République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A. Mira de Bejaia
Faculté des Sciences de la Nature et de la Vie
Département de Biologie Physico-Chimique

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

INTRODUCTION A LA BIOINFORMATIQUE APPLIQUEE A LA GENOMIQUE

Présenté par Dr AIT ALI Djida

Année universitaire 2020-2021

I. Introduction

I.1 Concept de bioinformatique

Lors de sa création, la bioinformatique correspondait à l'utilisation de l'informatique pour stocker et analyser les données de la biologie moléculaire. Cette définition originale a maintenant été étendue et le terme bioinformatique est souvent associé à l'utilisation de l'informatique pour résoudre les problèmes scientifiques posés par la biologie dans son ensemble (**figure 1**). Il s'agit dans tous les cas d'un champ de recherche multidisciplinaire qui associe informaticiens, mathématiciens, physiciens et biologistes.

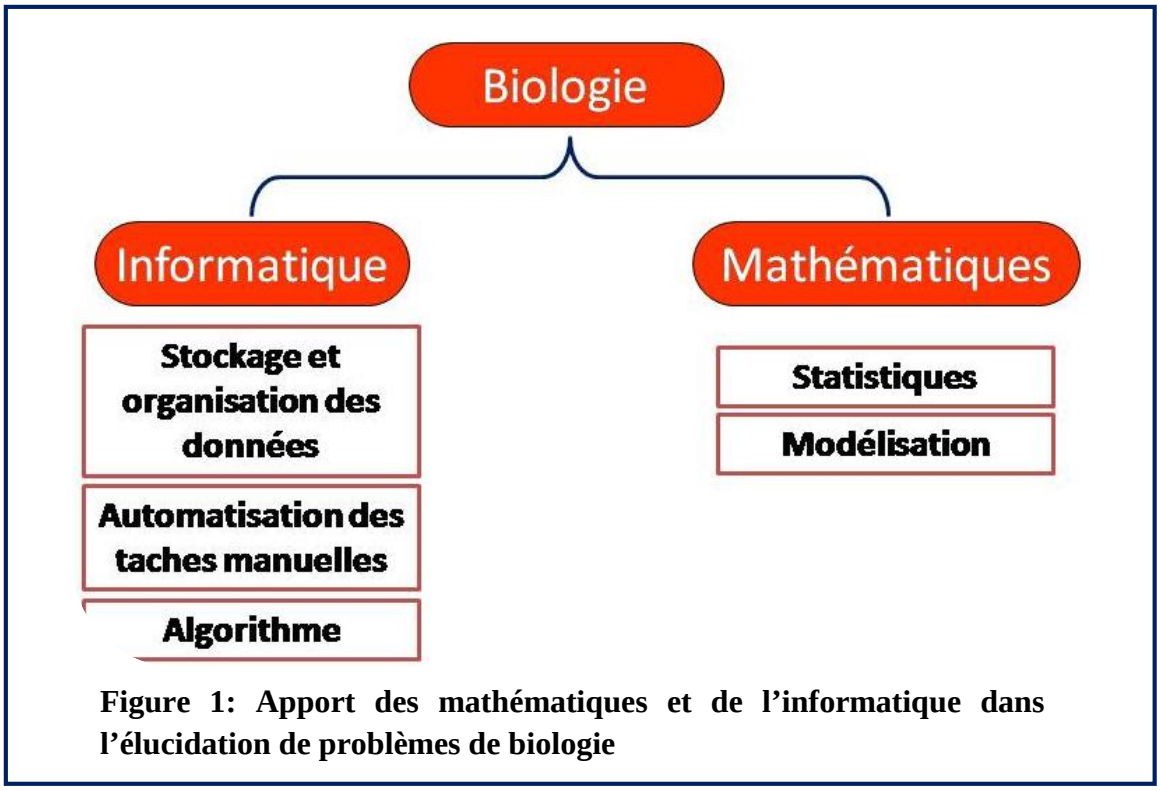


Figure 1: Apport des mathématiques et de l'informatique dans l'élucidation de problèmes de biologie

La bioinformatique est ainsi une discipline relativement nouvelle, qui évolue en fonction des nouveaux problèmes posés par la biologie moléculaire. En effet, avec le développement de la génétique et des nouvelles technologies à très haut débit, nous faisons actuellement face à la production de données à un niveau encore jamais atteints. En effet, il est aujourd'hui démontré que les données produites par les technologies de séquençage à haut débit seront plus importantes que tout ce qui a été produit dans le passe y compris le web lui-même. Les scientifiques font donc face à de multiples challenges tant pour le stockage de ces données (les nouvelles plateformes de séquençage peuvent produire jusqu'a 0,1 téraoctets de données par heure) que pour leur analyse.

En anglais on distingue deux termes:

- ***«Bioinformatics»*** Analyse "*in silico*" de l'information biologique contenue dans les séquences nucléiques et protéiques
- **«*Computational biology*»** Ensemble de méthodes et de logiciels qui permettent de gérer, manipuler, traiter et analyser les données biologiques

La bioinformatique est donc constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation de l'information génétique (séquences) et structurale (repliement 3-D). C'est le décryptage de la "bioinformation" ou *computational biology*. La bioinformatique est donc une branche théorique de la biologie. Son but, est d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories), d'énoncer des hypothèses généralisatrices (ex.: comment les protéines se replient ou comment les espèces évoluent), et de formuler des prédictions (ex.: localiser ou prédire la fonction d'un gène).

La bioinformatique permet notamment de:

- Formaliser des problèmes de biologie moléculaire;
- Développer des outils formels;
- Analyser les données;
- Prédire des résultats biologiques.

Cette discipline s'applique à tout type de données biologiques, en particulier moléculaires:

- Les séquences d'ADN et de protéines
- Les structures d'ARN et de protéines
- Les contenus en gènes des génomes
- Les puces à ADN (microarrays)
- Les réseaux d'interactions entre protéines
- Les réseaux métaboliques
- Les arbres de phylogénie

Elle permet ainsi de:

- Faire avancer les connaissances en biologie, en génétique humaine, en théorie de l'évolution
- Aider à la conception de médicaments
- Comprendre les maladies complexes

I.2 La Bioinformatique, une histoire

Voici une brève promenade historique le long de quelques évènements biologiques ou informatiques :

1646: Blaise Pascal invente une machine ("La Pascaline") capable d'effectuer des additions et des soustractions afin d'aider son père, collecteur d'impôts à Rouen.

1673: Gottfried Wilhelm von Leibniz construit une machine effectuant automatiquement les additions, soustractions, multiplications et les divisions.

1812: Charles Babbage, professeur de mathématiques, réalise les plans d'une machine capable d'exécuter n'importe quelle séquence de calculs au moyen d'une cinquantaine de roues dentées qui étaient activées grâce à des instructions lues sur une carte perforée.

1840: Collaboratrice de Charles Babbage et fille du poète Lord Byron, Ada Lovelace, mathématicienne, définit le principe des itérations successives dans l'exécution d'une opération. En l'honneur du mathématicien Arabe Al Khowarizmi, elle nomme le processus logique d'exécution d'un programme : algorithme.

1854: George Boole pose les axiomes et règles de l'algèbre booléenne, fondement des ordinateurs à arithmétique binaire.

1866: Gregor Mendel publie ses lois de l'hérédité à partie d'études menées chez le Pois.

1896 : Herman Hollerith crée la Tabulating machine et fonde une compagnie, qui deviendra IBM.

1901: De Vries redécouvre expérimentalement les lois de Mendel et publie "La théorie de la mutation".

1903: Walter S. Sutton (1903) et Boveri (1904) proposent pour la première fois d'associer les gènes au chromosome qui deviennent ainsi supports de l'hérédité.

1909: Wilhem Johannsen dénomme "gènes" les particules de l'hérédité proposées par Mendel puis redécouvertes par de Vries.

- Archibald Garrod propose la relation un gène-une enzyme à partir de l'étude d'une anomalie métabolique humaine: l'alcaptonurie (déficit en acide homogentisique-oxydase sur la voie du catabolisme de la tyrosine).

1913: Thomas Morgan et Alfred Sturtevant publient la première carte génétique du chromosome X avec la position respective de 3 gènes évaluée par le pourcentage de recombinaison (phénomène de crossing-over).

1915: Thomas Morgan publie avec Sturtevant, Muller et Bridge: "Le mécanisme de l'hérédité mendélienne".

1928: Fred Griffith fait les premières expériences de la transformation bactérienne.

1930: Georges Stibitz construit un additionneur binaire, appelée "Calculateur de Nombres Complexes", en s'appuyant sur les idées de Georges Boole.

1931: Konrad Zuse construit, le Z1 : premier calculateur digital électromécanique.

1935: Max Delbrück étudie le gène par le biais de l'effet induit par des rayonnements sur celui-ci. Il fonde le Groupe du phage, avec Salvador Luria et Alfred Hershey six ans plus tard.

1936: Alan Turing définit le concept de la machine de Turing et de là les notions de fonctions calculables.

1940: Alan Turing parvient à décrypter le code Enigma utilisé par l'Amirauté du Reich pour communiquer avec ses sous-marins sillonnant l'Atlantique.

1941: George Wells Beadle et Edward Tatum établissent la relation un "gène-une enzyme" chez Neurospora crassa.

1944: Oswald Avery démontre avec Colin McLeod et McLyn McCarthy que l'ADN transporte l'information génétique responsable de la transformation bactérienne.
- Erwin Schrödinger introduit la notion de programme et de code génétique.
- Howard Aiken termine la construction du Mark I: 1er ordinateur électronique à programme interne (à registre).

1946: L'annonce de l'ENIAC (*Electronic Numerical Integrator and Computer*) par J. Presper Eckert, marque le début de l'histoire moderne des calculateurs.

1947: Le DOE (agence fédérale responsable des programmes nucléaires aux Etats-Unis) s'engage dans les recherches génétiques.
- John Mauchly, J.P. Eckert, et John von Neumann travaillent à la conception d'un ordinateur électronique, l'EDVAC (Electronic Discret VAriable Computer) : 1er calculateur à programme enregistré. C'est le descendant direct de l'ENIAC (capacité mémoire est de 1024 mots de 44 bits).

1948: Claude Shannon publie "Une théorie mathématique de la communication" et est à l'origine de la théorie de l'information).

1949: John Mauchly présente "*Short Order Code*", le premier langage de programmation. EDSAC (*Electronic Delay Storage Automatic Computer*): 1er ordinateur numérique et électronique basé sur l'architecture de John von Neumann.

1950: Alan Turing publie le Test de Turing, pour définir l'IA (intelligence artificielle) d'une machine.

1952: Alfred Day Hershey et Chase démontrent que les bactériophages injectent leur ADN dans les cellules hôtes (corrélation entre l'ADN et l'information génétique).

1953: James Watson, Francis Crick et Maurice Wilkins (prix Nobel) découvrent la structure en double hélice de l'ADN.
- Début de l'IBM 650, le premier ordinateur "commercial".

1956: Frédérick Sanger établit la séquence en acides aminés de l'insuline.
- Vernon Ingram montre qu'une mutation liée à une altération héréditaire de l'hémoglobine se traduit par un changement d'un unique acide aminé dans la protéine.

1960: DEC présente le PDP1, premier ordinateur commercial avec écran/clavier.

1961: Marshall Nirenberg et J. Heinrich Matthaei déchiffrent le code génétique.

1965: Jacques Monod, François Jacob et André Wolf (prix Nobel) découvrent les mécanismes de la régulation génétique impliqués dans le dogme central de la biologie moléculaire, énoncé initialement par Crick.

1970: Programme d'alignement global de séquences (algorithme de Needleman & Wunsch).

1971: Annonce du microprocesseur INTEL 4004 : 1er microprocesseur.

1972: Clonage de fragments d'un plasmide bactérien dans le génome du virus SV40 (Paul Berg, David Jackson, Robert Symons).

1973: Découverte des enzymes de restriction.
- Obtention d'une méthode fiable de transfection (introduction d'un ADN étranger) des cellules eucaryotes grâce à un virus (vecteur). (Franck Graham et Alex Van der Eb).
- Développement de l'ALTO de Xerox suite aux recherches démarrées en 1970. Ce prototype, pensé pour devenir le bureau du futur, est le premier à introduire l'idée de fenêtres et d'icônes que l'on peut gérer grâce à une souris. Il ne sera introduit sur le marché qu'en 1981 sous le nom de Star 8010 qui connaîtra un échec commercial total.

1977: Frédérick Sanger met au point la méthode de Sanger pour établir le séquençage. Premier ensemble de programmes sur l'analyse des séquences (Staden).

- Création d'Apple Computer (Apple II) et de Microsoft.

1978: Mutagenèse dirigée. (Michael Smith)
- Séquençage du premier génome à ADN, le bactériophage phiX174 (5386pb) (Frederick Sanger)

1980: David Botstein et Ronald Davis introduisent les marqueurs moléculaires, notamment, les RFLP.
- Découverte de la technique de FISH (hybridation *in situ* sur chromosome), technique notamment utile dans la construction des banques génomiques (identification d'un fragment d'ADN sur un chromosome).
- Création de la banque EMBL : banque européenne généraliste de séquences nucléiques créée à Heidelberg et financée par l'EMBO (*European Moleculary Biology Organisation*).Elle est aujourd'hui diffusée par l'EBI (*European Bioinformatics Institute*, Cambridge, GB).

1981: Programme d'alignement local de séquences (algorithme de Smith et Waterman).
- Naissance du 1er animal transgénique (une souris).
- Découverte des oncogènes humains.

1982: Création de la banque Genbank : banque américaine généraliste de séquences nucléiques créée par la société IntelliGenetics et diffusée aujourd'hui par le NCBI (*National Center for Biotechnology Information*).

1984: Développement de la réaction de polymérisation en chaîne par Mullis de la PCR: outil devenu indispensable tant en recherche appliquée que fondamentale : séquençage génomique et cartographie, diagnostic génétique, analyse de l'expression des gènes.
- Création de la banque NBRF : banque américaine généraliste de séquences protéiques créée par la NBRF (*National Biomedical Research Foundation*).

1985: Programme FASTA (Fast Alignment, Pearson-Lipman) : recherche rapide d'alignements locaux dans unebanque.

1986: Création de la banque DDBJ : banque japonaise généraliste de séquences nucléiques créée par le NIG (*National Institute of Genetics*, Japon).
- Création de la banque SwissProt: banque généraliste de séquences protéiques créée à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (*via* ExPASy, *Expert Protein Analysis System*) et l'EBI.

1987: Réalisation et commercialisation du premier séquenceur automatisé par la société Applied Biosystems (Californie).

- Mise au point d'un nouveau vecteur: le YAC (*Yeast Artificial Chromosome*), premier vecteur permettant de cloner des fragments d'ADN 20 fois plus grands que les plasmides utilisés jusqu'alors.
- Publication de la 1ère carte génétique du génome humain.
- Apparition de la technologie des puces à ADN.

1988: Création du projet HUGO (*Human Genome Organization*) pour coordonner les efforts de cartographie et de séquençage entrepris dans le monde et éviter les doublons.

1989: INTERNET succède à ARPANET et BITNET.
- Découverte des marqueurs microsatellites.
- Découverte du système double hybride permettant d'étudier dans des cellules de levure (ou d'Escherichia Coli) l'interaction entre deux protéines hybrides fusionnées à des facteurs de transcription.

1990: Programme Blast: recherche rapide d'alignements locaux dans une banque.

1996: Séquençage du 1er génome eucaryote, Saccharomyces cerevisiae (12 Mb).

1998: Séquençage du 1er organisme pluricellulaire, Caenorhabditis elegans (100 Mb).

2000: Séquençage du 1er génome de plante, Arabidopsis thaliana.

2001: Annonce du décryptage presque complet du génome humain.

I.3 Les applications de la Bioinformatique

La bioinformatique est non seulement devenue essentielle pour la recherche fondamentale en génomique et en biologie moléculaire, mais elle a également un impact majeur sur de nombreux domaines de la biotechnologie et des sciences biomédicales. Elle a des applications dans la conception de médicaments basés sur la connaissance, l'analyse ADN médico-légale et la biotechnologie agricole.

Les études informatiques sur les interactions protéine-ligand fournissent une base rationnelle pour l'identification rapide de nouvelles pistes pour les médicaments synthétiques. La connaissance des structures tridimensionnelles des protéines permet de concevoir des molécules capables de se lier au site récepteur d'une protéine cible avec une grande affinité et spécificité. Cette approche basée sur l’informatique réduit considérablement le temps et le coût nécessaires au développement de médicaments plus puissants, avec moins d’effets secondaires et moins de toxicité que l’approche traditionnelle par essais et erreurs.

Dans le domaine de la santé, la génomique et la bioinformatique permettront de proposer une médecine personnalisée. Le séquençage génomique à haute vitesse, associé à une technologie informatique sophistiquée, permettra au médecin d'une clinique de séquencer rapidement le génome du patient, de détecter facilement les mutations potentiellement néfastes, de diagnostiquer rapidement et de traiter efficacement les maladies.

Les outils bioinformatiques sont également utilisés en agriculture. Les bases de données sur le génome des plantes et les analyses du profil d'expression génétique ont joué un rôle important dans la mise au point de nouvelles variétés de cultures plus productives et plus résistantes aux maladies

II. Le séquençage de génomes

II.1 Historique

La génomique est née avec l'apparition des techniques de séquençage de l'ARN, puis de l'ADN. En 1965, Holley et ses collaborateurs ont séquencé les deux premiers acides nucléiques de l'histoire, l'ARNt de l'alanine d'*Escherichia coli*, puis celui de la levure. C'est grâce à la capacité de purifier des ARNt particuliers et à la connaissance de RNAses dont la spécificité était connue que ces premiers séquençages ont pu avoir lieu. De plus, il a été possible de déterminer la structure secondaire de l'ARNt, puisque l'hybridation entre les bases était connue à l'époque. C'est en 1971 que la première molécule d'ADN a été séquencée. Cette molécule consistait en une séquence de12 nucléotides, soit la séquence des extrémités cohésives du phage lambda. Ces premières séquences ont été obtenues à l'aide de réactions chimiques spécifiques, comme la dépurination. Ces méthodes permettaient d'obtenir des séquences longues de 10 à 20 nucléotides.

En 1975, Sanger et Coulson ont introduit la méthode de terminaison des chaînes pour le séquençage de l'ADN. En 1977, Maxam et Gilbert ont conçu une méthode similaire à celle de Sanger, mais ils utilisaient plutôt des nucléotides qui ne permettaient pas l'élongation des chaînes. La même année, Sanger a introduit la méthode des didéoxynucléotides, méthode qui permettait de séquencer jusqu'à 100 nucléotides. Cette technique a permis le séquençage du génome du phage PhiX, aussi publié en 1977.

II.2 Préparation de la séquence cible

Ces trois méthodes utilisent des bases communes: le brin d'ADN à séquencer est extrait, amplifié, puis dénaturé (on sépare le double brin d'ADN en deux), et on utilise la polymérase, une enzyme dont le rôle est de dupliquer un élément. L'amplification, c'est à dire la réplication de la séquence, est réalisée à partir d'une technique appelée PCR ou "*Polymerase Chain Reaction*" (réaction de polymérisation en chaîne). Cela permet de dupliquer à plusieurs millions d'exemplaires un fragment d'ADN grâce à l'ADN polymérase.

Cette technique utilise des cycles de trois phases au cours desquelles chaque double brin est dénaturé et séparé de sa polymérase (qui est maintenu par des protéines). Une variation spécifique de la température permet aux nombreuses amorces présentes en suspension dans la solution de s'hybrider aux simples brins sans que ceux-ci ne s'hybrident entre eux (ce qui rendrait impossible l'accrochage de l'amorce).

Enfin, la polymérase assimile progressivement les nucléotides néoformés libres ajoutés également dans la solution jusqu'à l'obtention de doubles brins complets; c'est l'élongation. Le cycle, d'une durée maximum de quatre minutes, est répété en boucles de nombreuses fois (entre 35 et 45 par PCR) et double le nombre de brins à chaque fois (suite logique de puissance 2). La dénaturation (ou déshybridation), c'est à dire la séparation des deux brins composant la molécule d'ADN, est obtenue par élévation de température ce qui a pour effet de fragiliser les liaisons hydrogènes reliant les bases azotées (**figure 2**).

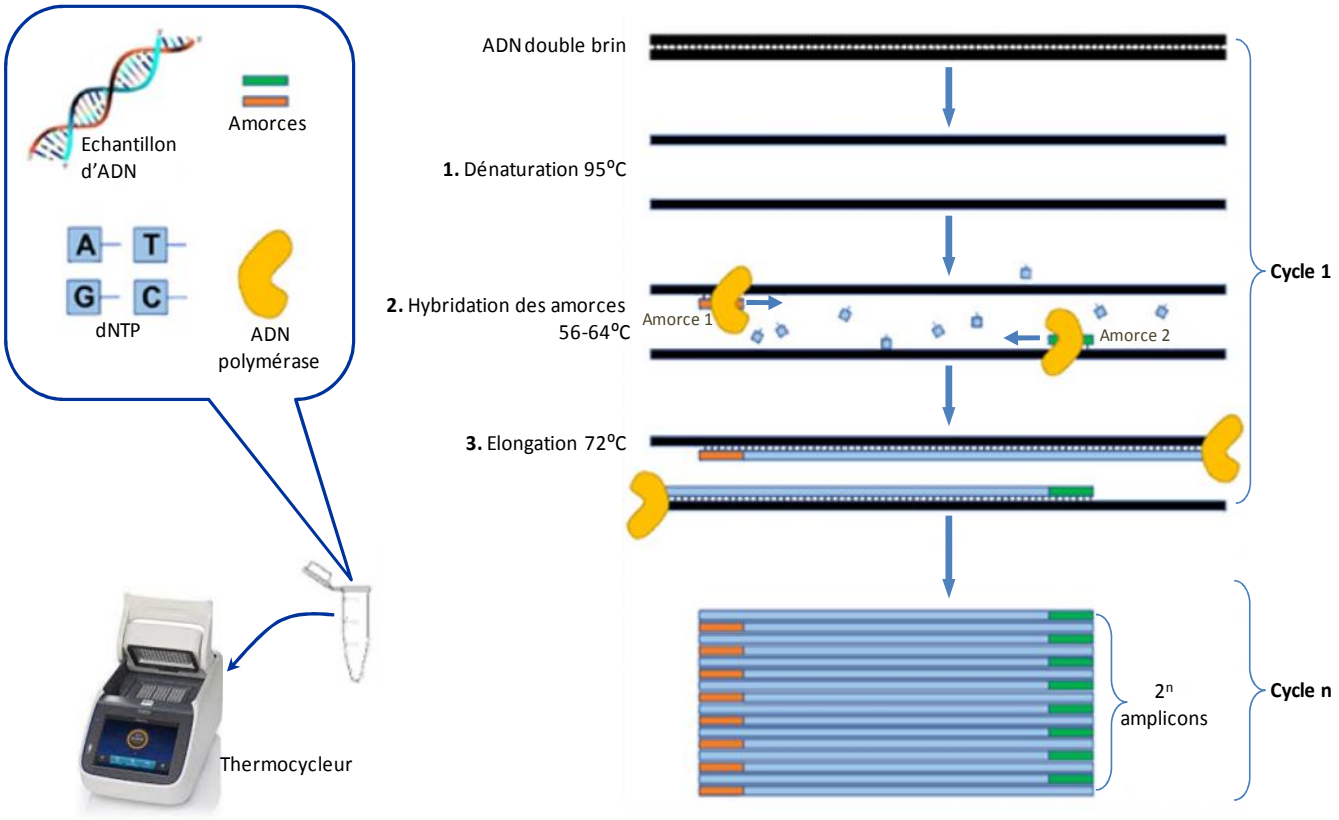


Figure 2: Etapes de la réaction de polymérisation en chaine (PCR)

II.3. Principe de la Méthode de Maxam & Gilbert

Cette méthode utilise des échantillons d'ADN double brin et ne nécessite donc pas le clonage de l'ADN dans un vecteur pour produire de l'ADN simple brin comme c'est le cas pour la méthode de Sanger. Cette méthode est basée sur une dégradation chimique de l'ADN et utilise les réactivités différentes des quatre bases A, T, G et C, pour réaliser des coupures sélectives. Les réactifs sont résumés dans le **tableau I**. En reconstituant l'ordre des coupures, on peut remonter à la séquence des nucléotides de l'ADN correspondant.

Tableau I : Principaux agents chimiques utilisés pour la méthode de séquençage Maxam et Gilbert

Bases	Altérations des bases	Suppression des bases	Coupure du brin
G	diméthylsulfate	piperidine	piperidine
A+G	acide	acide	piperidine
C+T	hydrazine	piperidine	piperidine
C	Hydrazine + alkali	piperidine	piperidine
A >C	alkali	piperidine	piperidine

Ce séquençage chimique peut-être décomposé en six étapes successives (**figure 3**):

1. Marquage. Les extrémités des deux brins d'ADN à séquencer sont marquées par un traceur radioactif (^{32}P). Cette réaction se fait en général au moyen d'ATP radioactif et de polynucléotide kinase.

2. Isolement du fragment d'ADN à séquencer. Celui-ci est séparé au moyen d'une électrophorèse sur un gel de polyacrylamide. Le fragment d'ADN est découpé du gel et récupéré par diffusion.

3. Séparation de brins. Les deux brins de chaque fragment d'ADN sont séparés par dénaturation thermique, puis purifiés par une nouvelle électrophorèse.

4. Modifications chimiques spécifiques. Les ADN simple-brin sont soumis à des réactions chimiques spécifiques des différents types de base. Walter Gilbert a mis au point plusieurs types de réactions spécifiques, effectuées en parallèle sur une fraction de chaque brin d'ADN marqué. Par exemple une pour les G (alkylation par le diméthyle sulfate), une pour G et les A (dépurination), une pour les C et une pour les C et les T (hydrolyse alkaline). Ces différentes réactions sont effectuées dans des conditions très ménagées, de sorte qu'en moyenne chaque molécule d'ADN ne porte que zéro ou une modification.

5. Coupure. Après ces réactions, l'ADN est clivé au niveau de la modification par réaction avec une base, la pipéridine.

6. Analyse. Pour chaque fragment, les produits des différentes réactions sont séparés par électrophorèse et analysés pour reconstituer la séquence de l'ADN. Cette analyse est analogue à celle que l'on effectue pour la méthode de Sanger.

La méthode de Maxam et Gilbert nécessite des réactifs chimiques toxiques et reste limitée quant à la taille des fragments d'ADN qu'elle permet d'analyser (<250 nucléotides). Moins facile à robotiser, cette technique est aujourd'hui très peu utilisée.

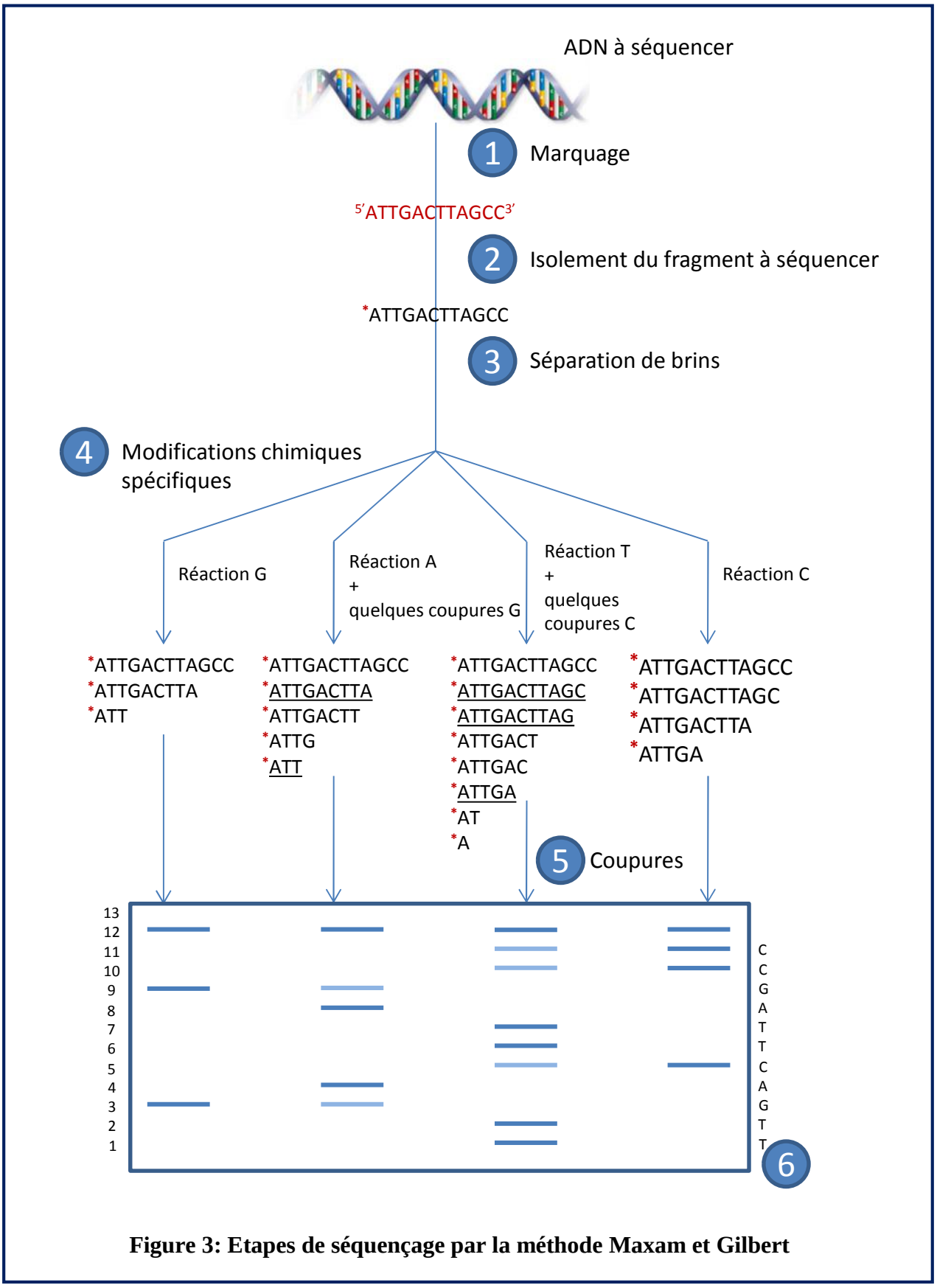


Figure 3: Etapes de séquençage par la méthode Maxam et Gilbert

II.4 Principe de la Méthode de Sanger

Cette méthode est basée sur l'interruption de la synthèse enzymatique d'un brin d'ADN complémentaire (arrêt d'élongation). L'ADN à séquencer est cloné et de nombreuses molécules d'ADN simple brin sont produites. Une courte amorce d'oligonucléotides (généralement synthétisée chimiquement et éventuellement marquée) est ajoutée à l'ADN. Le point de fixation de l'amorce sert de point de départ pour la synthèse du brin complémentaire.

Le recopiage d'un brin matrice par une ADN polymérase ADN dépendante (Taq) est initiée par la fixation d'un oligonucléotide spécifique (amorce), complémentaire du brin matrice. Cette ADN polymérase va permettre l'élongation d'un nouveau brin complémentaire du brin matrice dans le sens 5'→3'. L'ADN polymérase permet l'incorporation de nucléotides (dNTP: déoxynucléotides) libres présents dans le milieu réactionnel par la formation d'un pont phosphodiester entre le 3'OH de la chaîne et le 5' phosphate du dNTP suivant.

La réaction de Sanger repose sur l'incorporation aléatoire par cette ADN polymérase de didéoxynucléotides interrupteurs de chaîne (ddNTP) eux aussi présents dans le milieu réactionnel. Ces ddNTP diffèrent des dNTP par leur extrémité 3'. L'extrémité 3'OH des dNTP est remplacée par une extrémité 3'H (**figure 4**). Cette modification empêche la formation de la liaison phosphodiester entre le ddNTP incorporé dans la chaîne et le nucléotide suivant. L'allongement de la chaîne est alors interrompu.

Dans le milieu réactionnel il y a compétition entre les dNTP et les ddNTP. Le rapport spécifique ddNTP/dNTP et l'affinité de la Taq pour chaque nucléotide sont optimisés de telle façon qu'un ddNTP soit statistiquement incorporé à toutes les positions possibles. Une migration électrophorétique du produit de cette réaction de séquence sur un gel très résolutif (polyacrylamide) va séparer tous les fragments présents en fonction de leur masse moléculaire (taille). Les plus petits fragments vont migrer plus rapidement que les grands. La grande résolution de ce gel permet de distinguer des fragments différents entre eux d'une paire de base. L'identification du ddNTP présent à l'extrémité 3' de chaque fragment déterminera la séquence nucléotidique du brin matrice initial (**figure 5**).

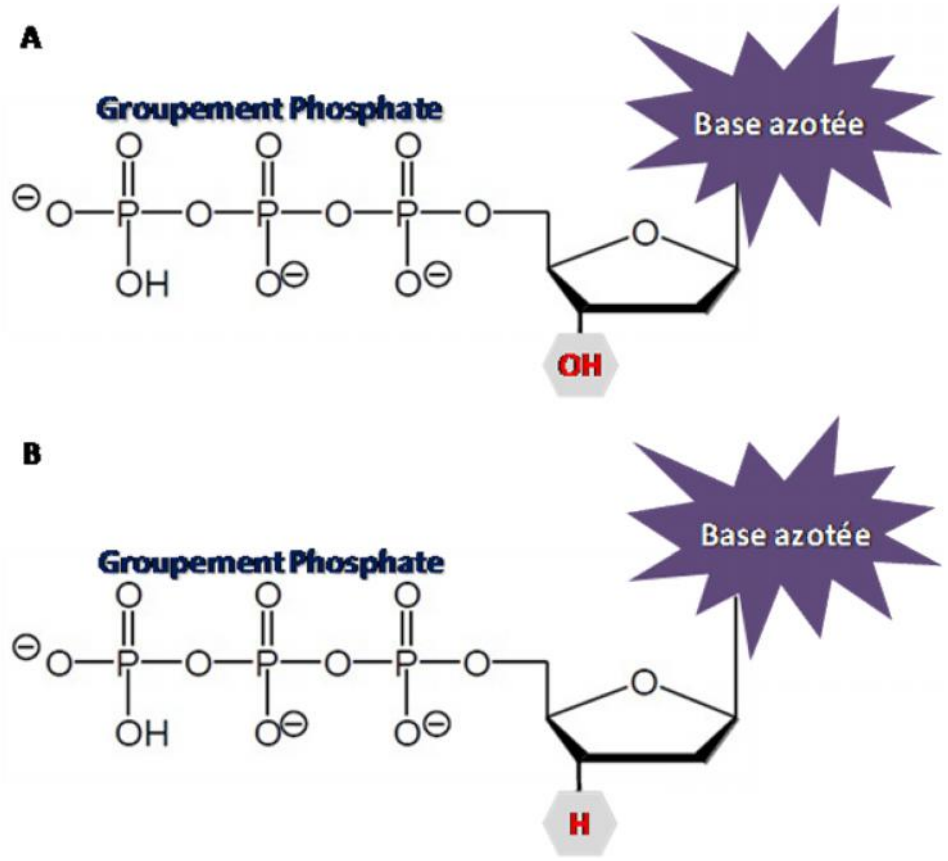


Figure 4: Les deux types de nucléotides triphosphates utilisés dans la méthode de Sanger. A. Groupement OH présent, désoxyribonucléotide (dNTP). **B.** Groupement OH absent didésoxyribonucléotide (ddNTP)

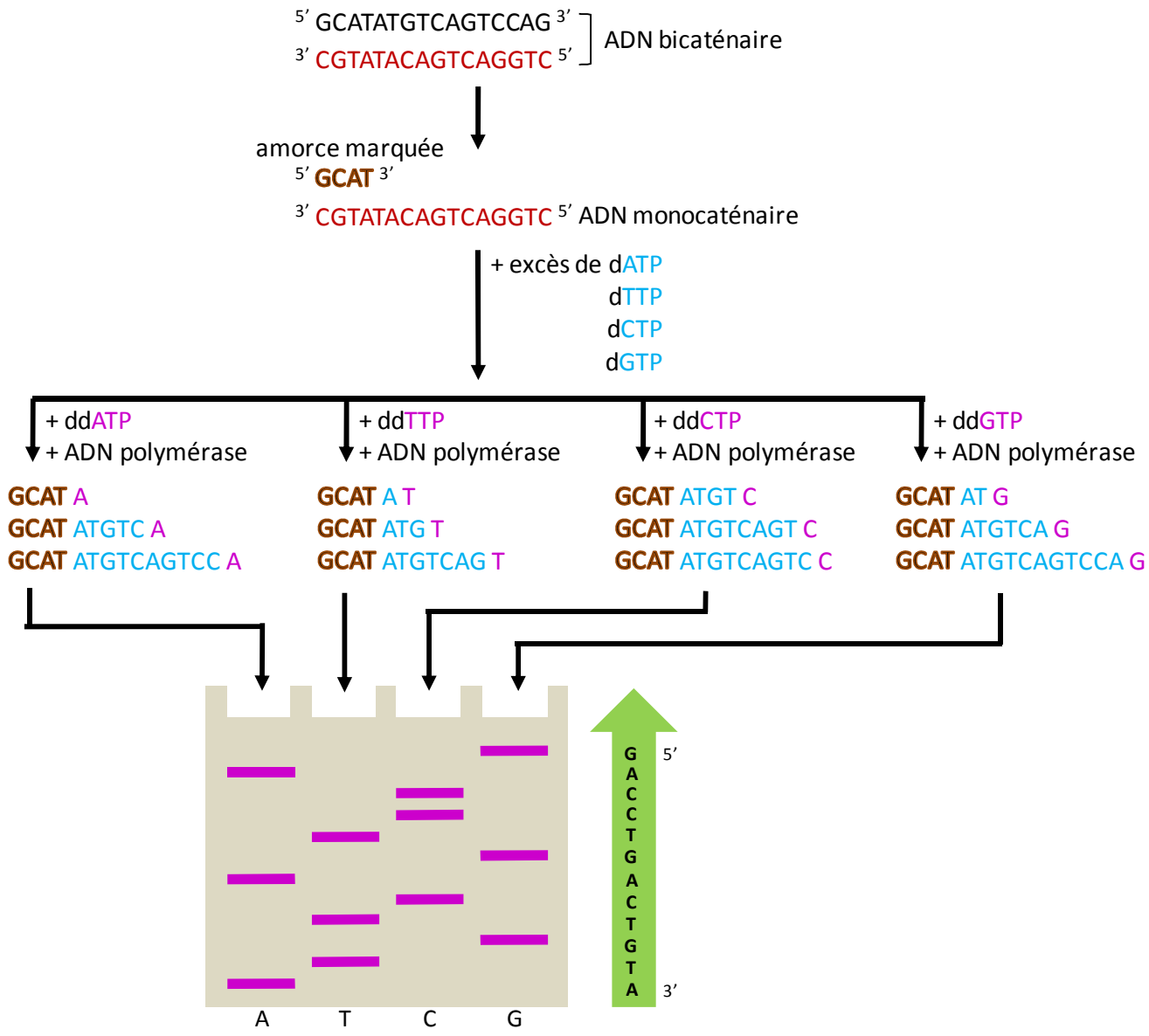


Figure 5:Réaction de séquençage par la méthode de Sanger

II.5 Automatisation du séquençage

Aujourd'hui, la plupart des séquençages sont réalisés par des séquenceurs industriels entièrement automatisés. Ceux-ci utilisent la technique de Sanger mais avec des méthodes de révélation différentes.

Les fragments d'ADN sont marqués par des marqueurs fluorescents; leur taille est ensuite déterminée par chromatographie ou électrophorèse assistée par ordinateur.

Avec ces techniques, on peut séquencer jusqu'à 1000 bases avec les meilleurs séquenceurs contre 200 à 300 *via* une méthode manuelle comme celles exposées ci-dessus. En effet, lors de l'électrophorèse manuelle, le nombre de bases est limité afin de ne pas rendre le

chromatogramme illisible par la surcharge des bandes et ainsi ne plus permettre une lecture horizontale.

C’est notamment grâce à la rapidité de ces appareils que le séquençage du génome humain fut réalisé en un temps record par rapport aux prévisions effectuées lors du démarrage du projet.

Plus adaptée à la robotisation, la méthode de Sanger a été largement développée grâce à plusieurs avancées technologiques importantes :

- La mise au point de vecteurs de séquençage adaptés, comme le phage M13 développé par Joachim Messing au début des années 1980.
- Le développement de la synthèse chimique automatisée des oligonucléotides qui sont utilisés comme amorces dans la synthèse.
- L'introduction de traceurs fluorescents à la place des marqueurs radioactifs utilisés initialement. Ce progrès a permis de sortir le séquençage des pièces confinées, réservées à l'usage des radioisotopes.
- L'adaptation de la technique PCR pour le séquençage.
- L'utilisation de séquenceurs automatiques de gènes.

L'utilisation de l'électrophorèse capillaire pour la séparation et l'analyse A côté des séquenceurs en gel plat, les séquenceurs capillaires ont apporté une plus grande automatisation dans les laboratoires de plus en plus demandeurs en séquençage de routine. La technique de Sanger est celle qui est mise en œuvre dans les premiers séquenceurs automatiques (le premier séquenceur automatisé a été mis au point en 1987, par la compagnie Applied Biosystem ABI).

En général l’automatisation requiert l’emploi:

- d'un système d'électrophorèse piloté par ordinateur,
- des marqueurs fluorescents de différentes couleurs qui sont révélés après excitation par un laser à l'aide d'une caméra CCD.
- des logicielles permettant l'analyse des signaux sortant de l'appareil et leur mise en forme sous forme de résultats (électrophorégramme et séquence).
- d'un robot passeur d'échantillon permettant d'enchaîner les échantillons les uns à la suite des autres (notamment passage de plaques de réaction à 96 puits (12x8)).

Les séquenceurs capillaires

Un séquenceur de gène à capillaire utilise des tubes capillaires de verre ayant quelques microns de diamètre, sur plusieurs dizaines de centimètres de longueur (30 à 50 cm en général),

pour réaliser la séparation des brins d'ADN durant l'électrophorèse. Avec généralement un nombre de capillaires multiples de 2 (2, 4, 8, 16...). On multiplie ainsi le nombre de migrations simultanées, ce qui permet de passer un plus grand nombre d'échantillons dans le même laps de temps.

Les instruments à capillaire les plus modernes de séquençage automatique de l'ADN sont capables de lire jusqu'à 384 échantillons d'un coup (1 *run*) et réaliser jusqu'à 24 *runs* en une journée. Ces instruments ne réalisent que la séparation des brins et la lecture des pics ; les réactions de séquençage, la purification et la suspension dans un tampon approprié doivent être réalisées séparément, de façon manuelle ou à l'aide d'un robot pipeter.

Les séquenceurs automatiques présentent de nombreux avantages: l'automatisation et l'utilisation d'une chromatographie au lieu d'une électrophorèse permet un gain de temps appréciable. Le coût de revient est bien moindre. De plus, ce type de séquenceurs permet de lire plusieurs centaines de nucléotides avec une très bonne qualité, jusqu'à 1000 à 1100 pb pour les appareils les plus performants.

II.6 Le pyroséquençage: méthode non Sanger de séquençage

Le pyroséquençage est de loin la technique non Sanger qui a connu le plus de succès. Cette technique de séquençage d'ADN introduite depuis1988, par Hyman *et al.*, et amélioré par un groupe suédois par introduction de la PCR. Il s'agit d'un séquençage par synthèse (*sequencing by sysnthesis,* SBS) et qui se caractérise par la révélation en temps réel de l'activité de l'ADN polymérase (*real time sequencing*) qui ajoute un seul nucléotide non fluorescent à la fois.

Principe du pyroséquençage

Le pyroséquençage se déroule en 5 étapes illustrées dans la **figure 6**:

Etape 1: Consiste à préparer le mélange réactionnel, avec les enzymes clefs et les différents substrats

Etape 2: Ici, les nucléotides ne sont pas ajoutés tous ensemble comme dans une réaction de séquençage normale mais l'un après l'autre. Si le nucléotide ajouté dans le milieu réactionnel correspond à celui attendu par la polymérase, il est incorporé dans le brin en cours de synthèse (d'élongation) et libère un pyrophosphate.

Etape 3:L'ATP sulfurylase vient alors transformer ce Pyrophophate (PPi) en ATP qui est alors utilisé, couplé à une Luciférine, par une Luciférase. On a alors production d'Oxyluciférine et d'un signal lumineux

Etape 4: L'Apyrase dégrade les nucléotides en surplus.

Etape 5: Le signale lumineux est capté par un capteur CCD (*Charge-Coupled Device*) puis reproduit sous forme d'un pic sur le pyrogramme. La hauteur de ce pic est fonction de l'intensité du signal lumineux, elle-même proportionnelle au nombre de nucléotides incorporés en même temps. On peut donc déduire la séquence à partir de la taille des pics obtenus. Par ailleurs, en cas de mélange de nucléotides à une même position (polymorphisme de séquence), la taille des pics permet d'avoir une quantification de la proportion de brins porteurs de l'un ou l'autre des nucléotides.

Avantages et limitations de la technique de pyroséquençage

Le pyroséquençage est une technique qui permet d'effectuer un séquençage rapide et à moindre coût qu'un séquençage par la méthode de Sanger. En effet, cette technique ne nécessite pas de clonage (donc gain de temps et d'argent), et permet une lecture directe de la séquence obtenue après le séquençage.

Toutefois, contrairement aux méthodes traditionnelles, le pyroséquençage ne peut séquencer que des brins d'une centaine de bases azotées (100 pb). Par conséquent, cette méthode n'est pas très courante. Elle est utilisée principalement pour détecter des mutations sur des séquences ciblées par comparaisons de brins.

Avec cette capacité de lecture le pyroséquençage trouve son application dans l'identification d'étiquettes des séquences (séquence tag), le minis-séquençage de SNP (*Single Nucleotid Polymorphism*) connu, et pour la cartographie relative des génomes par rapport à une séquence de référence.

Etape 1

Une amorce de séquençage est hybridée à la matrice d'ADN simple brin obtenu par PCR en présence:

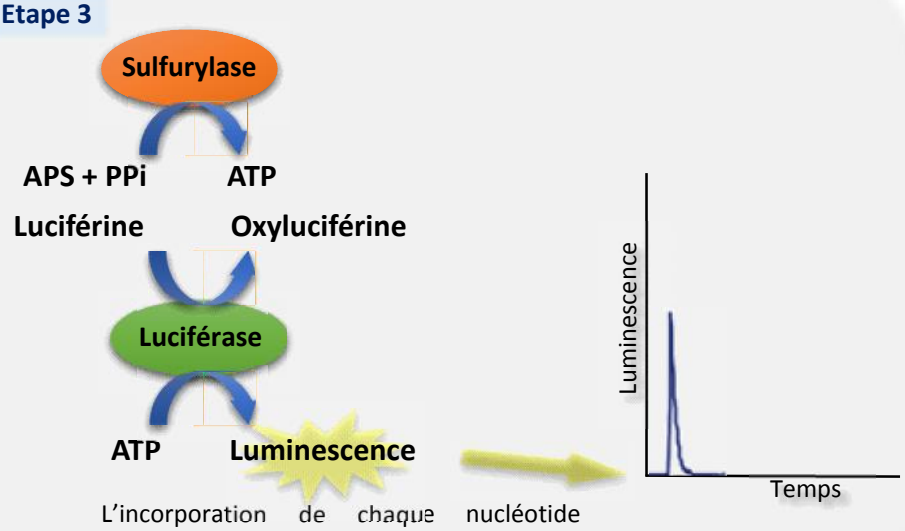
Enzymes ➔ ADN polymérase, ATP sulfurylase, Luciférase et Apyrase

Substrats ➔ Adénosine 5' phosphosulfate (APS) et Lucéférine

Etape 2

$(ADN)_n + dNTP \xrightarrow{\text{Polymérase}} (ADN)_{n+1} + PPi$

Etape 3



L'incorporation de chaque nucléotide génère une luminescence traduite par un pic dans le pyrogramme

Etape 4

$dNTP \xrightarrow{\text{Apyrase}} dNDP + dNMP + phosphate$

$ATP \xrightarrow{\text{Apyrase}} ADP + AMP + phosphate$

Etape 5

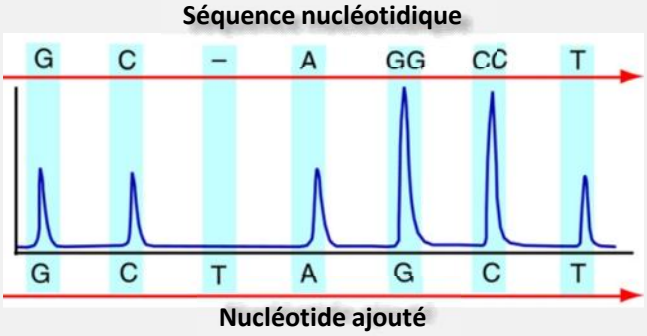


Figure 6: Etapes de déroulement du pyroséquençage

II.7 Le séquençage à haut débit (nouvelle génération de séquençage)

Depuis 2005, l'émergence de nouvelles technologies considérées comme la nouvelle ou seconde génération de séquençage (NGS, *next-generation sequencing*) a permis de séquencer avec des débits qui évoluent encore aujourd'hui de façon phénoménale.

Les trois (03) technologies de séquençage qui dominent actuellement le marché se démarquent les unes des autres par les chimies qui les constituent. Elles se décomposent cependant toutes en quatre (04) grandes étapes principales: la préparation des librairies qui contient une étape d'amplification par PCR, les cycles de réactions de séquençage, la prise d'image après chacun de ces cycles pour déterminer le nucléotide correspondant, puis l'analyse des données. Ces nouvelles générations de machines ont pour avantage leur capacité à analyser à analyser de grands génomes à haute résolution grâce à la parallélisation des réactions.

II.7.1 454/Roche

Le premier des séquenceurs de nouvelle génération a été commercialisé par 454 Life Sciences en 2005, depuis racheté par Roche, mais la plateforme est toujours connue sous le nom générique de 454.

La spécificité de cette technologie repose sur une PCR en émulsion pour l'amplification des fragments à séquencer. La PCR a lieu dans une microgoutte renfermant une microbille d'agarose en phase aqueuse, séparée des autres billes (plusieurs millions) par de l'huile. On obtient ainsi des copies d'un seul fragment d'ADN par bille (**figure 7**). Chacune des billes est ensuite déposée dans un des 1,6 millions de puits d'un support solide PTP (*PicoTiter Plate*).

Des réactions de pyroséquençage ont alors lieu à l'échelle du picolitre dans chaque puits: un flux de nucléotide (chaque nucléotide l'un après l'autre) traverse la PTP et lorsque l'un d'entre eux est incorporé par la polymérase, un pyrophosphate est libéré. Le reste de la réaction se déroule comme décrit précédemment (cf. Chap. II.6).

On obtient ainsi jusqu'à 900 Mégabytes de données en une dizaine d'heures, soit 15.000 fois plus qu'avec les méthodes de séquençage classiques. Cette technologie présente la plus grande taille de lecture (jusqu'à 700 pb par rapport à 100 pb à ses débuts) et sa grande précision la rend appropriée au séquençage *de novo.*

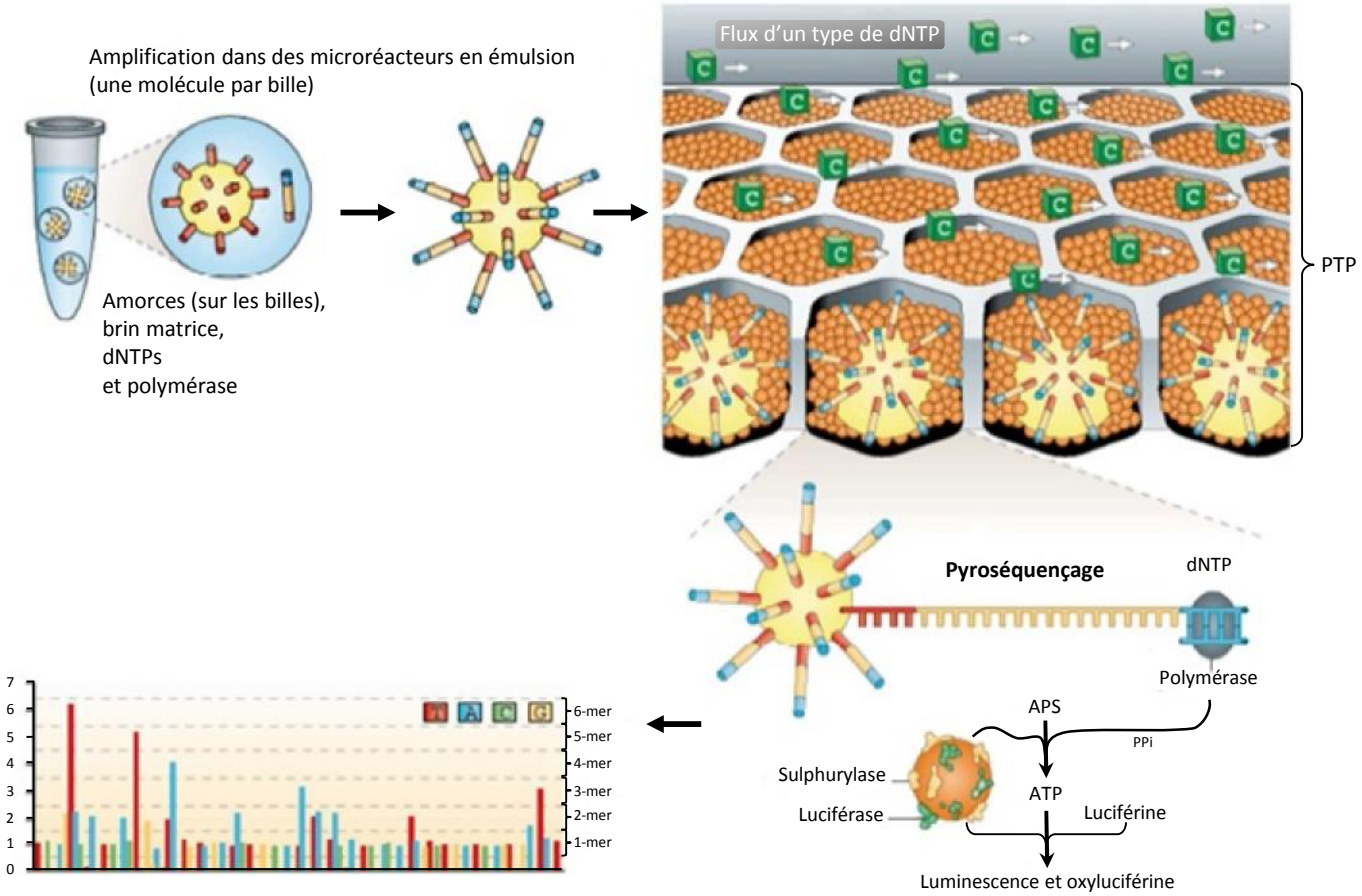


Figure 7: Séquençage haut débit par la technologie 454. PTP : PicoTiter Plate, PPi : pyrophosphate inorganique, APS : adénosine phosphosulphate, ATP : adénosine triphosphate

II.7.2 Solexa/Illumina

L'entreprise Solexa a commercialisée son premier séquenceur avec succès en 2006, le *Genome Analyzer* (GA). La spécificité de cette technologie repose sur une amplification en pont (*bridge* PCR) des fragments à séquencer.

La réaction se déroule sur une surface en verre appelée *flow cell* (FC), similaire à une lame de microscope, divisée en huit lignes (à l'origine, une ligne par échantillon). Les fragments de la librairie à séquencer possèdent des adaptateurs à leurs extrémités. Ceux-ci vont leur permettre de se fixer de façon aléatoire sur la FC, par hybridation sur les amorces qui en couvrent la surface (**figure 8**)

Un nouveau brin est alors synthétisé par une polymérase (**figure 8a**); il est fixé de façon covalente à la FC. Le brin d'origine est alors éliminé par dénaturation (**b**) et l'extrémité libre du brin restant s'hybride à une amorcé adjacente pour former un pont (**c**). La polymérase synthétise à nouveau le brin complémentaire pour former un pont d'ADN double brin (**d**) puis les deux copies sont libérées par dénaturation (**e**). Le cycle de dénaturation en pont (étapes **c** à **e**) recommence pour former un regroupement d'ADN clonal en une appelée *cluster* (**f**). Les brins anti-sens (correspondants aux amorces vertes) sont ensuite clivés (**g**); c'est la linéarisation.

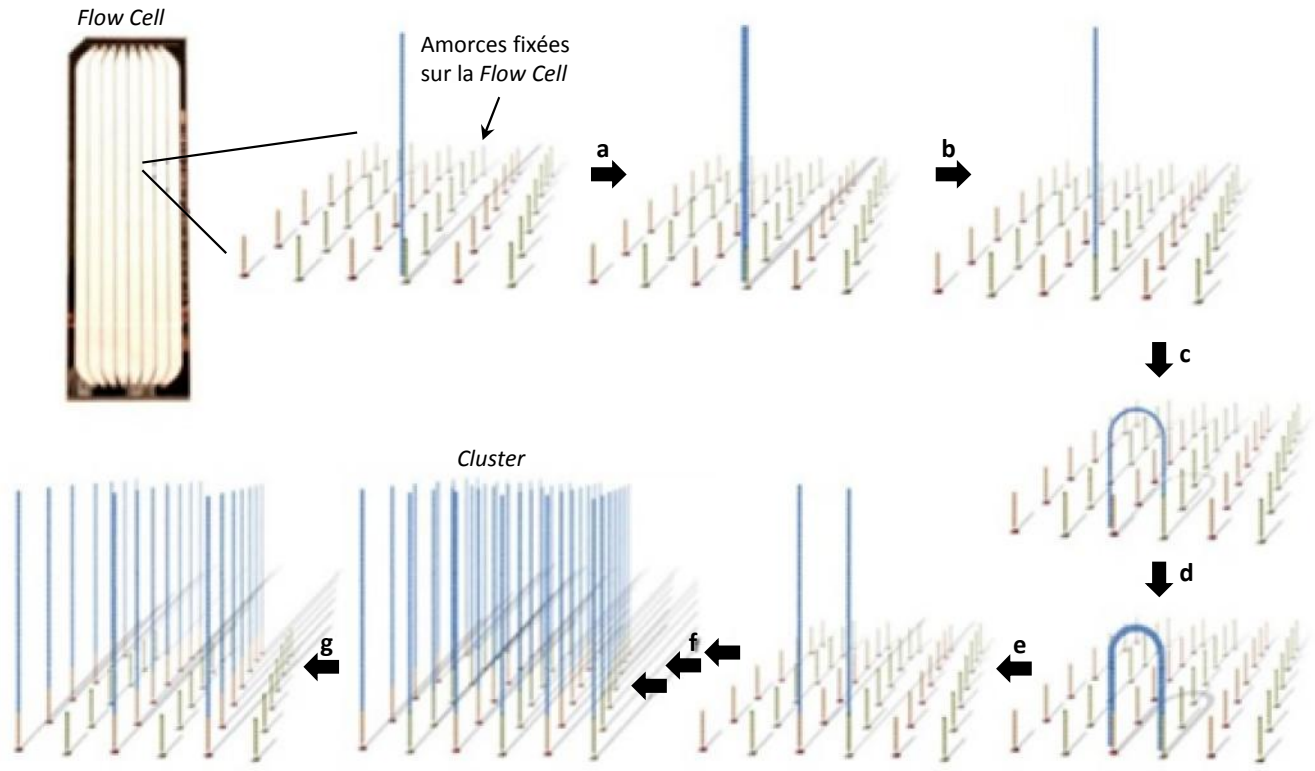


Figure 8: Amplification en pont de la technologie Solexa/Illumina. **a**) Synthèse du brin complémentaire par une polymérase. **b**) Dénaturation. **c**) Formation d'un pont. **d**) Synthèse du brin complémentaire par une polymérase. **e**) Dénaturation. **f**) Nombreux cycles c à e. **g**) Linéarisation.

L'extrémité 3' livre des fragments d'ADN est bloquée et l'amorce de séquençage s'y hybride (**figure 9**). Le séquençage s'effectue sur des centaines de millions de *clusters* simultanément, grâce à une chimie de terminateurs réversibles. Ainsi, des nucléotides bloqués marqués par fluorescence sont ajoutés, l'un d'entre eux est incorporé, la fluorescence est relevée puis le fluorophore et le bloqueur sont clivés permettant l'ajout d'un nouveau nucléotide. A chaque cycle d'incorporation une base peut être déterminée.

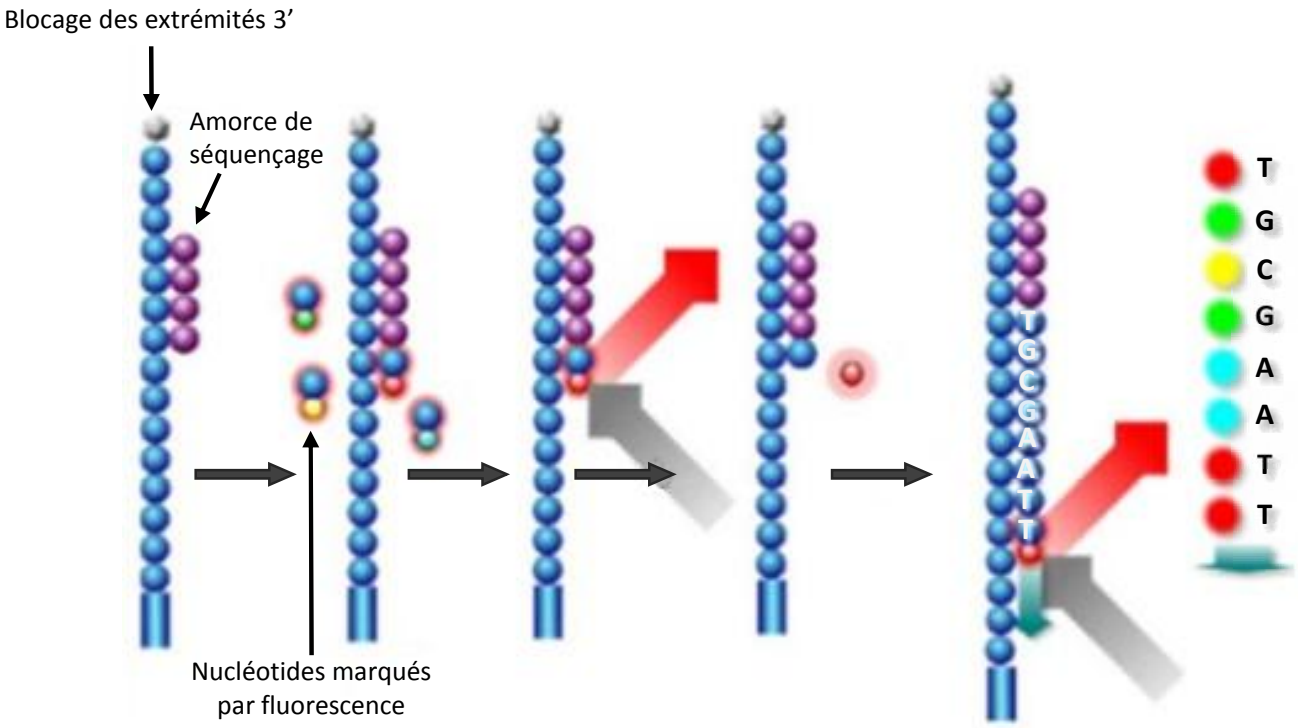


Figure 9: Détail de la technologie Solexa/Illumina.

Cette chimie a pour avantage de séquencer correctement les homopolymères. Toutefois, elle a pour inconvénient de lire peu de bases (36 à ses débuts, jusqu'à 150 bases de nos jours), ce qui le rend approprié à l'analyse de génomes dont on a une bonne annotation. La technologie Solexa/Illumina peut produire jusqu'à 600 Gigabites de données en une dizaine de jours, ce qui fait d'elle le leader sur le marché du séquençage.

II.7.3 SOLiD

Le SOLiD (*Sequencing by oligonucleotide Ligation and Detection*) a été la troisième plateforme de séquençage de nouvelle génération, commercialisée par *Applied Biosystems* (aujourd'hui *Life Technologies*) depuis 2007.

Cette technologie repose également sur une PCR en émulsion sur billes. Le séquençage ne s'effectue toutefois pas par synthèse comme sur les plateformes précédentes mais par ligation (**figure 10**).

Une amorce de séquençage universelle se fixe sur l'adaptateur puis des oligonucléotides dégénérés de huit (08) bases, marqués par fluorescence, sont ajoutés (**étape 1**). Dés que l'un d'entre eux correspond à la séquence adjacente à l'amorce, la ligase le fixe et de la fluorescence est émise, permettant d'identifier l'oligonucléotide fixé et d'interroger ainsi ses premières bases (**étapes 1, 2**). Un clivage chimique retire les bases 6 à 8 ainsi que le fluorophore puis les oligonuclèotides sont alors ajoutés à nouveau; on identifie les bases 6 à 7 du fragment à séquencer, puis dans un troisième temps les bases 11 et 12, et ainsi de suite jusqu'aux bases 31 et 32 (**étapes 3, 4**). Un deuxième cycle de ligation est alors entamé avec une amorce universelle se fixant en position n-1 et on identifie les bases en position 0 et 1, 5 et 6, et ainsi de suite jusqu'aux bases 30 et 31 (**étapes 5, 6**). Trois (03) nouveaux cycles de ligation sont effectués grâce à des amorces se fixant en n-2, n-3 puis n-4 (**étape 7**). Le nombre de cycles de ligation, détection et clivage détermine ainsi la longueur de lecture, de 35 bases dans le cas présenté ici à 75 bases.

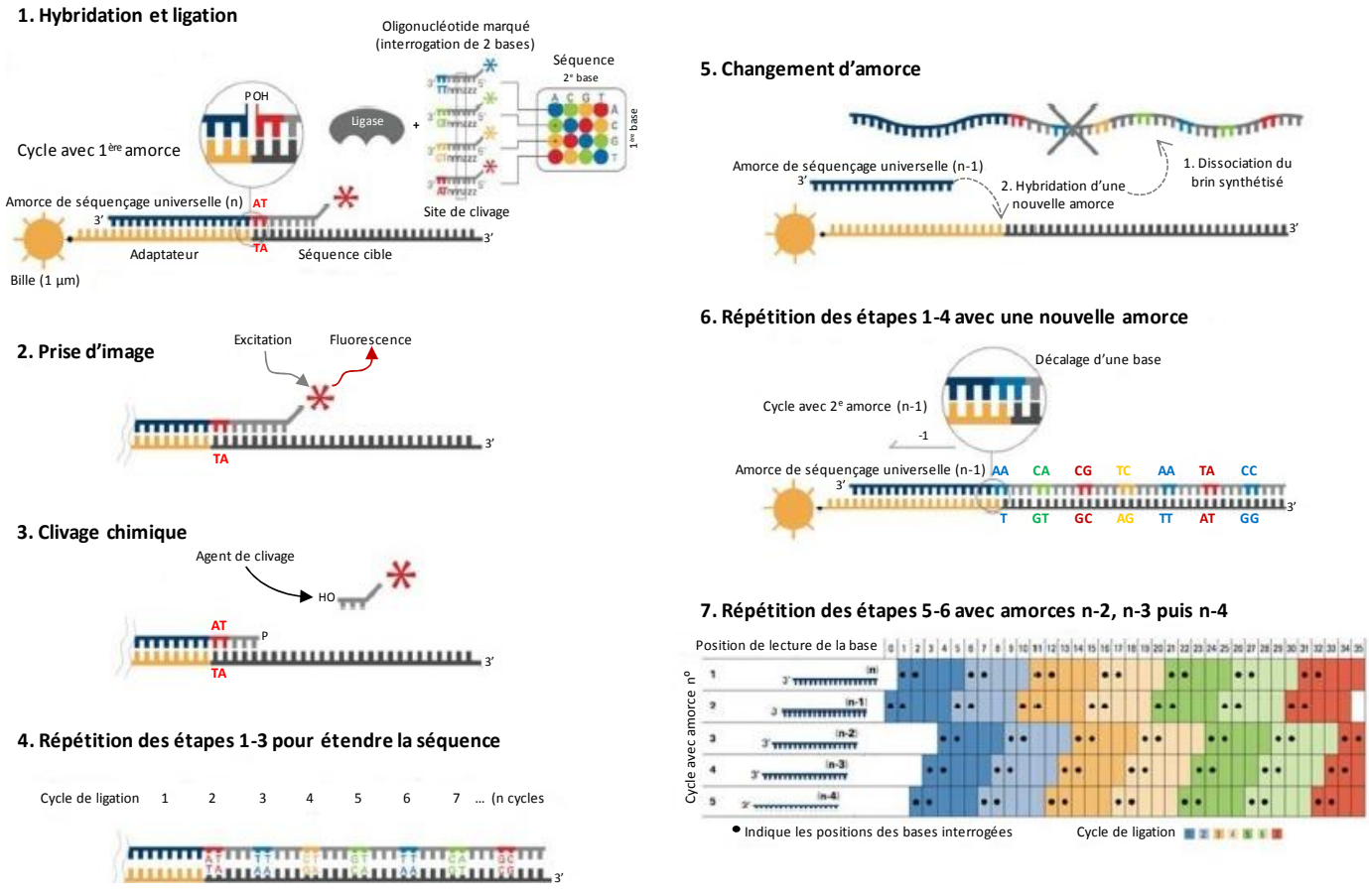


Figure 10: Etapes de séquençage par la technologie SOLiD.

Avec cette technologie, chaque base est lue deux fois, ce qui explique sa grande précision et qui la rend adaptée au reséquençage ou à l'analyse de polymorphismes. Néanmoins, la complexité de fonctionnement de cette plateforme de séquençage en est un inconvénient étant donné qu'elle implique un lourd travail d'analyse.

III Les Bases de données biologiques

Une base ou banque de données est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter leur utilisation (ajout, mise à jour, recherche et éventuellement analyse dans les systèmes les plus évolués) (**figure 11**).

L'ensemble des données relatives à un domaine particulier, sont organisées par traitement informatique et peuvent être accessible en ligne et à distance. Elles sont souvent stockées sous la forme d'un fichier texte formaté, respectant une disposition particulière.

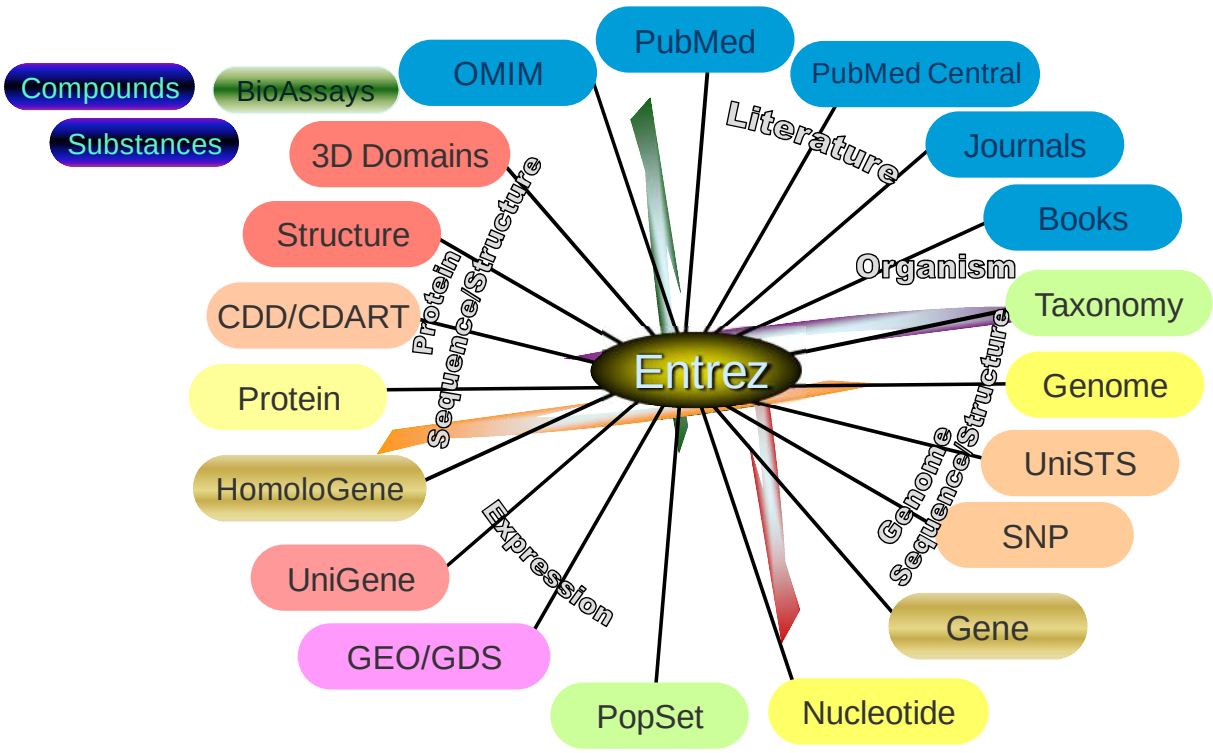


Figure 11: Exemple d'une base de données biologiques: NCBI (*National Center for Biotechnology Information*).

Il existe un grand nombre de banques ou bases de données d'intérêt biologique. Nous en distinguons deux types:

- ***Banques généralistes***: Celles-ci correspondent à une collecte des données la plus exhaustive possible et offrent finalement un ensemble plutôt hétérogène d'informations tels que:

- Banques de séquences nucléiques
- Banques de séquences nucléiques
- Banques de séquences protéiques
- Banques de structure 3D de macromolécules
- Banques d'articles scientifiques

Ce type de base de données présente comme avantage de tout permettre de consulter en une seule fois. Elles présentent toutefois des inconvénients tels qu'une difficulté à maintenir et à interroger.

- ***Banques spécialisées***: Elles correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe d'individus (ex. banques spécialisées pour un génome, banques de séquences d'immunologies, banques sur des séquences validées…)

- Avantage: facilité pour mettre à jour les données, vérifier leur intégrité, offre une interface adaptée.
- Inconvénients: ne cible pas toujours ce que l'on veut; toutes les banques possibles n'existent pas.

III.1 Les Banques de séquences biologiques

Il existe un grand nombre de bases de données d'intérêt biologique. Ici, nous nous limiterons à une présentation des principales banques de données publiques, basées sur la structure primaire des séquences nucléiques, et qui sont largement utilisées dans l'analyse informatique des séquences.

En effet, la séquence est l'élément central autour duquel les banques de données se sont constituées. Les séquences biologiques, dès qu'elles ont pu être établies, ont très tôt fait l'objet d'une compilation dans les banques de données. La première compilation de protéines est publiée en 1965 par Margaret Dayhoff: c'est l'*Atlas of Protein Sequences* qui contient alors 50 entrées. D'abord imprimé jusqu'en 1978, il fut ensuite proposé sous forme électronique.

C'est au début des années 80 que les premières banques de séquences sont apparues sous l'initiative de quelques équipes dont la première à l'initiative de Grantham et C. Gautier à Lyon. Très rapidement avec les évolutions techniques du séquençage, la collecte et la gestion des données ont nécessité une organisation plus conséquente. Plusieurs organismes ont pris en charge la production de telles bases de données.

Ainsi sont nées trois (03) banques nucléiques:

- **EMBL**: banque européenne créée en 1980 et financée par l'EMBO (*European Moleculary Biology Organization*), elle est aujourd'hui diffusée par l'EBI (*European Bioinformatics Institute, Cambridge, UK*).
- **GenBank**: créée en 1982 par la société *IntelliGenetics* et diffusée maintenant par le NCBI (*National Center for Biotechnology Information, US*).
- **DDBJ**: créée en 1986 et diffusée par le NIG (*National Institute of Genetics, Japan*)

Ces trois banques s'échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes : "The DDBJ/EMBL/GenBank*Feature Table Definition*", garantissant ainsi dans chacune d'elles un ensemble de données le plus complet possible Chaque enregistrement ou « entrée » correspond à une séquence nucléique.

Parallèlement, pour les protéines, deux banques principales ont été créées. La première, sous l'influence du *National Biomedical Research Foundation* (NBRF) à Washington, produit maintenant une association de données issues du MIPS (*Martinsried Institute for Protein Sequences*), de la base Japonaise JIPID (*Japan International Protein Information Database*) et des données propres de la NBRF. Elle se nomme la *Protein Identification Ressource* (PIR-NBRF). La deuxième, Swissprot a été constituée à l'Université de Genève à partir de 1986 et regroupe entre autres des séquences annotées de la PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL.

Dans la suite de ce document nous nous intéresserons particulièrement aux trois (03) banques de données nucléiques présentées ci-dessus.

III.2 Les Banques de séquences nucléiques

Comme mentionné ultérieurement les trois (03) banques nucléotidiques principales coexistent et coopèrent entre elles. Elles collectent des informations de séquences (ADN et ARN), associées ou non à une publication, par soumission directe des auteurs ce qui représente 95% de l'ensemble des données, mais également par balayage systématique de la littérature scientifique (principalement les brevets). Chaque enregistrement ou « entrée » correspond à une séquence nucléique. Afin d'identifier ces séquences, les différentes banques de données leur assignent des **Numéros d'Accession** (*Accession Number*) uniques au sein de leur collections respectives.

L'organisation de l'information suit un format général ou les données sont organisées séquentiellement. Les banques sont distribuées dans un fichier plat « *Flat file* » c'est-à-dire sous la forme d'un simple fichier texte (format .txt).

D'autre part, une fiche ou entrée dans une banque nucléotidique comporte deux (02) parties essentielles. La première contient des informations relatives à la séquence ou qu'on appelle **annotation**. Cette dernière est organisée sous forme de champs regroupant des informations de même type permettant ainsi de faciliter l'accès à l'information. La deuxième partie contient la **séquence** elle-même (Voir les exemples d'entrées dans les paragraphes suivant).

Afin d'éviter toute ambiguïté, les séquences sont représentés par leur codes IUPAC standard (**figure 12**). De plus toutes les séquences quelles soit ADN ou ARN sont écrites avec des T (Thymine) et sont toujours orientées de 5' à 3'.

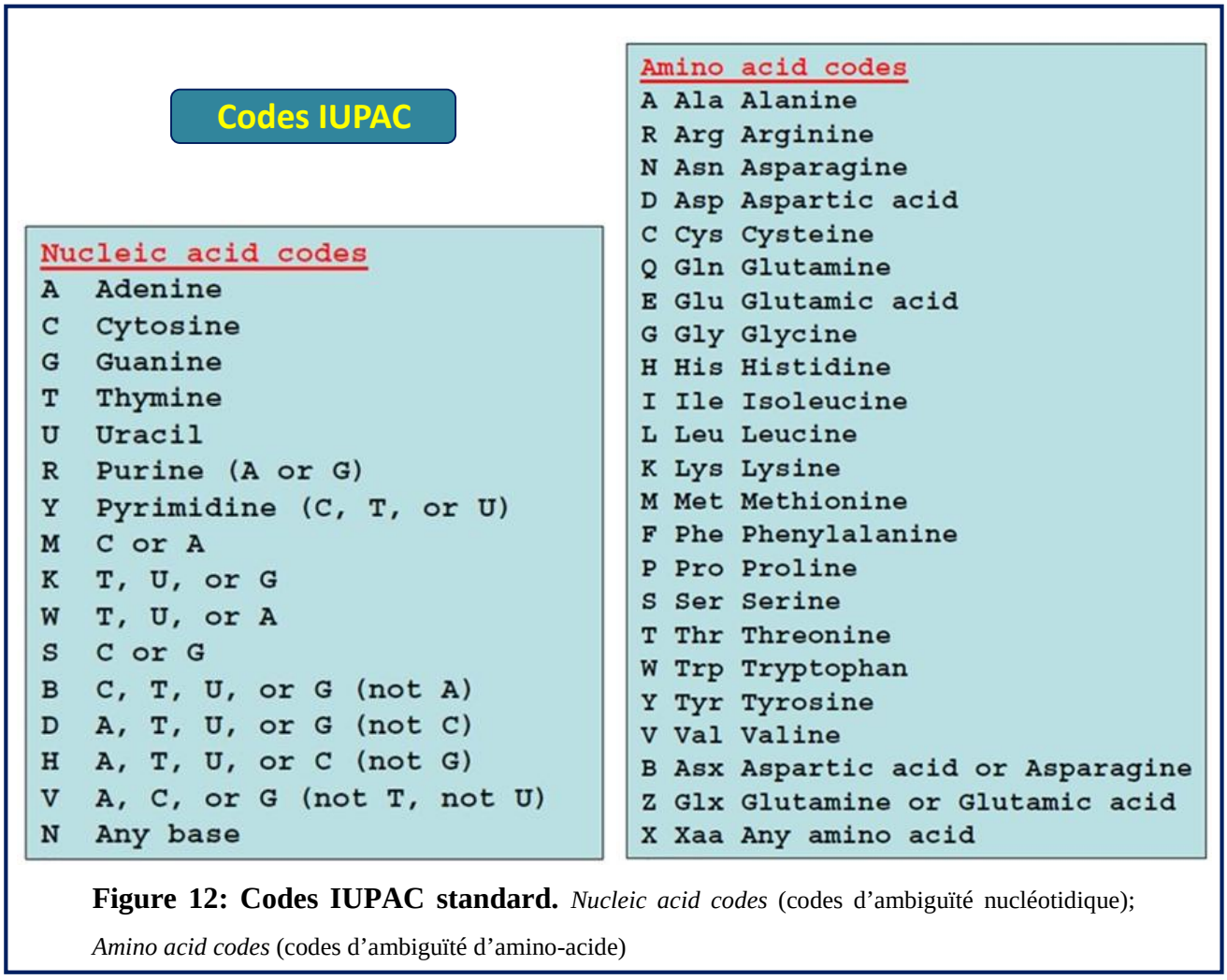


Figure 12: Codes IUPAC standard. *Nucleic acid codes* (codes d'ambiguïté nucléotidique); *Amino acid codes* (codes d'ambiguïté d'amino-acide)

III.2.1 GenBank

Du côté américain, soutenue par le NIH (*National Institute of Health*) une banque nucléique nommée GenBank a été créée en 1986. Cette base de données était distribuée par la société *IntelliGenetics* et est diffusée maintenant par le NCBI (*National Center for Biotechnology Information*).

Cette banque contient 776,291.211.106 nucléotides (**figure 14A**) dans 226,241.746 entrées (**figure 14B**) à la date de février 2021. A titre de comparaison, GenBank à ses débuts (décembre 1982) ne contenait que 680.338 nucléotides dans 606 entrées.

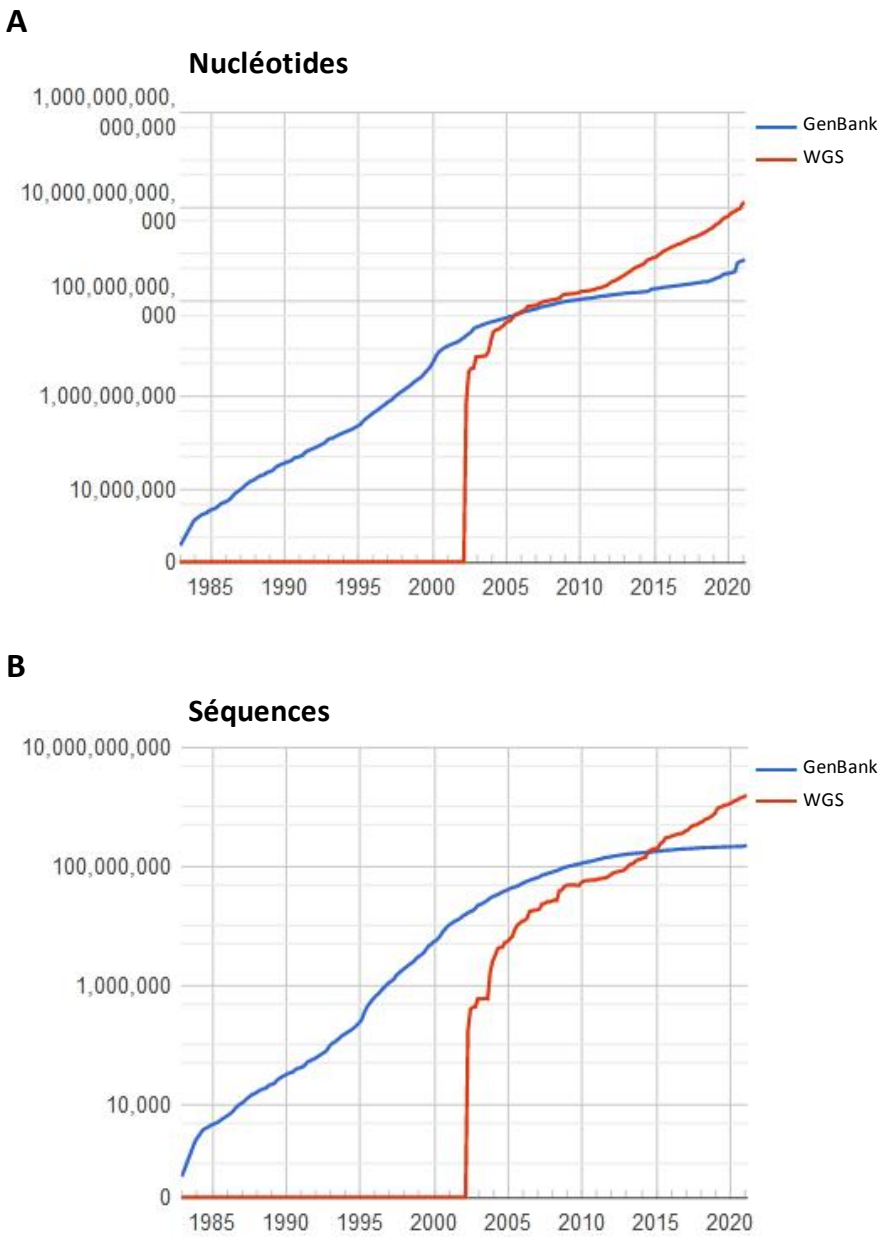
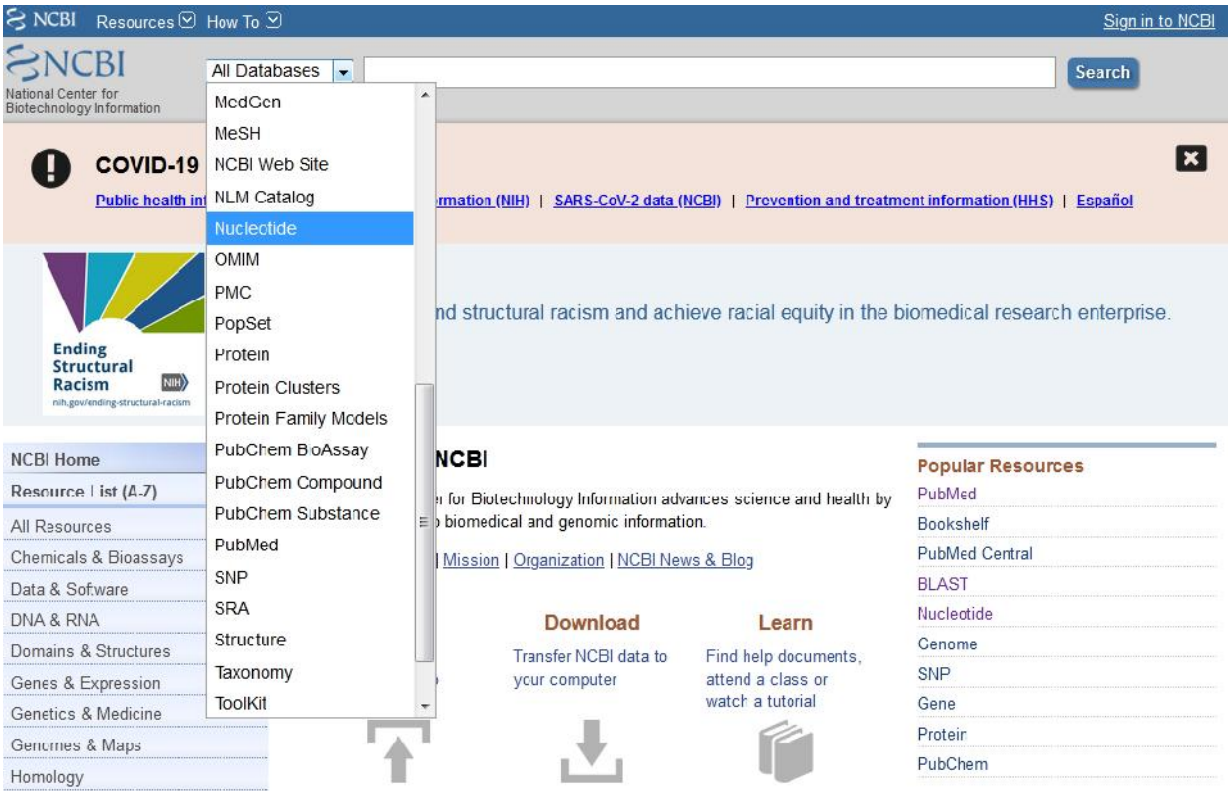
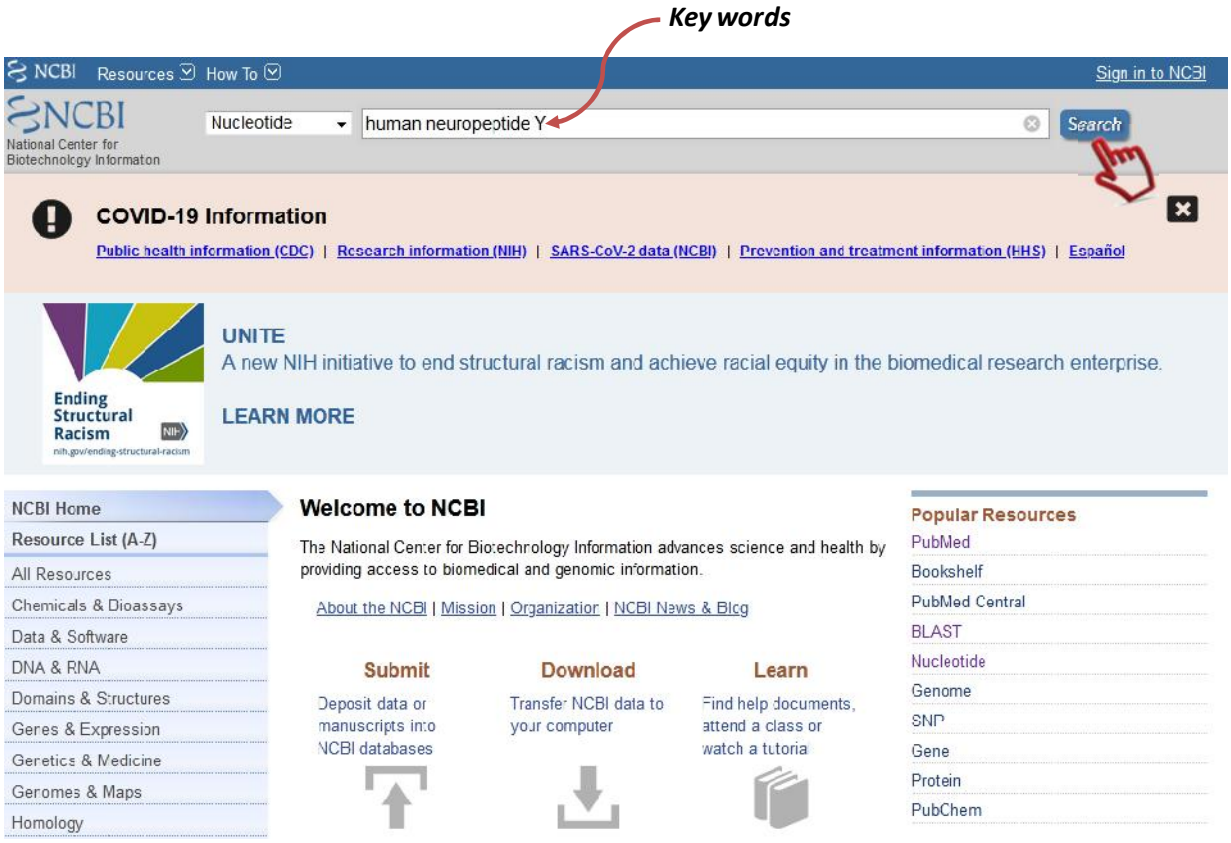


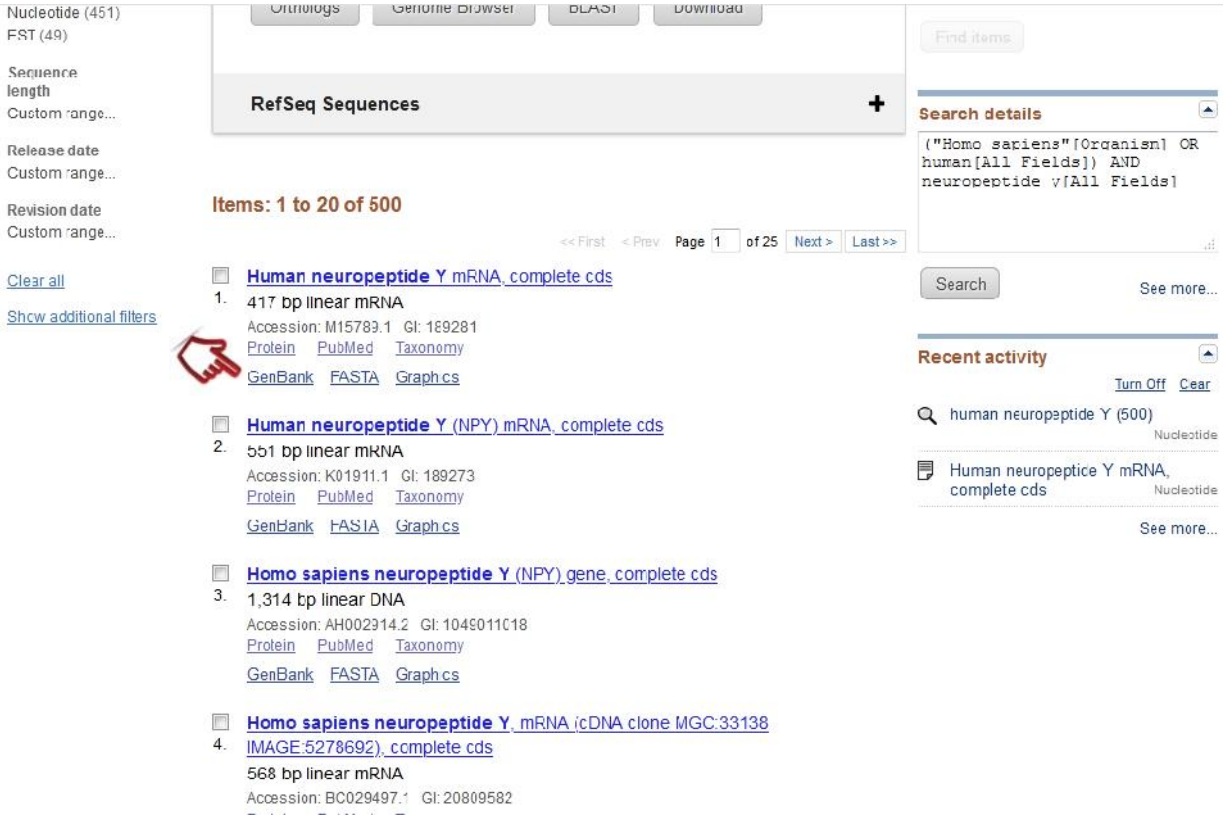
Figure 14: Statistiques d'évolution de la banque nucléiques GenBank: **A.** Evolution par nombre de nucléotides. **B.** Evolution par nombre de séquences.

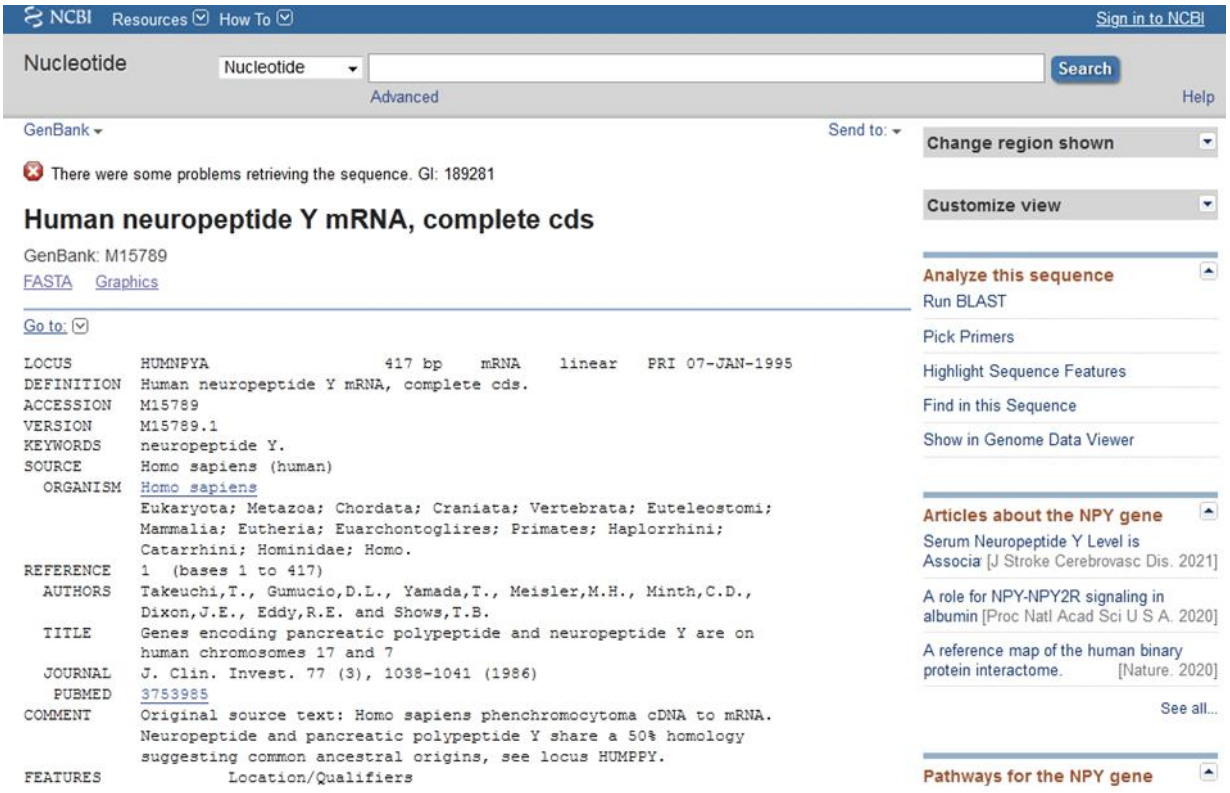
Ci-après vous trouverez les étapes de recherche d'une séquence nucléotidique dans la banque de donnée NCBI:

Allez dans: **https://www.ncbi.nlm.nih.gov**






Nucleotide (451)
EST (49)
Sequence length
Custom range...
Release date
Custom range...
Revision date
Custom range...
Clear all
Show additional filters
Orthologs
Genome Browser
BLAST
Download
RefSeq Sequences
Items: 1 to 20 of 500
<< First < Prev Page 1 of 25 Next > Last >>
1. Human neuropeptide Y mRNA, complete cds
417 bp linear mRNA
Accession: M15789.1 GI: 189281
Protein PubMed Taxonomy
GenBank FASTA Graphics
2. Human neuropeptide Y (NPY) mRNA, complete cds
551 bp linear mRNA
Accession: K01911.1 GI: 189273
Protein PubMed Taxonomy
GenBank FASTA Graphics
3. Homo sapiens neuropeptide Y (NPY) gene, complete cds
1,314 bp linear DNA
Accession: AH002914.2 GI: 1049011018
Protein PubMed Taxonomy
GenBank FASTA Graphics
4. Homo sapiens neuropeptide Y, mRNA (cDNA clone MGC:33138 IMAGE:5278692), complete cds
568 bp linear mRNA
Accession: BC029497.1 GI: 20809582
Find items
Search details
("Homo sapiens"[Organism] OR human[All Fields]) AND neuropeptide y[All Fields]
Search
See more...
Recent activity
Turn Off Clear
human neuropeptide Y (500) Nucleotide
Human neuropeptice Y mRNA, complete cds Nucleotide
See more...


NCBI Resources How To
Sign in to NCBI
Nucleotide
Nucleotide
Search
Advanced
Help
GenBank
Send to:
There were some problems retrieving the sequence. GI: 189281
Human neuropeptide Y mRNA, complete cds
GenBank: M15789
FASTA Graphics
Go to:
LOCUS HUMNPYA 417 bp mRNA linear PRI 07-JAN-1995
DEFINITION Human neuropeptide Y mRNA, complete cds.
ACCESSION M15789
VERSION M15789.1
KEYWORDS neuropeptide Y.
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 417)
AUTHORS Takeuchi,T., Gumucio,D.L., Yamada,T., Meisler,M.H., Minth,C.D., Dixon,J.E., Eddy,R.E. and Shows,T.B.
TITLE Genes encoding pancreatic polypeptide and neuropeptide Y are on human chromosomes 17 and 7
JOURNAL J. Clin. Invest. 77 (3), 1038-1041 (1986)
PUBMED 3753985
COMMENT Original source text: Homo sapiens phenchromocytoma cDNA to mRNA. Neuropeptide and pancreatic polypeptide Y share a 50% homology suggesting common ancestral origins, see locus HUMPPY.
FEATURES Location/Qualifiers
Change region shown
Customize view
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence
Show in Genome Data Viewer
Articles about the NPY gene
Serum Neuropeptide Y Level is Associa [J Stroke Cerebrovasc Dis. 2021]
A role for NPY-NPY2R signaling in albumin [Proc Natl Acad Sci U S A. 2020]
A reference map of the human binary protein interactome. [Nature. 2020]
See all...
Pathways for the NPY gene

Voici un exemple d'entrée (*Human* Neuropeptide Y) dans la banque de données GenBank:

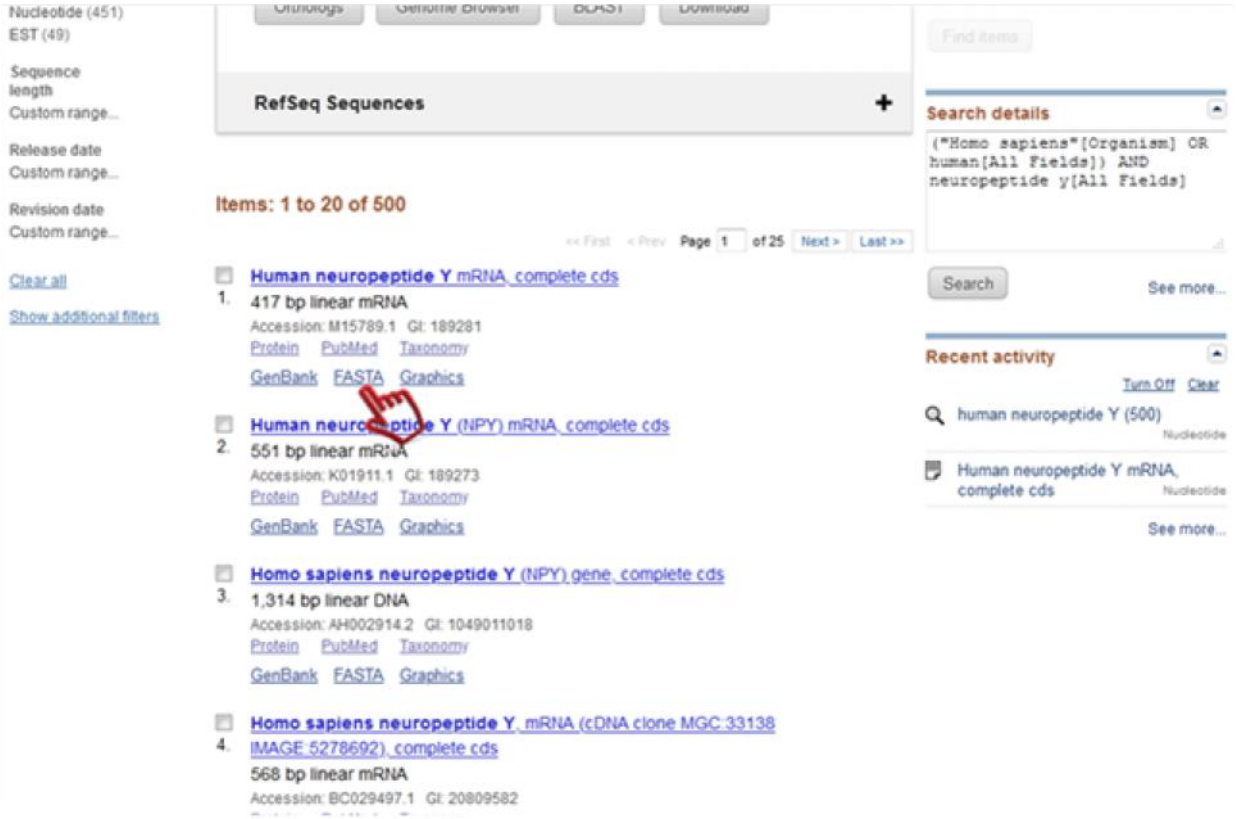
```
LOCUS       HUMNPY          551 bp    mRNA    linear   PRI 07-JAN-1995
DEFINITION  Human neuropeptide Y (NPY) mRNA, complete cds.
ACCESSION   K01911
VERSION     K01911.1
KEYWORDS    neuropeptide Y.

SOURCE      Homo sapiens (human)
  ORGANISM    Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi; Mammalia; Eutheria; Euarchontoglires;
Primates;Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE   1  (bases 1 to 551)
  AUTHORS   Minth,C.D., Bloom,S.R., Polak,J.M. and Dixon,J.E.
  TITLE   Cloning, characterization, and DNA sequence of a human cDNA
encoding neuropeptide tyrosine
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 81 (14), 4577-4581 (1984)
   PUBMED   6589611

COMMENT     Original source text: Human pheochromocytoma, cDNA to
mRNA, clone pNPY3-75. Neuropeptide Y (NPY) is one of
the most abundant peptides in the mammalian nervous
system, and its extensive distributionsuggests a
neuro-transmitter or -modulator role. NPY is also
found in some chromaffin cells of the adrenal medulla.

FEATURES             Location/Qualifiers
source          1..551
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /map="7pter-q22"
                     /tissue_type="pheochromocytoma"
gene            1..551
                     /gene="NPY"
mRNA<1..551
                     /gene="NPY"
                     /note="G00-119-456"
CDS87..380
                     /gene="NPY"
                     /codon_start=1
/product="neuropeptide Y"
                     /protein_id="AAA59944.1"
/db_xref="GDB:G00-119-456"
                     /translation="MLGNKRLGLSGLTLALSLLVCLGALAEAYPSK
                     PDNPGEDAPAEDMARYYSALRHYINLITRQRYGKRSSPETLISDLL
                     MRESTENVPRTRLEDPAMW"
sig_peptide     87..170
                     /gene="NPY"
                     /note="G00-119-456"
mat_peptide     171..278
                     /gene="NPY"
/product="neuropeptide Y"
                     /note="G00-119-456"
ORIGIN      51 bp upstream of RsaI site.
 1 accccatccgctggctctcacccctcggagacgctcgcccgacagcatagtacttgccgc
61 ccagccacgcccgcgcgccagccaccatgctaggtaacaagcgactggggctgtccggac
```

```
121 tgaccctcgccctgtccctgctcgtgtgcctgggtgcgctggccgaggcgtacccctcca
181 agccggacaacccgggcgaggacgcaccagcggaggacatggccagatactactcggcgc
241 tgcgacactacatcaacctcatcaccaggcagagatatggaaaacgatccagcccagaga
301 cactgatttcagacctcttgatgagagaaagcacagaaaatgttcccagaactcggcttg
361 aagaccctgcaatgtggtgatgggaaatgagacttgctctctggccttttcctattttca
421 gcccatatttcatcgtgtaaaacgagaatccacccatcctaccaatgcatgcagccactg
481 tgctgaattctgcaatgttttcctttgtcatcattgtatatatgtgtgtttaaataaagt
541 atcatgcatt c
//
```

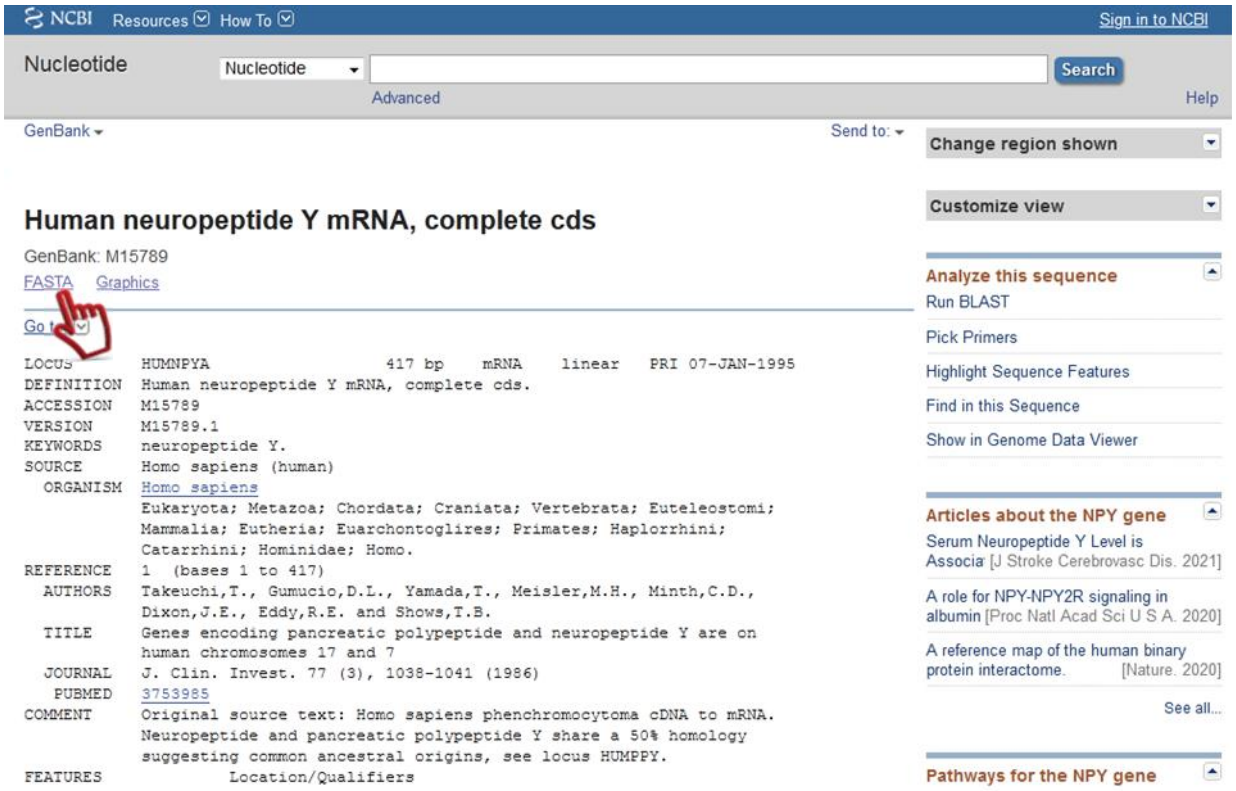
Le texte sous-ligné est une information par un lien (URL) vers un serveur Web. Les informations et leurs formats sont très similaires à celles de la banque EMBL, à ceci près que les étiquettes ne sont pas des abréviations mais le nom complet, directement explicite. Rappelons que depuis 1987 pour les "*Features*", un système de conventions communes a été adopté par les trois banques généralistes nucléiques : "The DDBJ/EMBL/GenBank *Feature Table Definition*".

Il est également possible de retrouver la séquence uniquement sous le format simplifié FASTA (*Fast Alignement*). Ce dernier est commun à toutes les bases de données nucléotidiques. Il permet ainsi de manipuler facilement des séquences dans les bases de données, à l'aide d'un format universel, compatibles avec les traitements de texte, sous forme de fichier texte dépourvu de chiffre, par simple copie/coller.

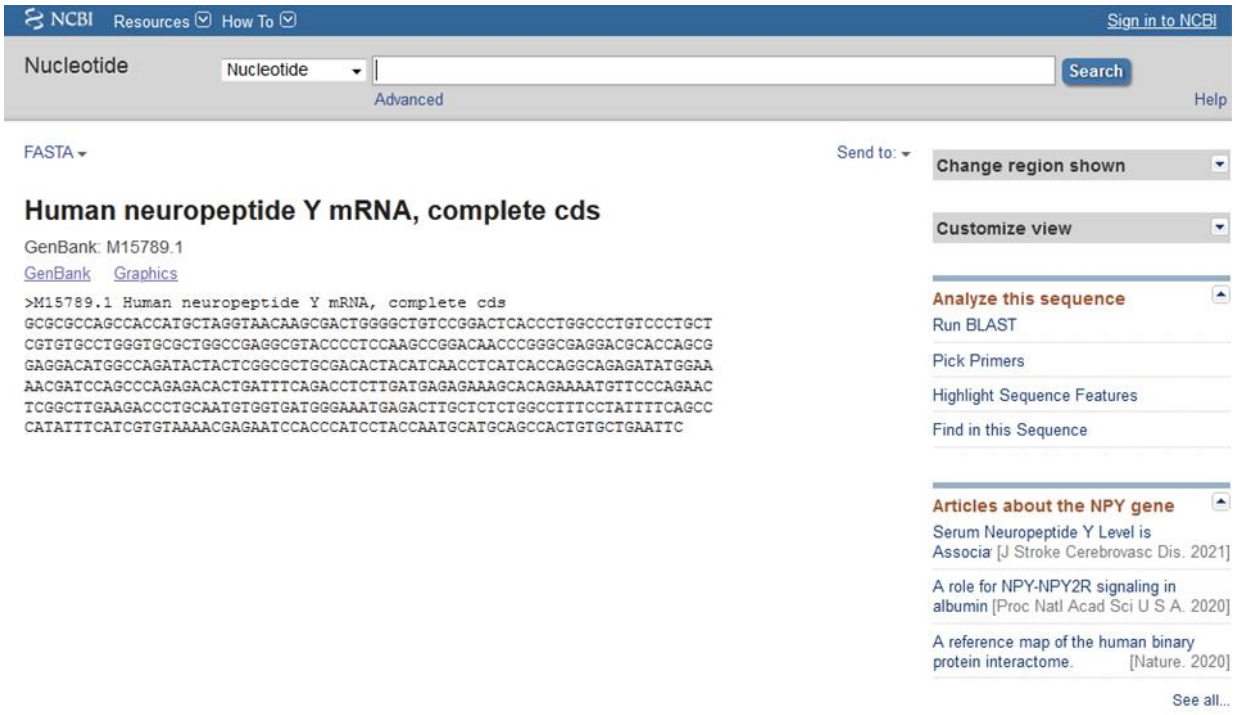
Voici à titre d'exemple la procédure pour accéder à la forme FASTA d'une séquence dans la banque GenBank. Deux (02) possibilités:



Ou


NCBI Resources How To
Sign in to NCBI
Nucleotide
Nucleotide
Search
Advanced
Help
GenBank
Send to:
Change region shown
Customize view
Human neuropeptide Y mRNA, complete cds
GenBank: M15789
FASTA Graphics
LOCUS HUMNPYA 417 bp mRNA linear PRI 07-JAN-1995
DEFINITION Human neuropeptide Y mRNA, complete cds.
ACCESSION M15789
VERSION M15789.1
KEYWORDS neuropeptide Y.
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 417)
AUTHORS Takeuchi,T., Gumucio,D.L., Yamada,T., Meisler,M.H., Minth,C.D.,
Dixon,J.E., Eddy,R.E. and Shows,T.B.
TITLE Genes encoding pancreatic polypeptide and neuropeptide Y are on
human chromosomes 17 and 7
JOURNAL J. Clin. Invest. 77 (3), 1038-1041 (1986)
PUBMED 3753985
COMMENT Original source text: Homo sapiens phenchromocytoma cDNA to mRNA.
Neuropeptide and pancreatic polypeptide Y share a 50% homology
suggesting common ancestral origins, see locus HUMPPY.
FEATURES Location/Qualifiers
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence
Show in Genome Data Viewer
Articles about the NPY gene
Serum Neuropeptide Y Level is Associa [J Stroke Cerebrovasc Dis. 2021]
A role for NPY-NPY2R signaling in albumin [Proc Natl Acad Sci U S A. 2020]
A reference map of the human binary protein interactome. [Nature. 2020]
See all...
Pathways for the NPY gene

Résultat


NCBI Resources How To
Sign in to NCBI
Nucleotide
Nucleotide
Search
Advanced
Help
FASTA
Send to:
Change region shown
Customize view
Human neuropeptide Y mRNA, complete cds
GenBank: M15789.1
GenBank Graphics
>M15789.1 Human neuropeptide Y mRNA, complete cds
GCGCGCCAGCCACCATGCTAGGTAACAAGCGACTGGGGCTGTCCGGACTCACCCTGGCCCTGTCCCTGCT
CGTGTGCCTGGGTGCGCTGGCCGAGGCGTACCCCTCCAAGCCGGACAACCCGGGCGAGGACGCACCAGCG
GAGGACATGGCCAGATACTACTCGGCGCTGCGACACTACATCAACCTCATCACCAGGCAGAGATATGGAA
AACGATCCAGCCCAGAGACACTGATTTCAGACCTCTTGATGAGAGAAAGCACAGAAAATGTTCCCAGAAC
TCGGCTTGAAGACCCTGCAATGTGGTGATGGGAAATGAGACTTGCTCTCTGGCCTTTCCTATTTTCAGCC
CATATTTCATCGTGTAAAACGAGAATCCACCCATCCTACCAATGCATGCAGCCACTGTGCTGAATTC
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence
Articles about the NPY gene
Serum Neuropeptide Y Level is Associa [J Stroke Cerebrovasc Dis. 2021]
A role for NPY-NPY2R signaling in albumin [Proc Natl Acad Sci U S A. 2020]
A reference map of the human binary protein interactome. [Nature. 2020]
See all...

III.2.2 EMBL

En Europe, financée par l'EMBO (*European Moleculary Biology Organisation*), une équipe s'est constituée pour développer une banque de séquences nucléiques (EMBL *data library*) et en assurer la diffusion. Cette équipe travaille au sein du Laboratoire Européen de Biologie Moléculaire qui est longtemps resté à Heidelberg et qui se trouve actuellement près de Cambridge au sein de l'EBI (*European Bioinformatics Institute*).

Cette banque contient 10,265.200.000 nucléotides dans 2,808.100 à la date de mai 2021 (**figure 13**).

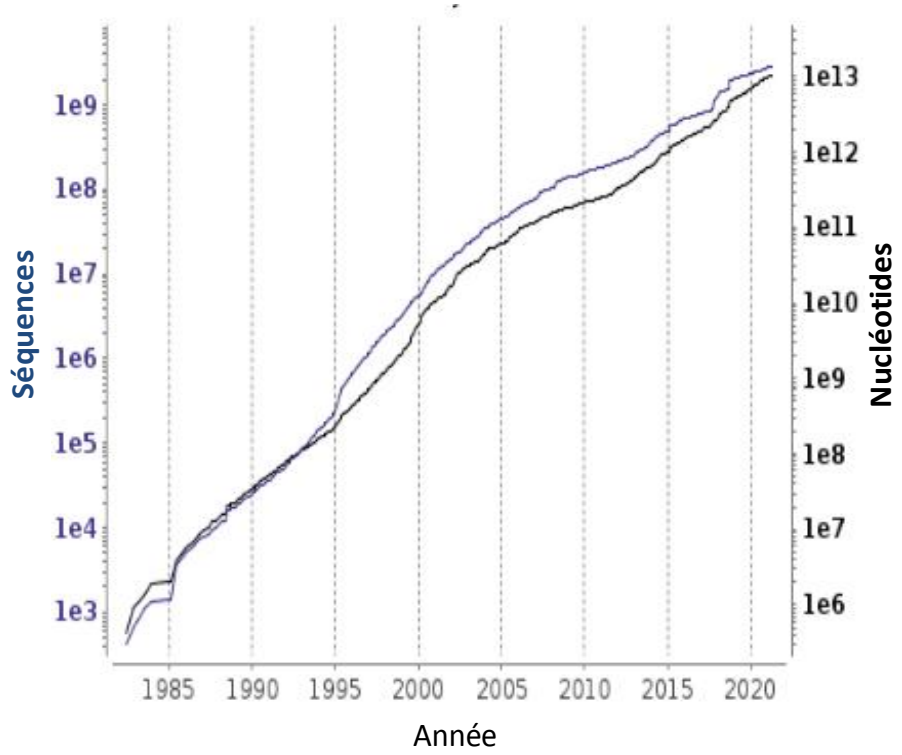


Figure 13: Statistiques d'évolution de la banque de données EMBL depuis sa création

Voici un exemple d'entrée dans EMBL (*Human* Neuropeptide Y):

```
ID   K01911; SV 1; linear; mRNA; STD; HUM; 551 BP.
XX
AC   K01911;
XX
DT   13-JUN-1985 (Rel. 06, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 5)
XX
DE   Human neuropeptide Y (NPY) mRNA, completecds.
XX
KW   neuropeptide Y.
```

```
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
OC   Euteleostomi; Mammalia;Eutheria; Euarchontoglires; Primates;
OC   Haplorrhini; Catarrhini; Hominidae;Homo.
XX
RN   [1]
RP   1-551
RX   DOI; 10.1073/pnas.81.14.4577.
RX   PUBMED; 6589611.
RA   Minth C.D., Bloom S.R., Polak J.M., Dixon J.E.;
RT   "Cloning, characterization, and DNA sequence of a human cDNA
RT   encoding neuropeptide tyrosine";
RL   Proc. Natl. Acad. Sci. U.S.A. 81(14):4577-4581(1984).
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..551
FT                   /organism="Homo sapiens"
FT                   /map="7pter-q22"
FT                   /mol_type="mRNA"
FT                   /tissue_type="pheochromocytoma"
FT                   /db_xref="taxon:9606"
FT   mRNA            <1..551
FT                   /gene="NPY"
FT                   /note="G00-119-456"
FT   sig_peptide     87..170
FT                   /gene="NPY"
FT                   /note="G00-119-456"
FT   CDS             87..380
FT                   /codon_start=1
FT                   /gene="NPY"
FT                   /product="neuropeptide Y"
FT                   /db_xref="GOA:P01303"
FT                   /db_xref="H-InvDB:HIT000191373.15"
FT                   /db_xref="HGNC:HGNC:7955"
FT                   /db_xref="InterPro:IPR001955"
FT                   /db_xref="InterPro:IPR020392"
FT                   /db_xref="PDB:1QFA"
FT                   /db_xref="PDB:1RON"
FT                   /db_xref="UniProtKB/Swiss-Prot:P01303"
FT                   /protein_id="AAA59944.1"
FT       /translation="MLGNKRLGLSGLTLALSLLVCLGALAEAYPSKPDNPGEDAPAEDM
FT       ARYYSALRHYINLITRQRYGKRSSPETLISDLLMRESTENVPRTRLEDPAMW"
FT   mat_peptide     171..278
FT                   /gene="NPY"
FT                   /product="neuropeptide Y"
FT                   /note="G00-119-456"
XX
SQ   Sequence 551 BP; 131 A; 171 C; 129 G; 120 T; 0 other;
accccatccgctggctctcacccctcggagacgctcgcccgacagcatagtacttgccgc 60
ccagccacgcccgcgcgccagccaccatgctaggtaacaagcgactggggctgtccggac 120
tgaccctcgccctgtccctgctcgtgtgcctgggtgcgctggccgaggcgtacccctcca 180
agccggacaacccgggcgaggacgcaccagcggaggacatggccagatactactcggcgc 240
tgcgacactacatcaacctcatcaccaggcagagatatggaaaacgatccagcccagaga 300
```

```
cactgatttcagacctcttgatgagagaaagcacagaaaatgttcccagaactcggcttg 360
aagaccctgcaatgtggtgatgggaaatgagacttgctctctggccttttcctattttca 420
gcccatatttcatcgtgtaaaacgagaatccacccatcctaccaatgcatgcagccactg 480
tgctgaattctgcaatgttttcctttgtcatcattgtatatatgtgtgtttaaataaagt 540
atcatgcattc 551
//
```

Chaque entrée de la base EMBL est composée de lignes ou champs qui commencent par une étiquette, code à deux (02) caractères indiquant le type d'information contenue dans la ligne et la fin de l'entrée est indiquée par **//**.

Ces étiquettes sont divisées en cinq parties:

1) General Information

- Etiquette **ID**: identificateur de l'entrée contenant la séquence. Cette ligne a la structure suivante: nom de l'entrée classe de la donnée ; molécule (ADN, ARN, ARNm, XXX si l'entrée n'a pas été annotée) ; division ; longueur de la séquence en paire de bases (pb).
- Etiquette **XX**: C'est une ligne vide qui sert à limiter les différents champs de l'entrée et à clarifier sa lecture
- Etiquette **AC**: numéro d'accession de l'entrée
- Etiquette **SV**: version de la séquence
- Etiquette **DT**: donne la date d'incorporation dans la base (1ère ligne) et la date de la dernière mise à jour de l'entrée (2ème ligne)

2) Description

- Etiquette **DE**: informations descriptives sur la séquence
- Etiquette **KW**: mot(s)-clé(s)
- Etiquette **OS**: organisme
- Etiquette **OC**: ordre dans la classification

3) References

- Etiquette **RN**: numéro de la référence (peut être utilisé dans les *features*)
- Etiquette **RC**: commentaires sur la référence
- Etiquette **RP**: région de la séquence
- Etiquette **RX**: lien (URL) vers des bases bibliographiques accessibles par le réseau (par exemple Medline, PubMed)
- Etiquette **RA**: auteurs de la publication
- Etiquette **RT**: titre de la publication

- Etiquette **RL** : référence : journal, volume, pages, année (peut aussi porter la mention: unpublished)

4) Additional Information:

Features (facultatif): La ligne est composée de l'étiquette **FT**, suivie d'un mot-clé (**Key**), lui-même suivi par le champ (**Location/Qualifiers**) qui est un couple de mot-clé/valeur, le mot-clé étant soit de type "location" soit de type "qualifier". Depuis 1987 un système de conventions communes a été adopté par les trois banques généralistes nucléiques : "The DDBJ/EMBL/GenBank *Feature Table Definition*".Lorsque le mot-clé (*key*) est absent, cela signifie que la ligne "FT *Location/Qualifiers*" est la suite de la précédente.

Un "qualifier" de type *db_xref* est un lien (URL) vers une banque ou base de données, c'est une référence croisée.

5) Sequence: Une seule étiquette: **SQ** qui contient le nombre de paires de base et la répartition entre les différents nucléotides. La séquence est suivie par l'étiquette // qui indique la fin de l'entrée.

III.2.3 DDBJ :

Créée en 1986 et diffusée par le NIG (*National Institute of Genetics*, Japon), elle contient 13,231.329.651.111 nucléotides dans 2,695.960.482 entrées à la date du Décembre 2020 (**figure 15**).

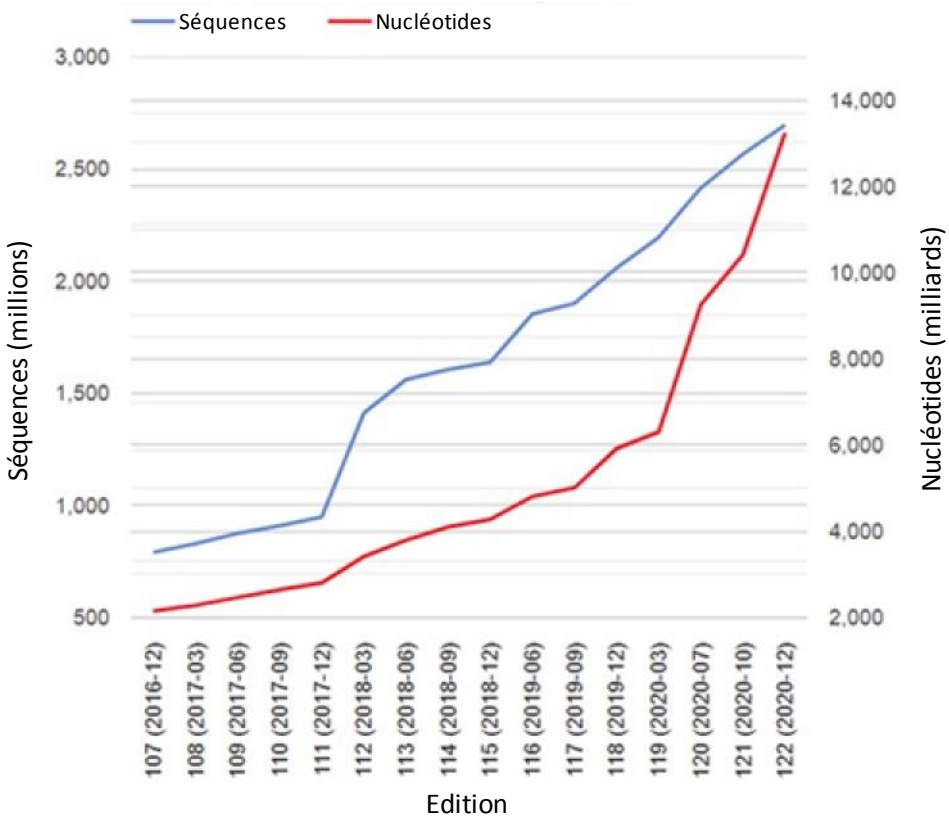


Figure 15: Statistiques d'évolution de la banque de données DDBJ

Voici un exemple d'entrée dans la banque de données DDBJ (*même gène que pour* GenBank et EMBL):

```
LOCUS       HUMNPY                551 bp    mRNA    linear   HUM 07-
JAN-1995
DEFINITION  Human neuropeptide Y (NPY) mRNA, complete cds.
ACCESSION   K01911
VERSION     K01911.1
KEYWORDS    neuropeptide Y.
SOURCE      Homo sapiens (human)
ORGANISM  Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi; Mammalia; Eutheria; Euarchontoglires;
Primates;Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 551)
  AUTHORS   Minth,C.D., Bloom,S.R., Polak,J.M. and Dixon,J.E.
  TITLE     Cloning, characterization, and DNA sequence of a human
cDNA encoding neuropeptide tyrosine
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 81 (14), 4577-4581 (1984)
   PUBMED   6589611
COMMENT     Original source text: Human pheochromocytoma, cDNA to
mRNA, clone
pNPY3-75.
            Neuropeptide Y (NPY) is one of the most abundant peptides
in themammalian nervous system, and its extensive
distribution suggests aneuro-transmitter or –modulator
role.NPY is alsofound in some chromaffin cells of the
```

```
adrenalmedulla.
FEATURES             Location/Qualifiers
source          1..551
/db_xref="H-InvDB:HIT000191373"
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /map="7pter-q22"
                     /tissue_type="pheochromocytoma"
gene            1..551
                     /gene="NPY"
mRNA<1..551
                     /gene="NPY"
                     /note="G00-119-456"
CDS             87..380
                     /gene="NPY"
                     /codon_start=1
/product="neuropeptide Y"
                     /protein_id="AAA59944.1"
                     /db_xref="GDB:G00-119-456"
/translation="MLGNKRLGLSGLTLALSLLVCLGALAEAYPSKPD
NPGEDAPAEDMARYYSALRHYINLITRQRYGKRSSPETLISDLLMRES
TENVPRTRLEDPAMW"
sig_peptide     87..170
                     /gene="NPY"
/note="G00-119-456"
mat_peptide     171..278
                     /gene="NPY"
/product="neuropeptide Y"
                     /note="G00-119-456"
BASE COUNT         131 a          171 c          129 g          120 t
ORIGIN      51 bp upstream of RsaI site.
  1 accccatccgctggctctcacccctcggagacgctcgcccgacagcatagtacttgccgc
 61 ccagccacgcccgcgcgccagccaccatgctaggtaacaagcgactggggctgtccggac
121 tgaccctcgccctgtccctgctcgtgtgcctgggtgcgctggccgaggcgtacccctcca
181 agccggacaacccgggcgaggacgcaccagcggaggacatggccagatactactcggcgc
241 tgcgacactacatcaacctcatcaccaggcagagatatggaaaacgatccagcccagaga
301 cactgatttcagacctcttgatgagagaaagcacagaaaatgttcccagaactcggcttg
361 aagaccctgcaatgtggtgatgggaaatgagacttgctctctggccttttcctattttca
421 gcccatatttcatcgtgtaaaacgagaatccacccatcctaccaatgcatgcagccactg
481 tgctgaattctgcaatgttttcctttgtcatcattgtatatatgtgtgtttaaataaagt
541 atcatgcatt c
//
```

Le texte sous-ligné est une information par un lien (URL) vers un serveur Web. Les informations et leurs formats sont très similaires à celles de la banque GenBank.

« ***Défauts » des banques nucléiques*:**

- **Aucun contrôle des banques:**
  - les auteurs sont responsables de la qualité des séquences soumises.

- **Hétérogénéité:**
  - ADN nucléaire, mitochondrial, chloroplastique, ARNm, ARNt, ARNs, ARNr, chromosomes entiers ...
  - gènes, fragments … (10 bp à 350000 bp)
- **Variabilité de l'état des connaissances sur les séquences:**
  - Annotation effectuée ou non
  - Annotation hétérogènes: automatique ou expérimentale
- **Erreurs dans les séquences (qualité inégale):**
  - origine du fragment
  - cultures infectés
  - présence de séquences de vecteurs de clonage
  - erreurs de saisie
- **Redondance des données: plusieurs entrées pour une même séquence:**
  - Certains gènes sont séquencés à la fois sous forme d'ARNm et de fragments génomiques.
  - Certaines séquences ont été saisies plusieurs fois dans la banque.

IV. La comparaison de séquences

La recherche de similitude entre séquences constitue souvent la première étape des analyses de séquences. La comparaison de séquences biologiques, ainsi que leur alignement, nécessite la mise en œuvre de procédures de calcul et de modèles biologiques permettant de quantifier la notion de ressemblance ou similitude entre ces séquences (% match, score, E-value…).

Une ressemblance entre séquences peut indiquer par exemple:

- une fonction biologique proche
- une structure tridimensionnelle semblable
- une origine et/ou une histoire évolutive commune

Une similitude entre séquences est souvent un argument en faveur d'une homologie : deux séquences sont homologues si elles ont un ancêtre commun. Remarquons quand-même qu'il n'y a pas de d'équivalence entre similitude et homologie : deux séquences peuvent avoir un degré de similitude conséquent sans être homologues et deux séquences peuvent être homologues avec un degré de similitude faible.

Cette notion d'homologie reflète le dogme fondamental de l'évolution biologique :

- les régions fonctionnelles des gènes ou de leurs produits (sites catalytique, de fixation, etc.) sont soumises à la sélection : elles sont relativement préservées par l'évolution cardes mutations trop radicales sont désavantageuses et pourraient leur feraient perdre leurs fonctions.
- les régions non fonctionnelles, qui ne subissent aucune sélection, et donc divergent rapidement à mesure que s'accumulent les mutations.
- les nouveaux gènes apparaissent surtout par remaniement de gènes ancestraux : on peut souvent déduire la fonction de la plupart des gènes par comparaison avec les gènes «homologues » d'autres espèces.

Lors des comparaisons de séquences, on retrouve fréquemment la terminologie suivante:

- **Identité:** Représente la proportion de paires de résidus identiques entre deux séquences alignées (exprimée généralement en **%**).
- **Similitude:** C'est la mesure de la ressemblance entre deux séquences. Le degré de similitude est quantifié par une score basé sur le % de similarité (% identité + % substitutions conservatives) des séquences.
- **Gaps/Indels:** Proportion d'Indels entre deux séquences alignées (exprimés en **%**).
- **Homologie:** Deux séquences sont homologues si elles ont un ancêtre commun.

*Attention : Il n'y a pas de degré d'homologie, c'est soit **oui** ou **non** !!*

On ne dit pas très homologue, faible homologie…

IV.1 Les alignements

Il y a plusieurs objectifs possibles dans une opération d'alignement, en voici quelques uns:

• **Comparaison**: Comparer deux (02) séquences (deux (02) gènes proches chez une même espèce, deux(02) mêmes gènes chez deux (02) espèces différentes)

• **Identification**: Identifier une séquence en cherchant les séquences les plus proches dans des bases de données de séquences connues.

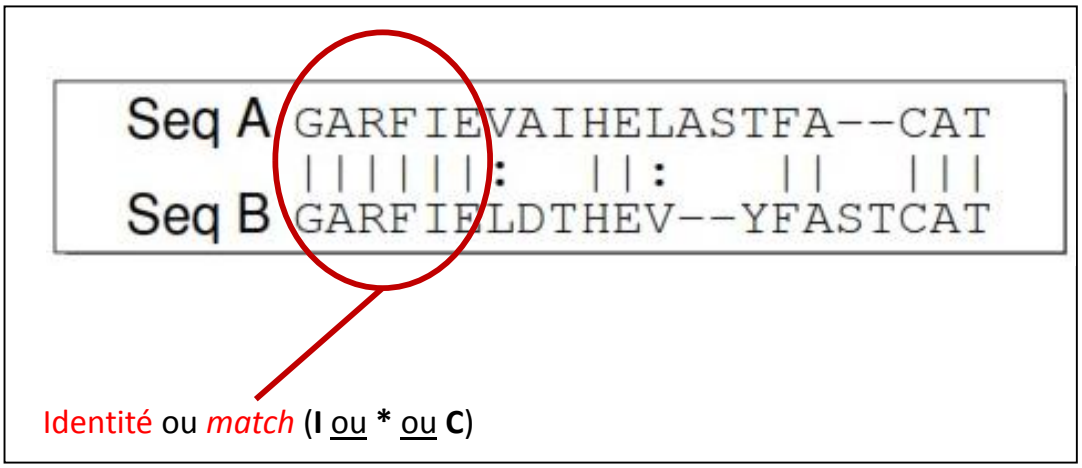
• **Prolongation**: trouver la séquence complète d'un gène dans une base de données à partir d'un simple fragment (cela peut économiser du séquençage).

• **Localisation génomique**: localiser un gène sur un génome (alignement ARN/ADN).

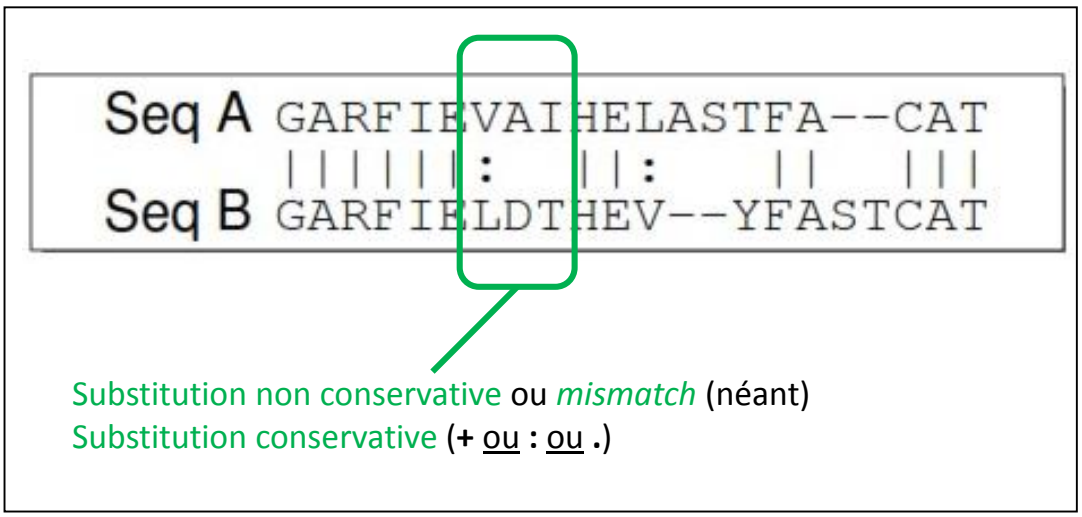
Un programme d'alignement est un programme qui permet de comparer deux (02) séquences (ou plus) en minimisant les éditions (insertion/délétion/substitution) permettant d'aligner au mieux tout ou partie des symboles composants les séquences:

Lors d'un alignement trois (03) situations sont possibles pour une position donnée de l'alignement:

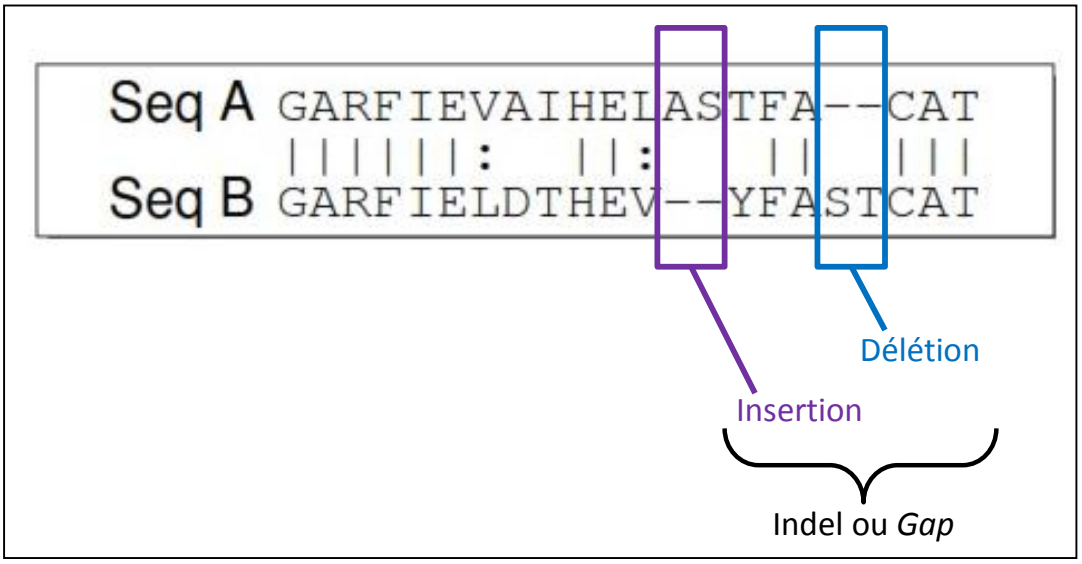
- Les caractères sont les mêmes: **Identité** ou **match**



- les caractères ne sont pas les mêmes: **Substitution**



- l'une des positions est un espace: **Insertion/Délétion**



Il existe trois types d'alignement:

- **Alignement global de séquences:** *Un alignement global de séquences* fait correspondre chacun des nucléotides d'une séquence avec un nucléotide de l'autre séquence en tentant de maximiser le nombre de matchs. Il est utilisé lorsque la t aille des séquences est semblable et les séquences assez similaires sur toute leur longueur. Il permet par exemple d'identifier des régions conservées.
- **Alignement global multiple de séquences:** Un *alignement global multiple* est une généralisation d'un alignement global à plus de deux séquences. Cela peut permettre de mettre en évidence des relations (une histoire évolutive) entre deux séquences qui sont assez divergentes lorsque comparées deux à deux, mais dont l'évolution peut devenir visible avec la présence de séquences intermédiaires.
- **Alignement local de séquences:** Un *alignement local* cherche à identifier les sous-séquences les plus similaires, sans tenir compte de la position. L'alignement local est utilisé lorsque la taille des séquences diffère beaucoup, pour rechercher des motifs partagés, ou pour faire des recherches dans les bases de données.

Dans la suite de document on décrira le mode d'utilisation d'un logiciel le plus utilisés en bioinformatique pour la recherche d'alignements locaux en l'occurrence le logiciel BLAST.

IV.2 BLAST

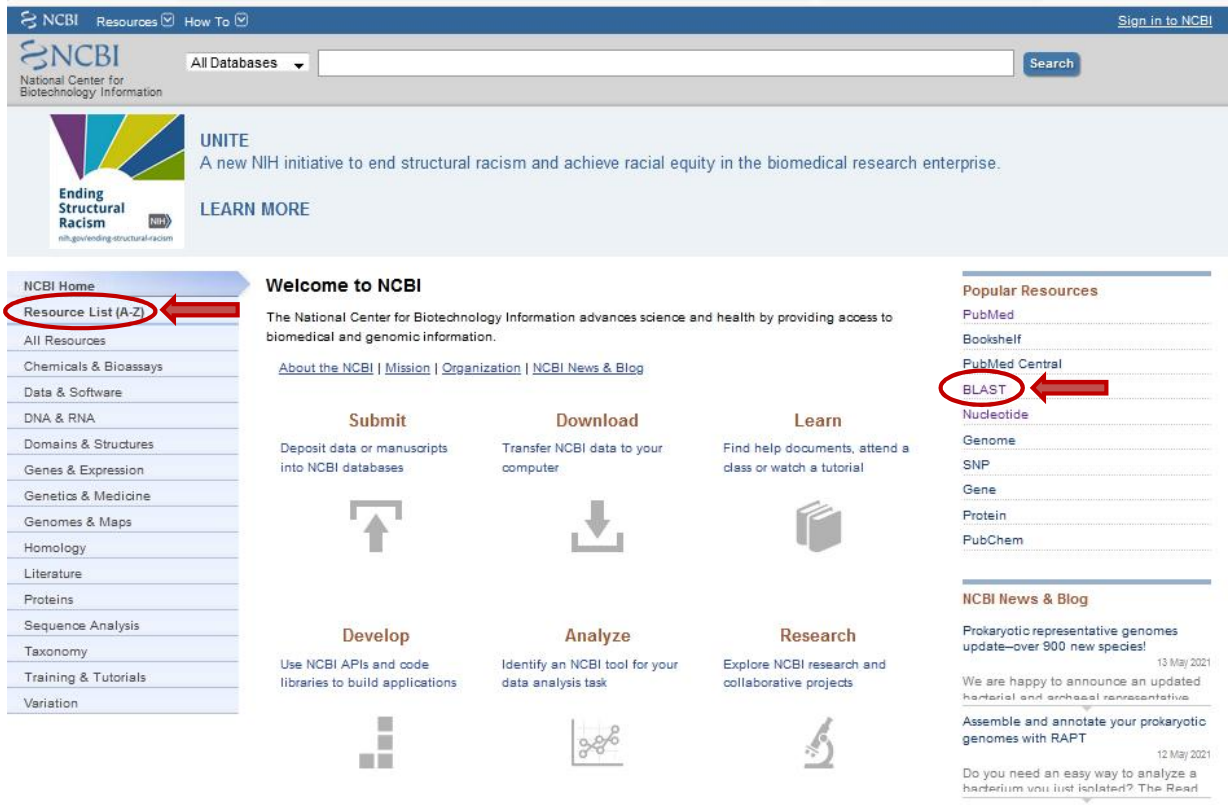
Basic Local Alignment Search Tool (BLAST) est l'un des choix les plus populaires pour la recherche et l'alignement de séquences. Il a l'avantage d'utiliser une technique de recherche rapide qui lui a permis de devenir l'outil le plus utilisé pour les recherches dans les bases de données.

Le programme BLAST est apparu en 1990, il est basé sur un algorithme heuristique et permet de trouver des séquences similaires à une séquence requête dans une banque de données. Il compare une séquence de nucléotides ou de protéines comme entrée contre une base de données de séquences de nucléotides ou des séquences de protéines, et calcule la signification statistique des résultats. BLAST peut être utilisé pour déduire les relations fonctionnelles et évolutives entre les séquences ainsi que pour aider à identifier les membres de familles de gènes.

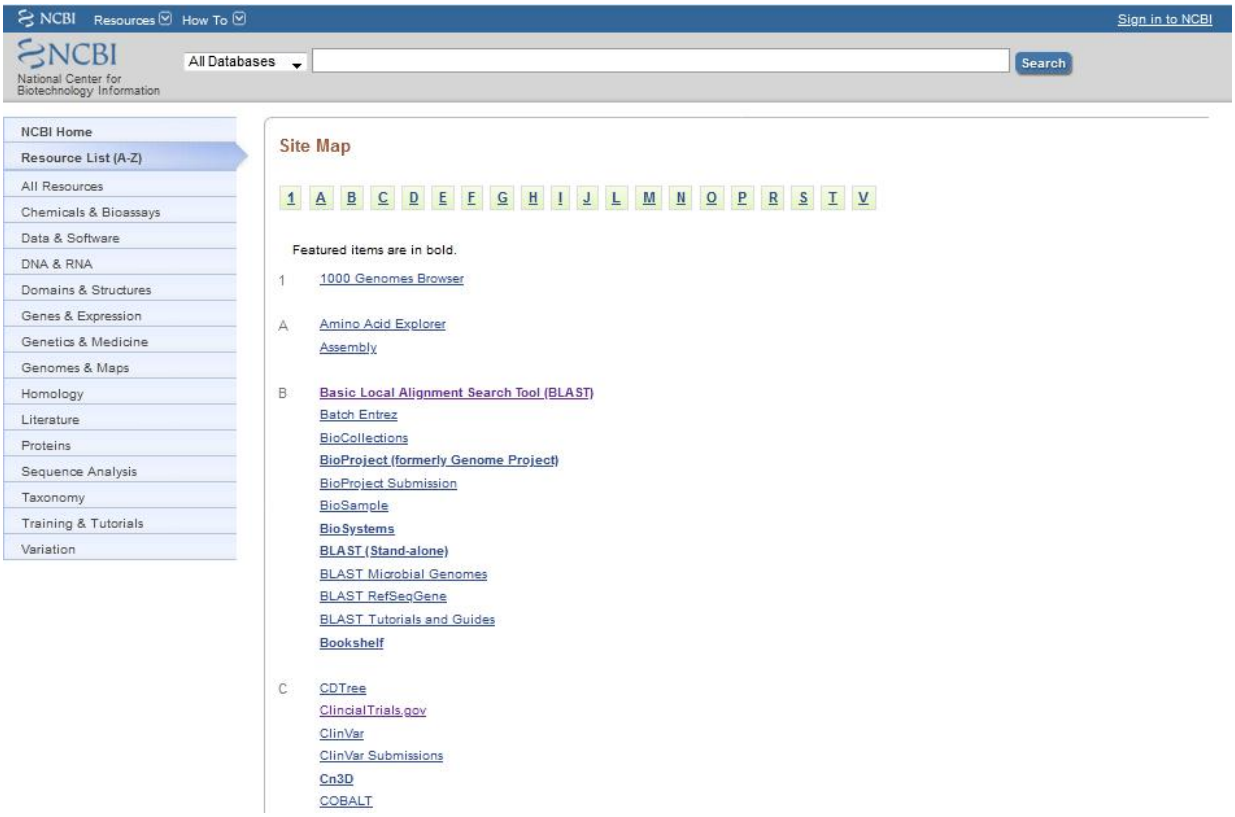
***NB*:** En informatique l'heuristique est un algorithme qui permet de rapidement dénouer des difficultés complexes d'optimisation, sans concevoir une modélisation formelle. Cela ne se résout pas forcément par une solution optimale.

Dans ce qui suit vous trouverez la procédure de recherche avec une séquence nucléotidique en utilisant l'outil BLAST de la bases de sonnées NCBI.

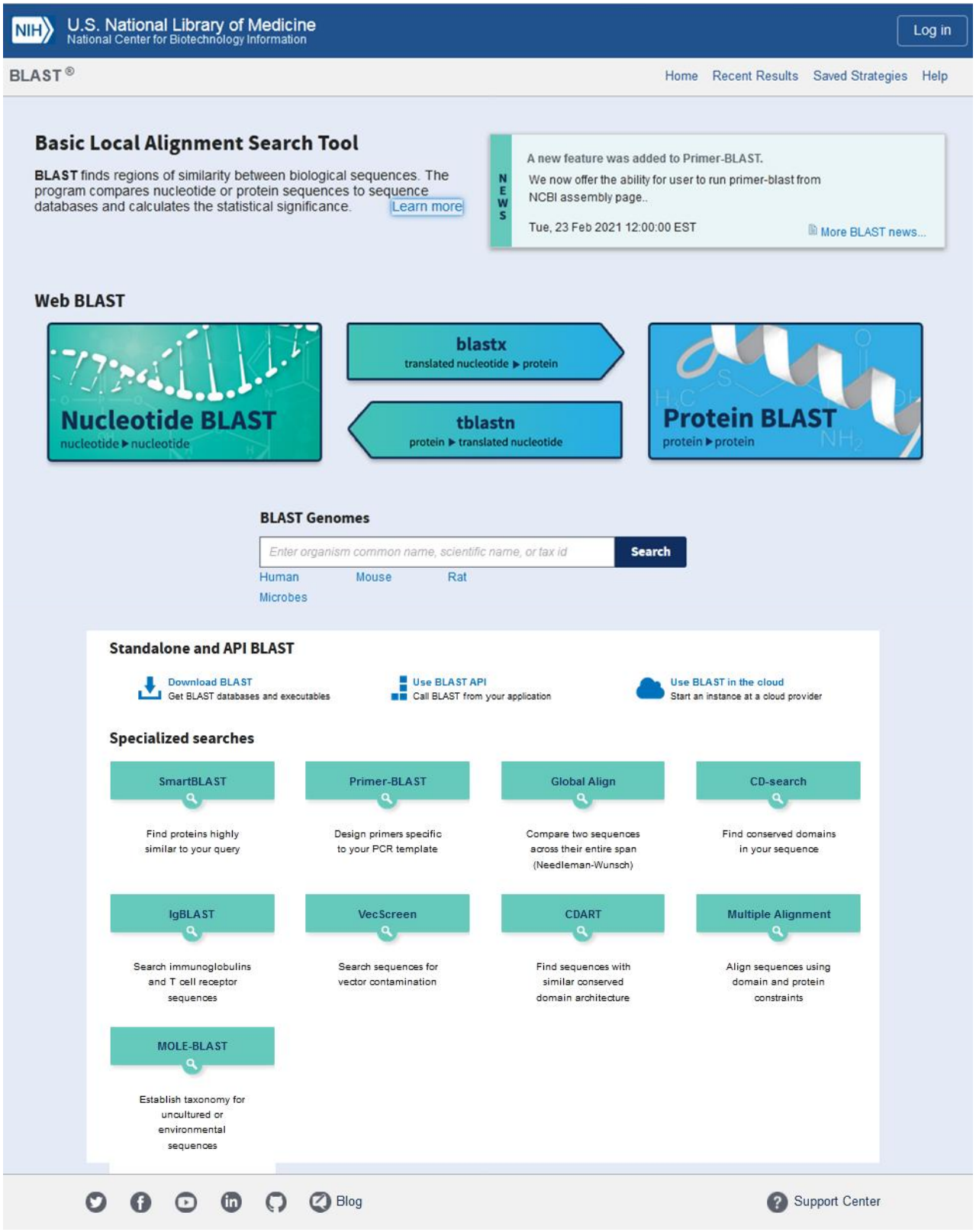
1) Vous pouvez accéder à cet engin de recherche soit *via* l'adresse: https://blast.ncbi.nlm.nih.gov/Blast.cgi soit à partir du menu des «***Popular Resources*»** ou ressources populaires sur la droite, ou par le biais du lien «***Resource list* (A-Z)**», dans le menu à gauche, dans la page d'accueil de NCBI.



2) « ***Resource List*** **(A-Z)** »: Sur cette page se retrouve la majorité des liens composant la base de données NCBI.



3) En cliquant sur le lien Blast, vous devriez obtenir la page suivante:



Comme vous le voyez ci-dessus, « **BLAST** » est une collection d'engins de recherches de similarités conçus pour examiner toutes les bases de données de séquences indépendamment qu'elle soit protéine ou ADN (**figure 16**).

- « ***Nucleotide* blast** » ou « **Blastn** » compare une séquence nucléique d'intérêt aux séquences d'une base de données d'acides nucléiques.
- « ***Protein* blast** » ou « **Blastp** » compare une séquence d'acides aminés d'intérêt aux séquences d'une base de données de protéines.
- « **Blastx** » compare une séquence nucléique d'intérêt traduite dans tous les cadres de lecture aux séquences d'une banque de données de protéines. Vous pourriez utiliser cette option pour trouver les produits de traduction d'une séquence nucléique inconnue.
- « **Tblastn** » compare une séquence d'acides aminés d'intérêt aux séquences d'une base de données d'acides nucléiques dont la traduction a été faite dans tous les cadres de lecture.
- « **Tblastx** » compare les traductions dans les six cadres de lectures d'une séquence nucléique d'intérêt aux séquences d'une base de données d'acides nucléiques traduites.

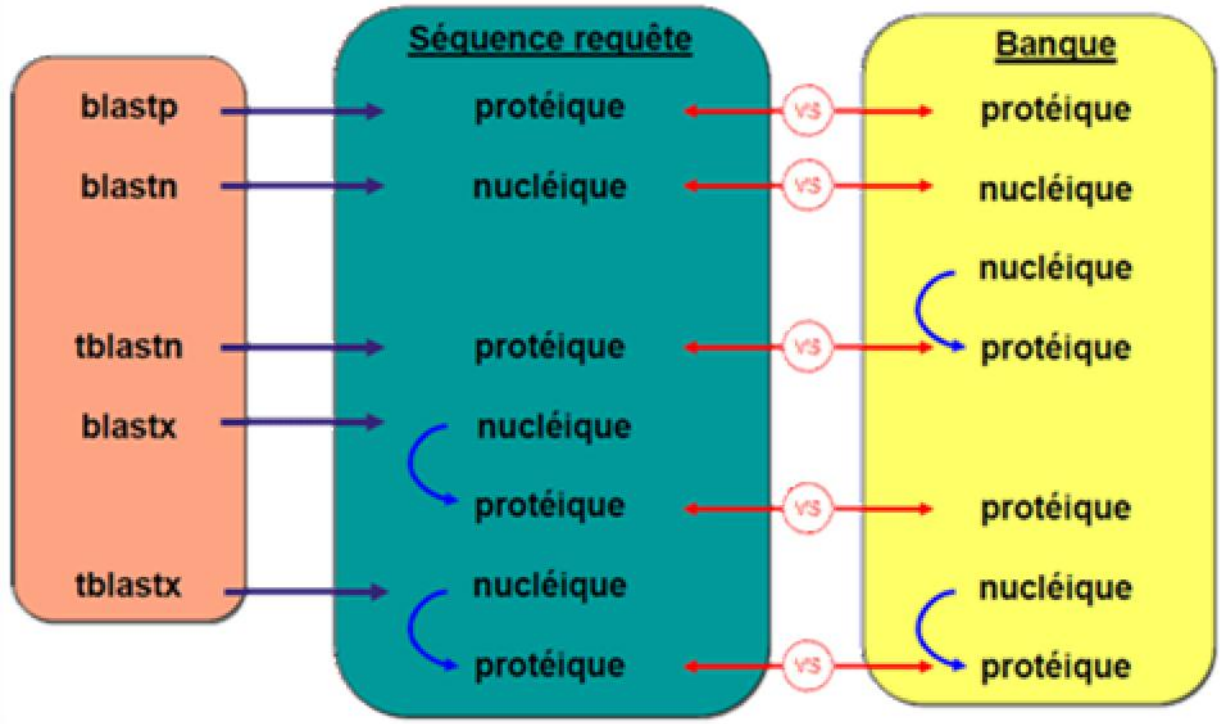
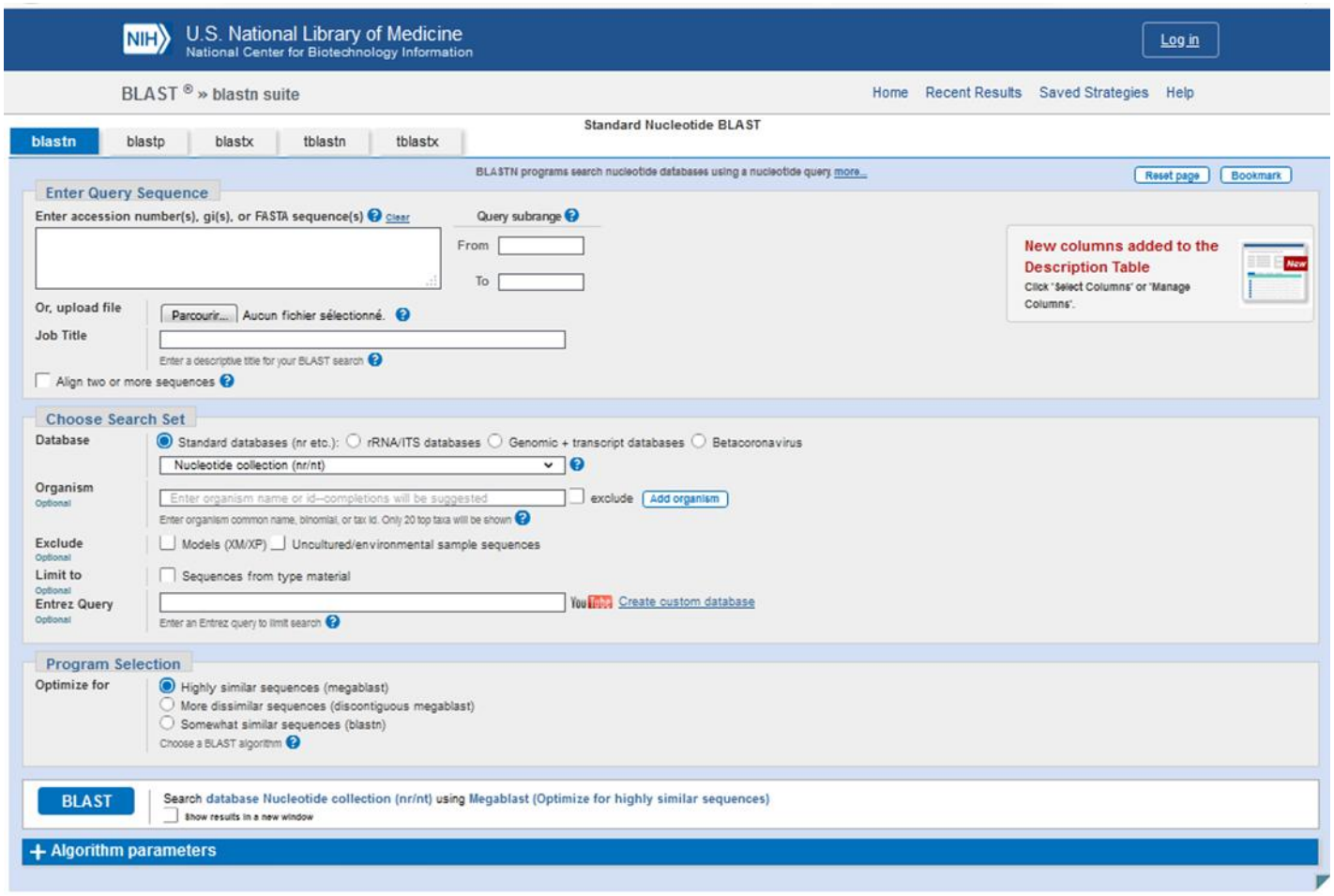


Figure 16: Différents modes de recherches de similarité du logiciel Blast

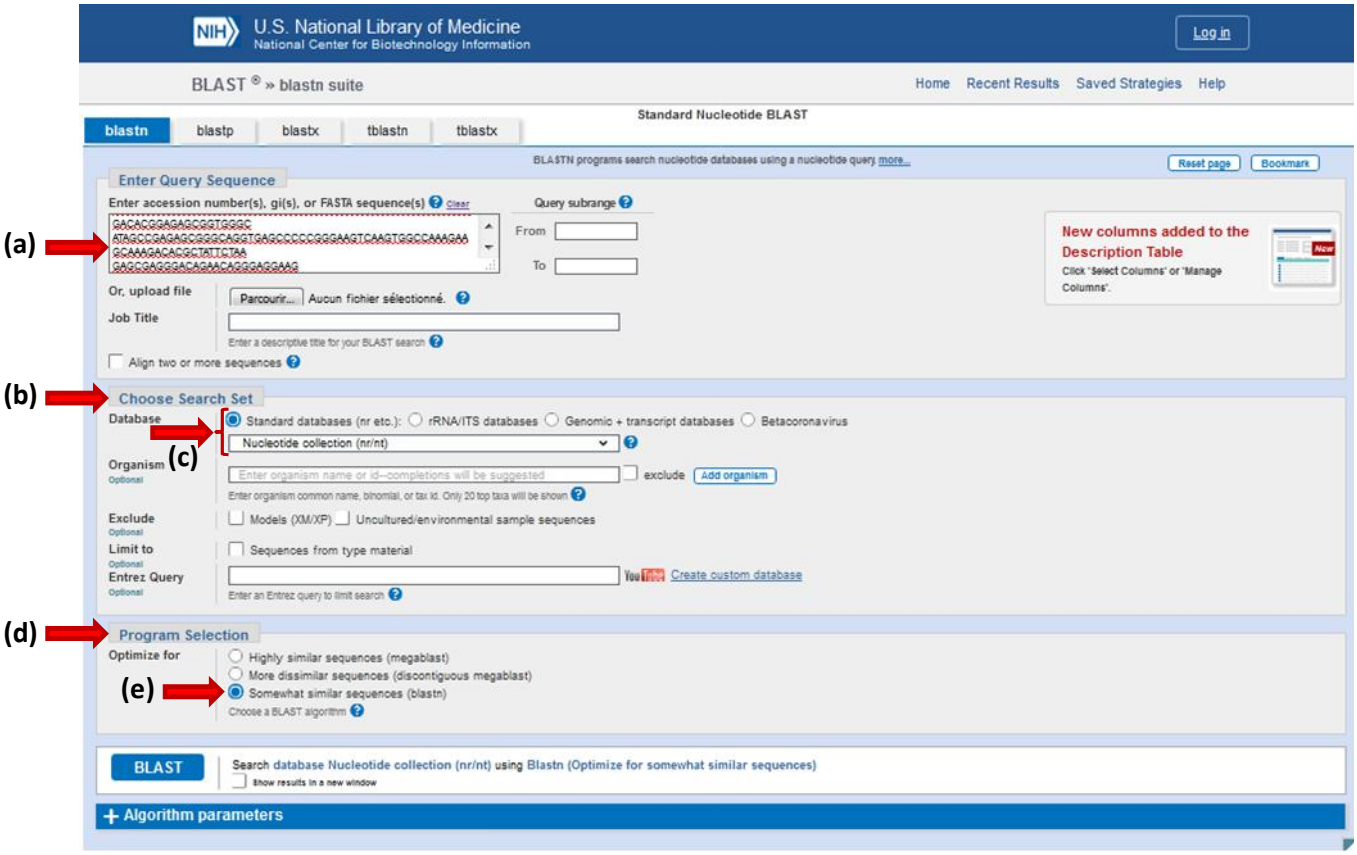
4) Cliquer sur l'option « ***Nucleotide blast*" (blastn)** ». Vous devriez obtenir la page suivante:



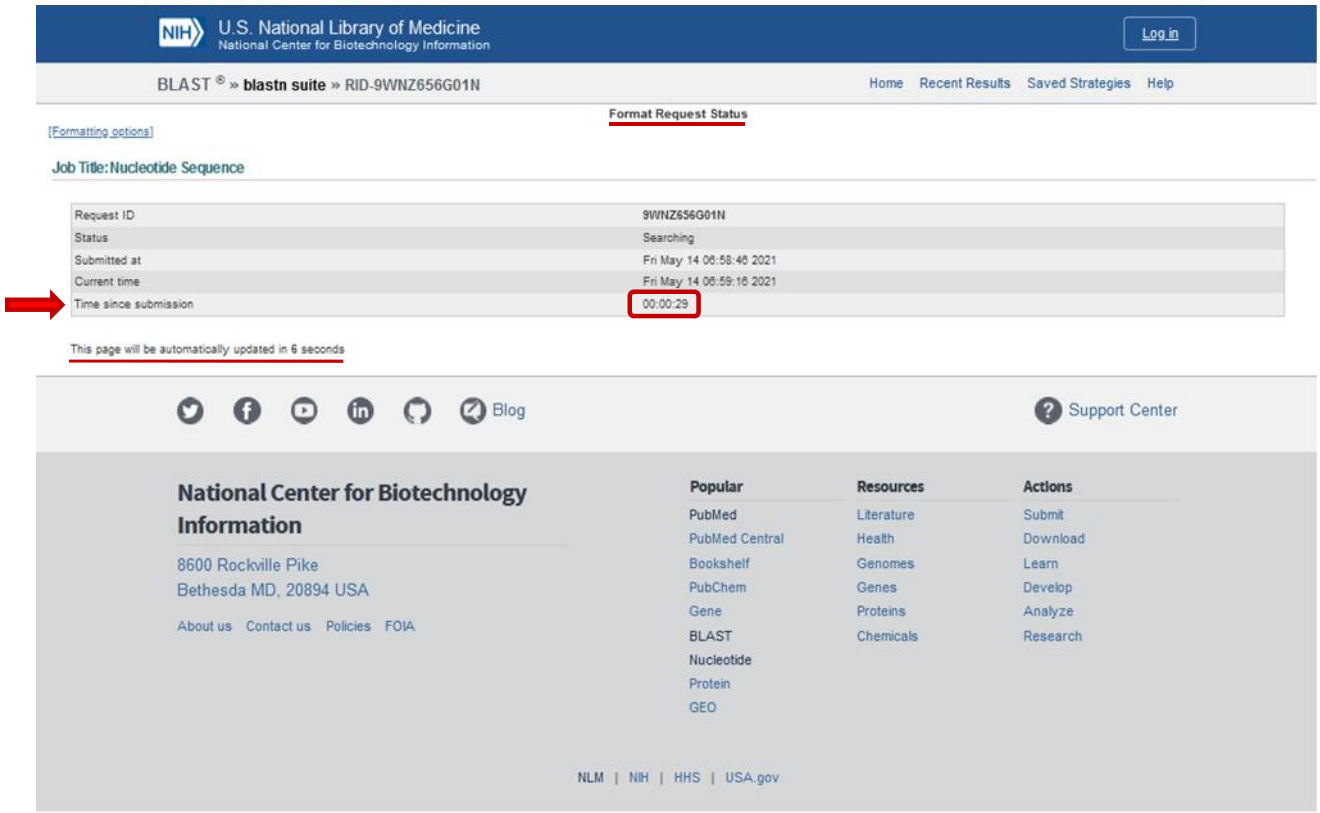
5). Avant de pouvoir faire l'entrée de la séquence d'intérêt, vous devez vous assurer que le format de celle-ci est compatible avec le logiciel. Comme mentionnez précédemment, la majorité des logiciels qui traitent des séquences peuvent comprendre un format appelé FASTA.

7) Copier et coller la séquence dans la boite de requête de « ***Nucleotide* blast** » **(a)**. Choisir la banque de données sur laquelle la recherche sera exécutée dans le menu « ***Choose Search Set*** » **(b)**. Choisir « **Standard databases (nr/nt)** » si l'origine de la séquence est inconnu « ***nucleotide collection*** **(nr/nt)** » **(c)**.

8) Maintenant, choisir le logiciel qui fera la recherche à partir du menu « ***Program Selection*** » **(d)**. Choisir « ***Somewhat similar sequences*** **(blastn)** » **(e)**. (Voir l'image ci-dessous)

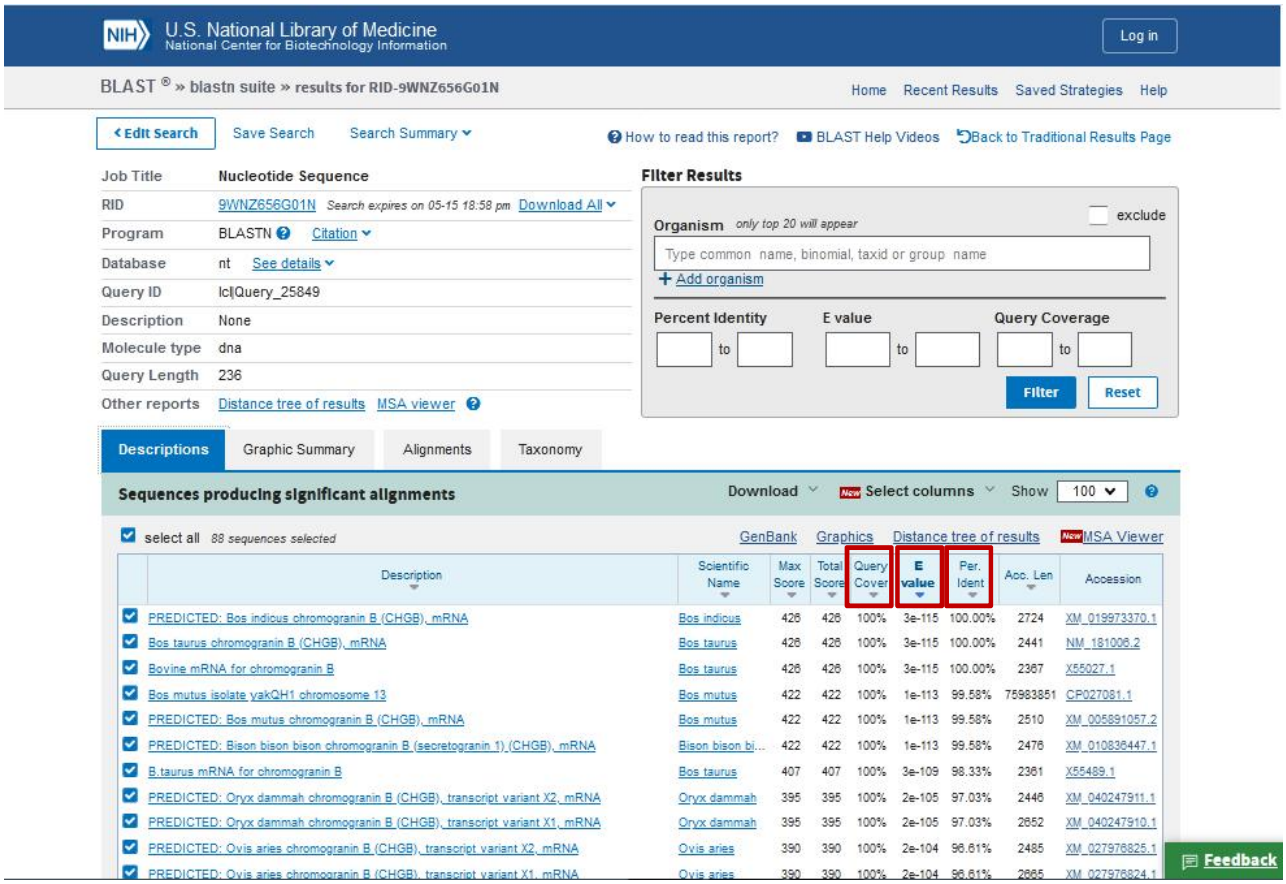


9) Cliquer sur **BLAST**. Une nouvelle page apparaîtra vous indiquant d'attendre pour que la requête soit complétée. Cela pourrait être très rapide ou très long en fonction de la charge sur le serveur de NCBI. Selon le temps d'attente vous verrez cette page:



10) Une fois que votre requête aura été complétée, une nouvelle page sera ouverte indiquant les résultats de votre recherche:

- Sous format textuel



Sous ce format, parmi l'information qui peut être retrouvée sont les valeurs suivantes:

- « **Max Score** », « **Total Score** »: En pratique, plus le score d'alignement est élevé, plus les séquences sont similaires et présenteront des propriétés et des fonctions proches (**figure 17**).
- « ***Query coverage*** » : Cette valeur indique l'étendue de la séquence (de requête) qui correspond à la séquence trouvée. Par exemple, si la requête est de 236 nucléotides et BLAST peut aligner tous les 236 nucléotides de cette requête à une correspondance, alors cela serait une couverture de 100%. A noter, « ***Query coverage*** » ne prends pas en considération la longueur de la séquence retrouvée, mais seulement le pourcentage de la requête qui s'aligne avec la correspondance.
- **« E value »** ***Expect value*** ou **valeur prévue** représente le nombre de correspondances (*Hits*) que vous devriez trouver lors d'une recherche d'une base de données de séquences aléatoires. Lorsque les valeurs E sont inférieures à 1, elles sont équivalentes à la probabilité

que deux séquences ont une certaine correspondance. Cela signifierait que si nous avons une « valeur E » de 0,01, il y a 1% de chances qu'on trouverait une correspondance dans une base de données de séquences aléatoires. Sa valeur décroit proportionnellement alors que le score augmente. Souvent, les valeurs E sont très faibles.

En fait, si nous avons une correspondance parfaite, la "valeur E" peut être donnée comme zéro. Deux facteurs supplémentaires ont une forte influence sur les valeurs E. Ce sont la longueur de la séquence et la taille de la base de données. C'est parce qu'il est plus facile de trouver une correspondance parfaite à une séquence plus courte. Il est également plus facile de trouver une correspondance dans une base de données plus grande.

- « ***Per Ident.*** » : BLAST calcule le pourcentage d'identité entre la requête et le résultat pour un alignement de nucléotide à nucléotide.

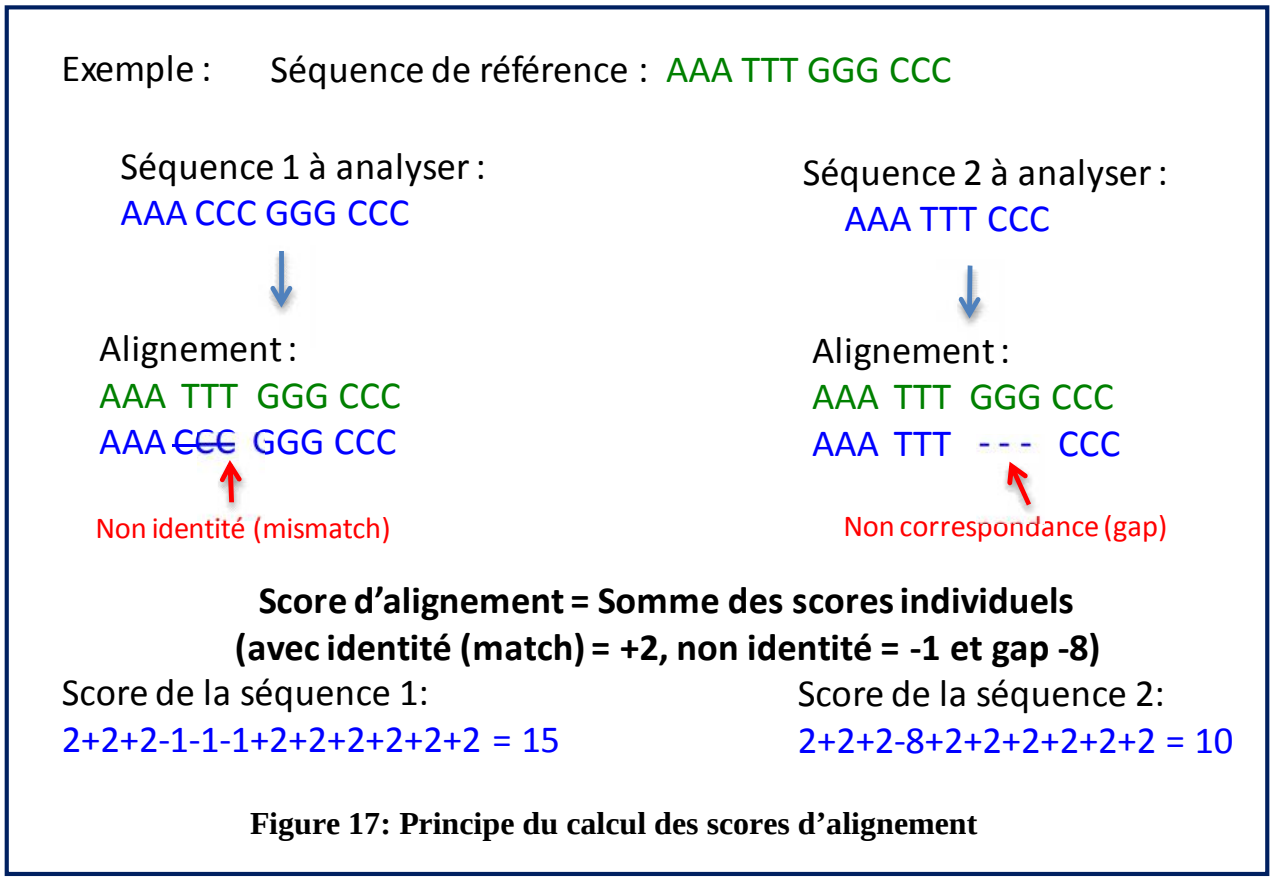
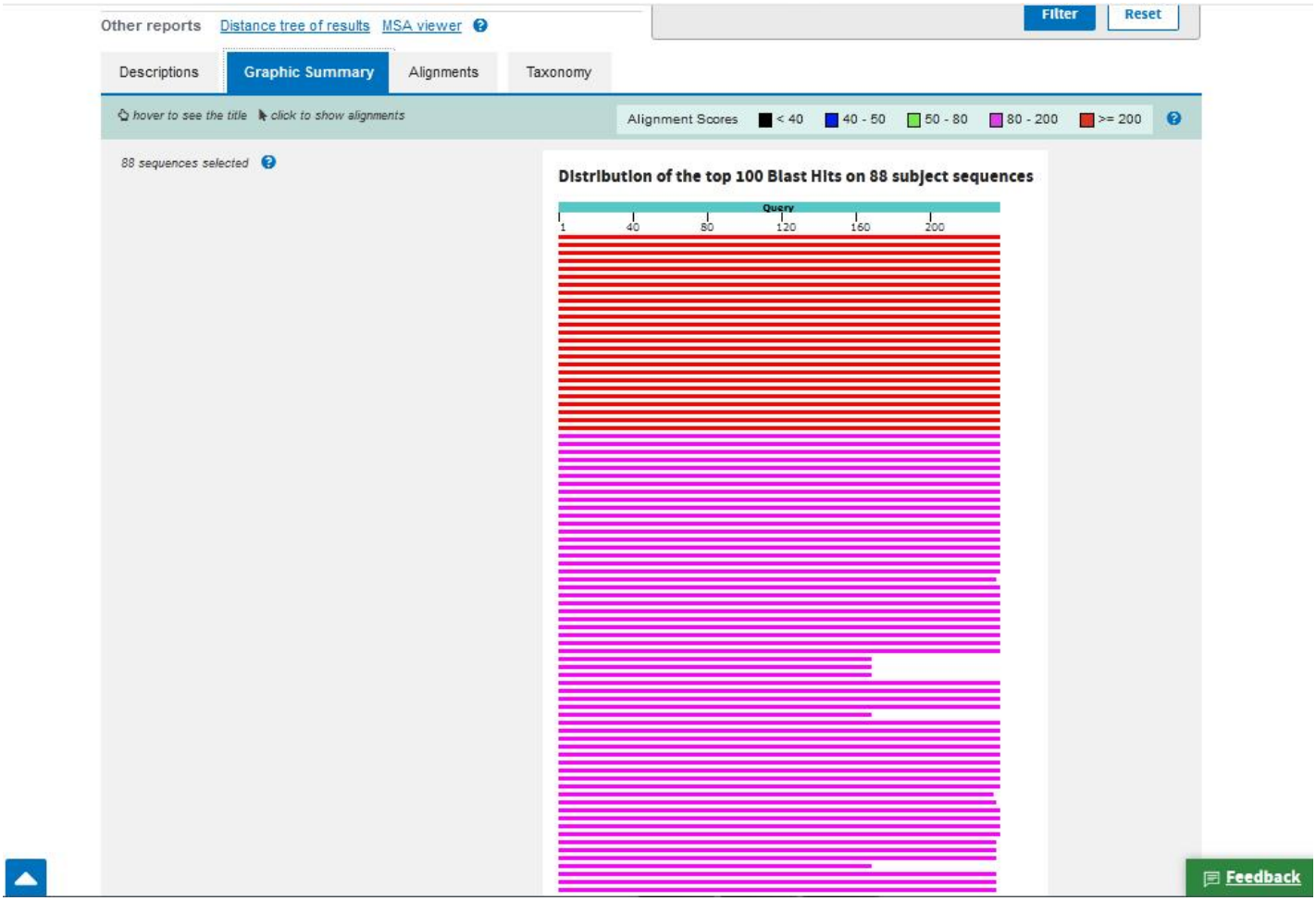


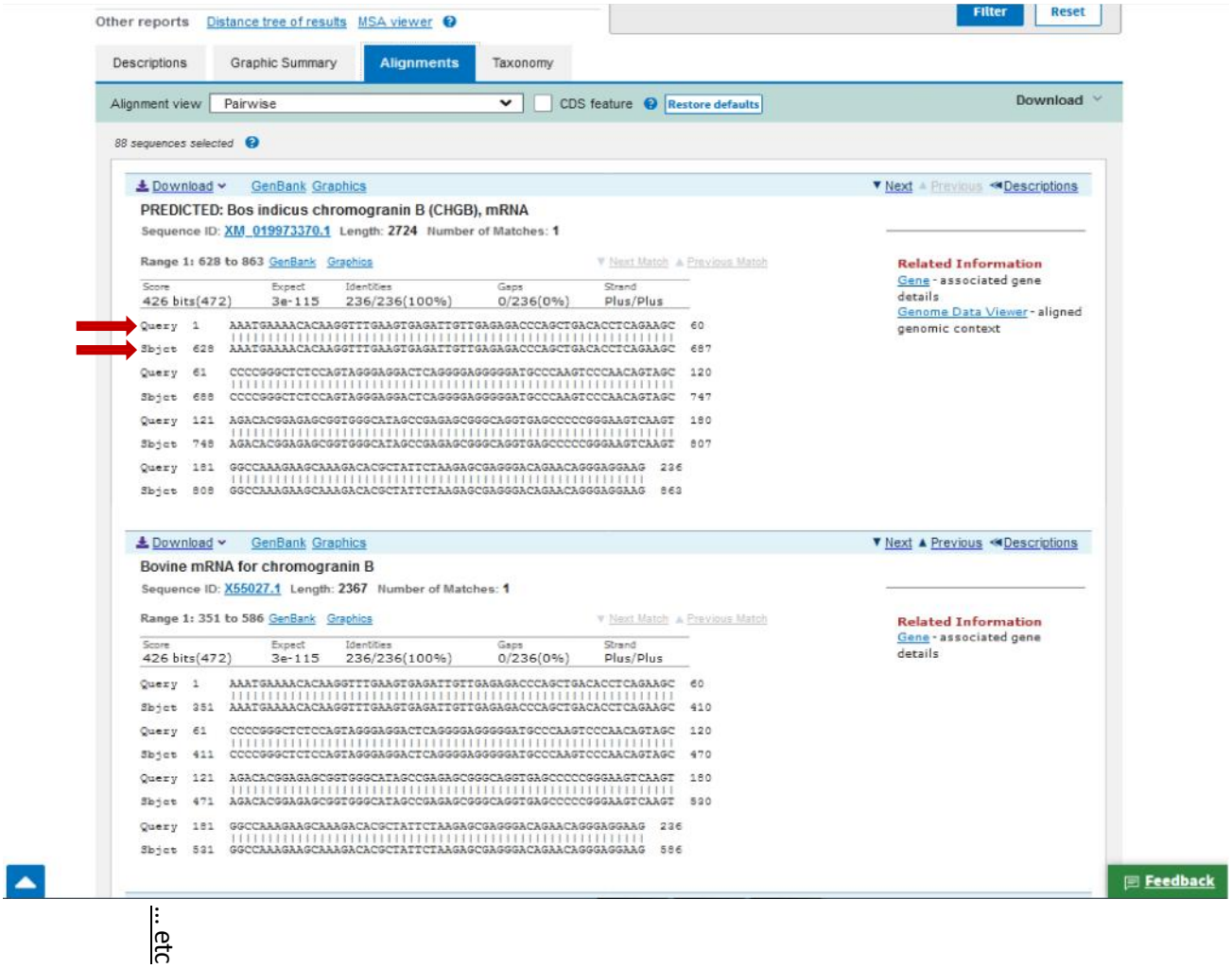
Figure 17: Principe du calcul des scores d'alignement

➢ En cliquant sur « ***Graphic Summary*** » vous obtiendrez le format graphique suivant:



Une liste des séquences trouvées sous forme graphique, avec un lien vers toutes les informations connues sur la séquence correspondante (en cliquant sur les lignes du graphique), est affichée, dans l'ordre de similarité décroissant.

➢ En cliquant sur « ***Alignments*** », vous trouverez les alignements des séquences:



... etc

Le nom et les références de la séquence similaire trouvée dans la banque sont rappelés, ainsi que le score d'alignement (voir **figure 17**), la E-value et un ensemble d'autres informations dépendantes du type d'alignement:

- Le brin (Strand)
- Le nombre de base identique
- Nombre de Gaps…

NB*:** La séquence « ***Query » est la séquence que vous avez soumis à la banque, la séquence « ***Subject*** » est la séquence trouvée dans la banque.

Comment interprétez les résultats d'un Blast:

Important: Il ne faut pas oublier que ces algorithmes d'alignement sont des heuristiques ils sont loin d'être parfaits mais ils sont très rapides. Cela implique un certain nombre d'imprécisions dans les alignements en particulier aux extrémités. Le principe de l'heuristique est de retrouver des zones identiques (ou très proches) *via* un système très rapide de recherche par index puis d'étendre les alignements à posteriori.

Il faut garder à l'esprit que la signification statistique ne reflète pas forcément la signification biologique et inversement!!!

- 30 < ID* <= 50 ➔ Séquences **faiblement similaires**
- 50 < ID <= 70 ➔ Séquences **similaires**
- 70 < ID <= 100 ➔ Séquences **fortement similaires**

***ID: Identités**

On peut déjà parler de séquences homologues au delà de 70% de similarité, mais cela reste à confirmer par d'autres hypothèses: présence de motifs communs, etc….

Si la Evalue est très faible ($<10^{-20}$), c'est probablement le signe d'une similarité entre les séquences. Mais, il ne faut jamais se fier uniquement à la Evalue.