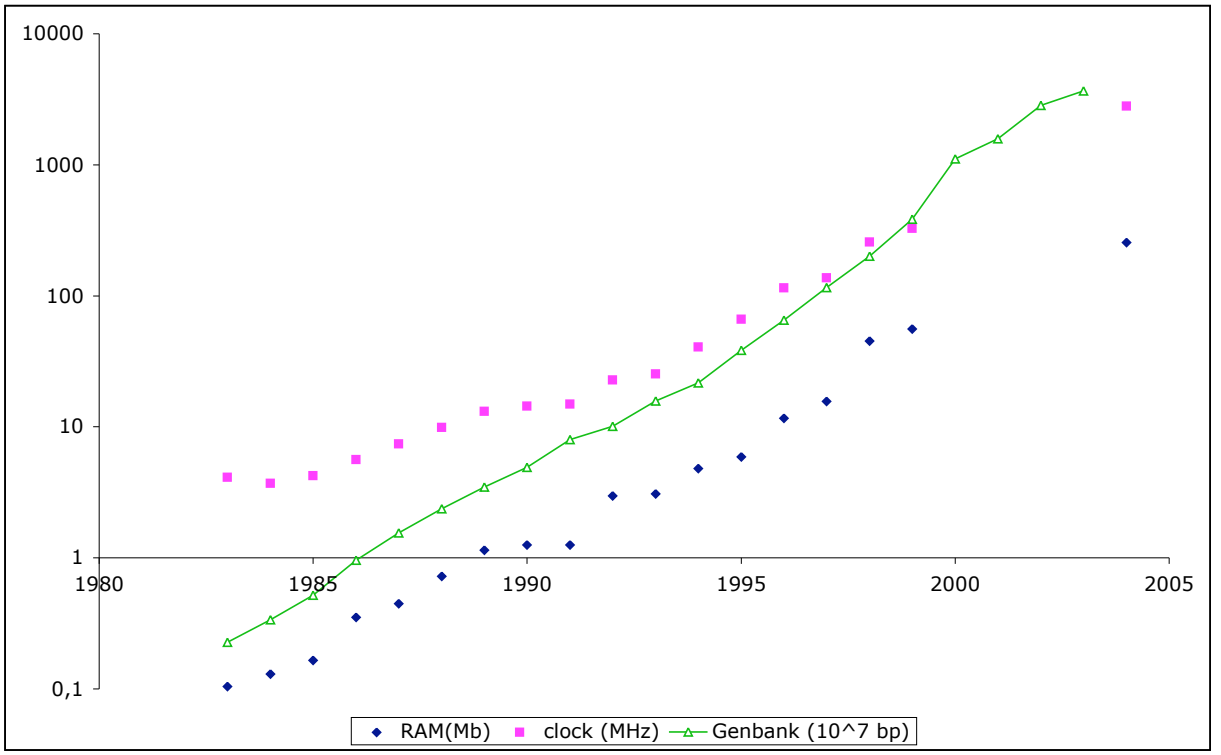
Bases de données de séquences : beaucoup d'information.

Exemple : GenBank

- 38 milliards de nucléotides ; 32.5 millions de séquences
- croissance exponentielle (taille doublée tous les 14 mois)



Need for speed



Banques de données

NCBI : «National Center for Biotechnology Information» — États-Unis

Interface [Entrez] à plusieurs bases de données :

- séquences d'acides nucléiques
- séquences protéiques
- PubMed : publications
- structures
- taxonomie
- ...

GenBank

Séquences d'ADN : GenBank (É-U), DDBJ (Japon), EMBL (Europe)

GenBank «flatfile» : exemple (HUMXIHB) :

```
LOCUS       HUMXIHB                  458 bp    mRNA    linear   PRI 14-JAN-1995
DEFINITION  Human zeta hemoglobin mRNA, complete cds.
ACCESSION   M24173
VERSION     M24173.1  GI:340391
KEYWORDS    zeta-globulin.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 458)
  AUTHORS   Cohen-Solal,M.M., Authier,B., deRiel,J.K., Murnane,M.J. and
            Forget,B.G.
  TITLE     Cloning and nucleotide sequence analysis of human embryonic
            zeta-globin cDNA
  JOURNAL   DNA 1 (4), 355-363 (1982)
  MEDLINE   83182021
   PUBMED   6963223
COMMENT     Original source text: Human erythroleukemia cell line K562, cDNA to
            mRNA, clones 1 (1g7-8), 2 (4p7-7), and 3 (5a3-3).
```

GenBank - champs 1

`LOCUS`

- 1–10 caractères alphanumériques ; jadis l'identificateur de la séquence (p.e. l'abbreviation du gène), préservée pour compatibilité seulement.
- longueur et type de la séquence (DNA, mRNA, tRNA, rRNA)
- code de la division (p.e. PRI) et date de dernière modification.

`DEFINITION` «sommaire» de la séquence : espèce et le nom de la séq

`ACCESSION` nombre d'accession : clé dans la base de donnée. Identificateur unique parmi les BDs. Forme AA999999. L'accno est généré automatiquement lors de la soumission d'une séquence à la BD.

GenBank - champs 2

`KEYWORDS` et `SOURCE` : moins d'importance (pour nous)

`VERSION` donne ⟨accno⟩.⟨version⟩ et gi : identificateur de GenInfo. Ce sont des identificateurs des *séquences* (qui peut changer pour le même accno).

`REFERENCE`

GenBank — exemple cont.

```
FEATURES             Location/Qualifiers
     source          1..458
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /map="16p13.3"
     gene            1..458
                     /gene="HBZ"
     CDS             30..458
                     /gene="HBZ"
                     /note="zeta hemoglobin"
                     /codon_start=1
                     /protein_id="AAA61306.1"
                     /db_xref="GI:340392"
                     /db_xref="GDB:G00-119-302"
                     /translation="MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYF
                     PHFDLHPGSAQLRAHGSKVVAAVGDAVKSIDDIGGALSKLSELHAYILRVDPVNFKLL
                     SHCLLVTLAARFPADFTAEAHAAWDKFLSVVSSVLTEKYR"
BASE COUNT       80 a    173 c    127 g     78 t
ORIGIN      79 bp upstream of BglII site.
        1 actccagtgc agctgcccac cctgccgcca tgtctctgac caagactgag aggaccatca
          .......
      421 cggtcgtatc ctctgtcctg accgagaagt accgctga
//
```

GenBank - champs 3

`FEATURES` annotation de la séquence : un «feature» comprend un mot-clé, sa position, et des qualifieurs

position : sous-mot [p.e., `2..280`], entre deux bases [p.e., `91^92`], ...,
et opérations : `complement`(·), `join`(·,...,·)

mots-clé :
- `source` information taxonomique
- `CDS` partie traduite en une séquence protéique
- `exon`, `intron`, `gene`
- `repeat_region`
- ...

Exemple : U96726

GenBank - entrées virtuelles

Exemple : U00089

```
LOCUS       U00089                816394 bp    DNA     circular CON 06-DEC-2002
DEFINITION  Mycoplasma pneumoniae M129, complete genome.
...
CONTIG      join(AE000016.2:1..19313,AE000015.2:59..17535,AE000014.2:22..12521,
            AE000013.2:53..10328,AE000012.2:59..10228,AE000011.2:59..15387,
            ... [plusieurs lignes]
            AE000019.2:59..10270,AE000018.2:59..11147,AE000017.2:62..15963)
//
```

FASTA

Un autre format très répandu, utilisé originalement par les logiciels du package FASTA.

```
>CRA|agCP11170 /len=264 /protein_uid=197000044174854 /org=Anopheles_gambiae
ATSFTMPQNEYIERHIKLYGRRLDYEERKRKREAREPKKRAAMARKLRGMKAKLFQKQRR
NEKIQMKRKIQAHEEKKVKKTTEKVEDGALPPYLMDRGIQSNAKVLSNMIKQKRKEKAGK
WDVPIPKVRAQADAEVFKVIRSGKTKRKAWKRMVTKVTYVGENFTRKPPKYERFIRPMAL
RMNKAHVTHPELKATFHLPIIGVKKNPSSPMYTSLGVITKGTVIEVNISELGLVTQSGKV
VWGKYAQVTNNPENDGCINAVLLV
>gi|30697195|ref|NP_200732.2|gnl|TIGR|At5g59240 structural constituent of ribosome [A. thaliana]
MGISRDSIHKRRATGGKQKMWRKKRKYELGRQPANTKLSSNKTVRRIRVRGGNVKWRALR
LDTGNFSWGSEAVTRKTRILDVAYNASNNELVRTQTLVKSAIVQVDAAPFKQGYLQHYGV
DIGRKKKGEAVTTEEVKKSNHVQRKLEMRQEGRALDSHLEEQFSSGRLLACIASRPGQCG
RADGYILEGKELEFYMKKLQKKKGKNAGAA
...
```

defline

une ou plusieurs séquences (ADN ou protéine)

en-tête pour chaque séquence (ligne `>...`)

syntaxe souvent utilisé : `/propriété=valeur`

defline : références à des banques de séquences

fomat général : ⟨db⟩ | ⟨ident⟩

NCBI

GenBank «flatfile» généré automatiquement à partir des bases de données.

Entrez : interface integré : recherche par identificateurs, mots clés, auteurs, etc.

BLAST : famille d'outils pour trouver des occurrences inexactes d'une séquence S dans le «texte» T

choix de T : `nr`, `est`, `month`, etc.