Introduction à l'informatique génomique

Informatique Génomique - Master 1

Guillaume Blin

IGM-LabInfo UMR 8049,
Bureau 4B066
Université de Marne La Vallée
gblin@univ-mlv.fr
http://igm.univ-mlv.fr/gblin

2007-08

Plan

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Plan

La bio-informatique en quelques mots

Les supports de la bio-information
- L'ADN
- L'ARN
- Les protéines

La bio-information en informatique

Du support à l'ordinateur

Les apports de l'informatique
- Le stockage
- Les formalismes
- Le traitement des données

Conclusion

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Définition

La bio-information

- L'information liée aux molécules biologiques :
  - leurs structures
  - leurs fonctions
  - leurs interactions
  - ...
- Obtenue à partir de divers domaines d'études en biologie

Une première définition

- La bio-informatique est l'analyse de la bio-information

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Description

La bio-informatique

- Domaine de recherche récent (quelques dizaines d'années)
- Interdisciplinaire par nature
- Puisant ses fondements
  - de la biologie
  - des mathématiques
  - de l'informatique
  - et de la physique-chimie
- Utilisant le potentiel de l'informatique : théorie (algorithmes, formalismes, ...), puissance de calcul, capacité de stockage et de partage ...

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

La bio-informatique : une poule aux oeufs d'or ?

Les 4 grandes étapes

1. Acquisition de données biologiques par les approches biologiques classiques
   - in situ (dans le milieu naturel)
   - in vivo (dans l'organisme vivant)
   - in vitro (en éprouvette)

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

La bio-informatique : une poule aux oeufs d'or ?

Les 4 grandes étapes

1. Acquisition de données biologiques par les approches biologiques classiques
2. Organisation en banques de données
   - Généralistes - généralement stockage massif sans expertise de l'information contenue
   - Spécialisées - dédiées à un thème précis

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

La bio-informatique : une poule aux oeufs d'or ?

Les 4 grandes étapes

1. Acquisition de données biologiques par les approches biologiques classiques
2. Organisation en banques de données
3. Traitement des données
   - But : détecter et définir une fonction ou une structure biologique importante
   - Résultat : de nouvelles données biologiques obtenues in silico (à l'aide de l'ordinateur)

La bio-informatique : une poule aux oeufs d'or ?

Les 4 grandes étapes

1. Acquisition de données biologiques par les approches biologiques classiques
2. Organisation en banques de données
3. Traitement des données
4. Intégration des connaissances in silico
   - Combinaison des données initiales et des données obtenues in silico

La bio-informatique : une poule aux oeufs d'or ?

Les 4 grandes étapes

1. Acquisition de données biologiques par les approches biologiques classiques
2. Organisation en banques de données
3. Traitement des données
4. Intégration des connaissances in silico

Ces nouvelles connaissances

- aboutissent au développement de nouveaux concepts biologiques
- nécessitent l'élaboration de nouvelles théories et outils informatiques

Une petite chronologie

de Pascaline au Human Genome Project

- 1646 : La Pascaline
- 1840 : Naissance de l'algorithmique
- 1854 : Algèbre de Boole
- 1858 : Premier câble télégraphique transatlantique
- 1866 : Lois de Mendel
- 1876 : Le téléphone
- 1890 : La machine à recenser
- 1900 : La redécouverte des lois de Mendel

1646

La Pascaline

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

Une petite chronologie

de Pascaline au Human Genome Project

- 1646 : La Pascaline
- 1840 : Naissance de l'algorithmique
- 1854 : Algèbre de Boole
- 1858 : Premier câble télégraphique transatlantique
- 1866 : Lois de Mendel
- 1876 : Le téléphone
- 1890 : La machine à recenser
- 1900 : La redécouverte des lois de Mendel

1840

L'algorithmique

- Ada Lovelace, mathématicienne britanique, définit le principe des itérations successives d'opérations dans l'exécution d'un programme
- En l'honneur du mathématicien Al Khowarizmi (780-850), elle nomme le processus logique d'exécution d'un programme : algorithme

Une petite chronologie

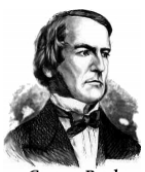
de Pascaline au Human Genome Project

- 1646 : La Pascaline
- 1840 : Naissance de l'algorithmique
- 1854 : Algèbre de Boole
- 1858 : Premier câble télégraphique transatlantique
- 1866 : Lois de Mendel
- 1876 : Le téléphone
- 1890 : La machine à recenser
- 1900 : La redécouverte des lois de Mendel

1854

Algèbre de Boole

- Georges Boole développe une nouvelle forme de logique, à la fois symbolique et mathématique
- C'est une algèbre binaire n'acceptant que deux valeurs numériques : 0 et 1 ; et munie de deux lois de composition interne (le ET et le OU)
- A l'origine des ordinateurs à arithmétique binaire



George Boole

Une petite chronologie

de Pascaline au Human Genome Project

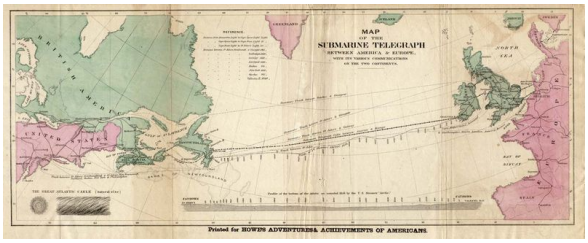
- 1646 : La Pascaline
- 1840 : Naissance de l'algorithmique
- 1854 : Algèbre de Boole
- 1858 : Premier câble télégraphique transatlantique
- 1866 : Lois de Mendel
- 1876 : Le téléphone
- 1890 : La machine à recenser
- 1900 : La redécouverte des lois de Mendel

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1858

1er câble transatlantique

- Reliant Valentia une petite île irlandaise à Trinity Bay au Canada (3 200 km de distance)
- Vitesse de transmission en morse : 2,75 mots par minute



G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

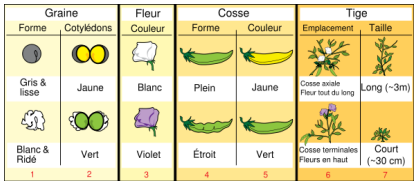
- 1646 : La Pascaline
- 1840 : Naissance de l'algorithmique
- 1854 : Algèbre de Boole
- 1858 : Premier câble télégraphique transatlantique
- 1866 : Lois de Mendel
- 1876 : Le téléphone
- 1890 : La machine à recenser
- 1900 : La redécouverte des lois de Mendel

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1866

Lois de Mendel

- Le moine botaniste tchèque Gregor Mendel (1822-1884) réalise des expériences sur l'hybridation des plantes
- Il publie les lois de la transmission des caractères héréditaires en se basant sur l'observation de la transmission des caractéristiques morphologiques (au nombre de 7) de pois à travers plusieurs générations



G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1646 : La Pascaline
- 1840 : Naissance de l'algorithmique
- 1854 : Algèbre de Boole
- 1858 : Premier câble télégraphique transatlantique
- 1866 : Lois de Mendel
- 1876 : Le téléphone
- 1890 : La machine à recenser
- 1900 : La redécouverte des lois de Mendel

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1876

Le téléphone

- Graham Bell, américain, dépose le brevet du téléphone en 1876
- Le développement du téléphone sera énorme aux Etats Unis (25.000 postes en 1879)
- Graham fondera la Bell Telephon Company

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1646 : La Pascaline
- 1840 : Naissance de l'algorithmique
- 1854 : Algèbre de Boole
- 1858 : Premier câble télégraphique transatlantique
- 1866 : Lois de Mendel
- 1876 : Le téléphone
- 1890 : La machine à recenser
- 1900 : La redécouverte des lois de Mendel

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction

1890

La machine à recenser

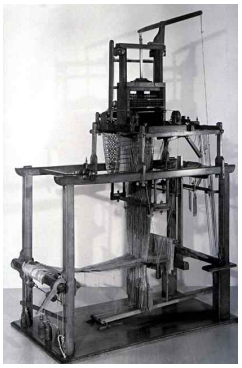
- Herman Hollerith *invente* la machine à cartes perforées pour accélérer le recensement de 1890 aux États-Unis traité en trois ans seulement (au lieu des 9 de celui de 1880)
- Il fonde une compagnie qui sera rebaptisée en 1917 International Business Machine

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction

1801

Le métier à tisser Jacquard

- Le métier Jacquard est un métier à tisser mis au point par le lyonnais Joseph Marie Jacquard en 1801
- La machine Jacquard combine les techniques des aiguilles de Basile Bouchon, les cartes perforées de Falcon et du cylindre de Vaucanson
- Cette utilisation de cartes perforées fait qu'il est parfois considéré comme l'ancêtre de l'ordinateur



Une petite chronologie

de Pascaline au Human Genome Project

- 1646 : La Pascaline
- 1840 : Naissance de l'algorithmique
- 1854 : Algèbre de Boole
- 1858 : Premier câble télégraphique transatlantique
- 1866 : Lois de Mendel
- 1876 : Le téléphone
- 1890 : La machine à recenser
- 1900 : La redécouverte des lois de Mendel

1900

La redécouverte des lois de Mendel

- La revue de la « Société Allemande de Botanique » publie un texte du botaniste hollandais Hugo de Vries intitulé « La loi de disjonction des hybrides » qui traite de la transmission des caractères héréditaires au moment de l'hybridation des végétaux
- Sans avoir lu les conclusions de Mendel sur la transmission des caractères des plantes, il aboutit aux mêmes conclusions
- Il fut avec Carl Correns et Erich von Tschermak-Seysenegg, l'un des trois scientifiques qui redécouvrirent les lois de Mendel

Une petite chronologie

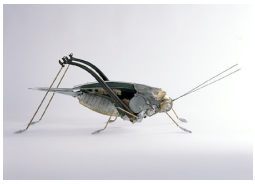
de Pascaline au Human Genome Project

- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

1902

La théorie chromosomique de l'hérédité

- L'américain Walter S. Sutton, s'appuyant sur une étude morphologique des chromosomes de sauterelles, suggère que les chromosomes vont par paires
- Il suggère également que les chromosomes sont le support de l'hérédité et qu'une copie de chaque chromosome est héritée d'un parent durant la méiose



Une petite chronologie

de Pascaline au Human Genome Project

- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

1909

Le gène

- Wilhelm Johannsen, un botaniste danois, introduit le terme gène, provenant du mot grec qui signifie *donner naissance à*, pour décrire les unités d'informations génétiques transmises d'une génération à une autre

Une petite chronologie

de Pascaline au Human Genome Project

- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

1910

Les mouches du vinaigre

- Thomas Hunt Morgan, biologiste américain, consolide la théorie chromosomique de l'hérédité de Sutton en étudiant les mouches du vinaigre...
- Morgan se consacre aux drosophiles sauvages (mouches aux yeux rouges) et en découvre avec des yeux blancs
- Il en déduit qu'à la suite d'une mutation, une modification d'un caractère héréditaire peut subvenir

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

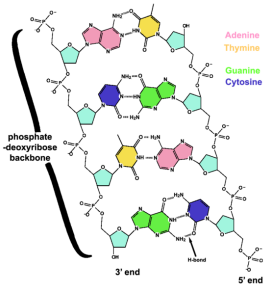
- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1929

ADN

- C'est le chimiste Phoebus Aaron Levene qui découvre l'Acide DésoxyriboNucléqique (ADN)
- Il détermine que l'ADN contient les éléments d'adénine, de guanine, de thymine, de cytosine, le désoxyribose, et un groupe phosphate



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1943

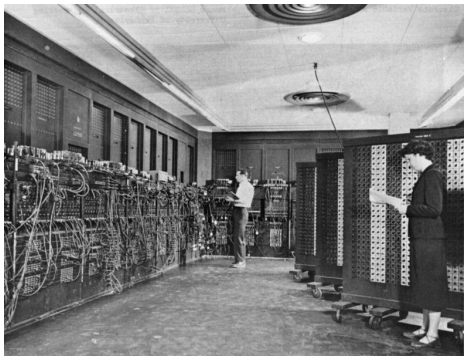
Eniac

- John Mauchley a construit *un des plus impressionnant calculateur* – Electronic Numerical Integrator and Calculator ; *le premier ordinateur électronique*
- Financé par l'armée américaine, il pesait 30 tonnes, mesurait 10m sur 17 et tombait souvent en panne
- Il fonctionnait à l'aide de cartes perforées et ne permettait pas de faire autant de choses qu'une simple calculatrice programmable actuelle
- Lorsqu'il fonctionnait, il lui suffisait de 0.2 millisecondes pour traiter une addition

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1943

Eniac



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1943

Eniac

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

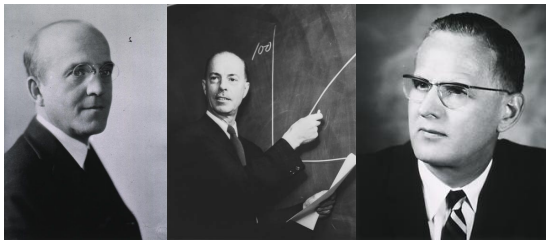
de Pascaline au Human Genome Project

- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

1944

L'ADN - transporteur d'informations

- Le bactériologiste et physicien Oswald Avery démontre avec Colin Mc Leod et Mc Lyn McCarthy que c'est l'ADN qui contient l'information héréditaire



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

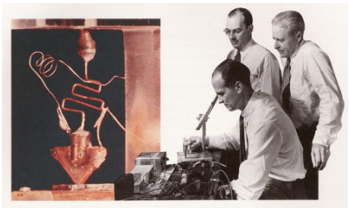
- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1951

Le transistor

- John Bardeen, Walter Brattain et William Shockley inventèrent au Bell Labs le transistor
- C'est un composant électronique actif fondamental en électronique utilisé principalement comme interrupteur commandé et pour l'amplification



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

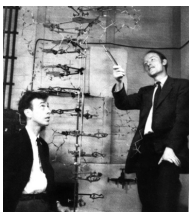
- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1953

L'ADN est une double hélice

- Les biologistes James Watson et Francis Crick, en se basant sur les analyses cristallographiques aux rayons X de l'ADN, proposent la structure en double hélice de la molécule d'ADN, publiée le 25 avril 1953 dans la revue Nature



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1953

Premier ordinateur commercial

- IBM lance son premier ordinateur commercial en série : l'IBM 650
- Bien que lent, peu fiable et coûteux, un millier d'exemplaires seront fabriqués

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1902 : La théorie chromosomique de l'hérédité
- 1909 : Le gène
- 1910 : Les mouches du vinaigre
- 1929 : L'ADN
- 1943 : Le premier calculateur électronique
- 1944 : L'ADN - transporteur d'informations
- 1951 : Transistor
- 1953 : Structure de l'ADN
- 1953 : Premier ordinateur commercial
- 1957 : Dogme de la biologie moléculaire

1957

Dogme de la biologie moléculaire

- Francis Crick et George Gamov définissent le dogme de la biologie moléculaire
- Ce dogme explique les mécanismes permettant de passer de l'information contenue dans les gènes aux protéines

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

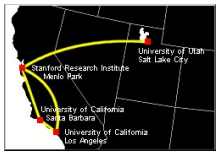
- 1969 : ARPANET
- 1970 : Algorithme d'alignement global de séquences
- 1975-76 : Début d'Apple et Microsoft
- 1977 : Méthode de séquençage
- 1978 : Séquencage premier génome
- 1980 : Première banque de données EMBL
- 1981 : 1er PC d'IBM (DOS)
- 1982 : Genbank

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1969

ARPANET

- Len Kleinrock aidé de deux étudiants du UCLA, ont créé le premier réseau, en reliant deux ordinateurs via un câble de 4,5 mètres
- Arpanet est créé en reliant quatre ordinateurs situés chacun dans des centres universitaires différents (chiffre qui passe à 35 en 1973). C'est l'ancêtre d'internet
- Le premier message envoyé fut LO (début du mot LOGIN - car il y eu un problème lors de l'envoi du G)



G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

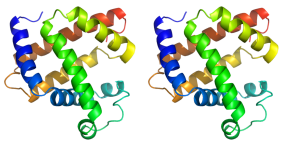
- 1969 : ARPANET
- 1970 : Algorithme d'alignement global de séquences
- 1975-76 : Début d'Apple et Microsoft
- 1977 : Méthode de séquençage
- 1978 : Séquencage premier génome
- 1980 : Première banque de données EMBL
- 1981 : 1er PC d'IBM (DOS)
- 1982 : Genbank

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1970

Algorithme d'alignement global de séquences

- Saul Needleman et Christian Wunsch proposent une méthode d'alignement global de séquences applicable à la recherche de similarités entre deux protéines



G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1969 : ARPANET
- 1970 : Algorithme d'alignement global de séquences
- 1975-76 : Début d'Apple et Microsoft
- 1977 : Méthode de séquençage
- 1978 : Séquencage premier génome
- 1980 : Première banque de données EMBL
- 1981 : 1er PC d'IBM (DOS)
- 1982 : Genbank

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1975-76

Début d'Apple et Microsoft

- Le 4 avril 1975, Microsoft Corporation, entreprise développant des logiciels pour ordinateurs, est créée à Albuquerque (Nouveau Mexique) par deux étudiants américains : Bill Gates et Paul Allen
- Le 1er avril 1976 grâce à Steve Wozniak et Steve Jobs, la société Apple Computer voit le jour

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction
En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

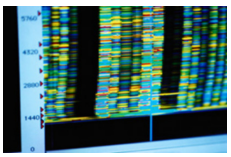
- 1969 : ARPANET
- 1970 : Algorithme d'alignement global de séquences
- 1975-76 : Début d'Apple et Microsoft
- 1977 : Méthode de séquençage
- 1978 : Séquencage premier génome
- 1980 : Première banque de données EMBL
- 1981 : 1er PC d'IBM (DOS)
- 1982 : Genbank

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction

1977

Méthode de séquençage

- Walter Gilbert (USA) et Frederick Sanger (UK) développent indépendamment deux méthodes permettant de trouver l'ordre des éléments A,C, G et T sur un simlpe brin d'ADN



Une petite chronologie

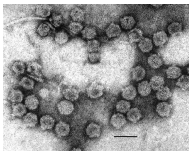
de Pascaline au Human Genome Project

- 1969 : ARPANET
- 1970 : Algorithme d'alignement global de séquences
- 1975-76 : Début d'Apple et Microsoft
- 1977 : Méthode de séquençage
- 1978 : Séquencage premier génome
- 1980 : Première banque de données EMBL
- 1981 : 1er PC d'IBM (DOS)
- 1982 : Genbank

1978

Premier génome séquencé

- Le génome du bactériophage (un virus n'infectant que des bactéries) phiX174 est séquencé par F. Sanger



Une petite chronologie

de Pascaline au Human Genome Project

- 1969 : ARPANET
- 1970 : Algorithme d'alignement global de séquences
- 1975-76 : Début d'Apple et Microsoft
- 1977 : Méthode de séquençage
- 1978 : Séquencage premier génome
- 1980 : Première banque de données EMBL
- 1981 : 1er PC d'IBM (DOS)
- 1982 : Genbank

1980

Première banque de données EMBL

- Création de la première base de données de séquences biologiques
- Banque européenne généraliste

European Molecular Biology Laboratory
Heidelberg Grenoble
Hamburg Hinxton Monterotondo

Une petite chronologie

de Pascaline au Human Genome Project

- 1969 : ARPANET
- 1970 : Algorithme d'alignement global de séquences
- 1975-76 : Début d'Apple et Microsoft
- 1977 : Méthode de séquençage
- 1978 : Séquencage premier génome
- 1980 : Première banque de données EMBL
- 1981 : 1er PC d'IBM (DOS)
- 1982 : Genbank

1981

1er PC d'IBM

- Le premier micro-ordinateur lancé par IBM :le 5150 ou PC/G

Une petite chronologie

de Pascaline au Human Genome Project

- 1969 : ARPANET
- 1970 : Algorithme d'alignement global de séquences
- 1975-76 : Début d'Apple et Microsoft
- 1977 : Méthode de séquençage
- 1978 : Séquencage premier génome
- 1980 : Première banque de données EMBL
- 1981 : 1er PC d'IBM (DOS)
- 1982 : Genbank

1982

Genbank

- Création de la base de données GenBank
- Banque américaine généraliste créée par la société IntelliGenetics et diffusée aujourd'hui par le National Center of Biotechnilogy Information

Une petite chronologie

de Pascaline au Human Genome Project

- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

1985

Idée de l'HGP

- L'idée de décrypter le génome humain naît pour la première fois au Imperial Cancer Research à Londres
- Il faudra attendre 5 ans avant le début du dit projet

Une petite chronologie

de Pascaline au Human Genome Project

- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

1989

Début de HGP

- Le Human Genome Project est un consortium regroupant des laboratoires de différents pays
- Ce projet est financé par des fonds publiques et caritatifs

Une petite chronologie

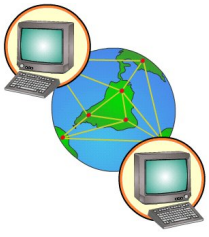
de Pascaline au Human Genome Project

- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

1989

Internet

- Le réseau ARPAnet s'ouvre au trafic commercial
- C'est la naissance d'Internet (Inter Network) – un réseau informatique à l'échelle du monde rendant accessible au public des services comme le courrier électronique et le web



Une petite chronologie

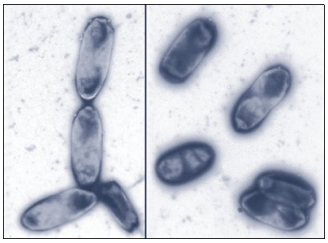
de Pascaline au Human Genome Project

- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

1995

Haemophilus Influenzae

- Haemophilus Influenzae est la première bactérie séquencée - 2 MB
- C'est une bactérie qui représente prés de 40 % des causes bactériennes des otites moyennes



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

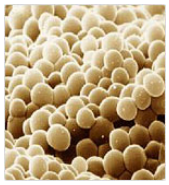
- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1996

Saccharomyces cerevisiae

- Saccharomyces cerevisiae est une levure utilisée depuis l'aube de l'humanité dans l'élaboration du pain, du vin et de la bière
- C'est le premier organisme unicellulaire eucaryote (avec noyau) séquencé - 12 MB



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1998

Caenorhabditis elegans

- Caenorhabditis elegans est un petit ver transparent d'environ un millimètre de long
- C'est le premier organisme pluricellulaire séquencé - 100 MB

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

1998

Celera Genomics

- C'est une entreprise fondée en mai 1998
- Créée dans le but de générer puis commercialiser des informations génétiques afin d'accélérer la compréhension des processus biologiques

CELERA
an Applera Corporation Business

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Une petite chronologie

de Pascaline au Human Genome Project

- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

2000

Arabidopsis Thaliana

- L'arabette des dames (Arabidopsis Thaliana) fut le premier génome végétal séquencé - 157 MB

2001

1er brouillon du génome humain

- La publication officielle des deux séquences "brutes" (1 par le consortium international et 1 par Celera Genomics)
- Celera Genomics annonce avoir utilisé non seulement ses propres données, mais aussi celles publiées en ligne au fur et à mesure par le consortium international
- Les séquences publiées en 2001 étaient des ébauches ; il y restait encore un grand nombre de trous et d'imperfections
- La séquence complète a été terminée en 2004 par le consortium international public

Les acides nucléiques et les protéines

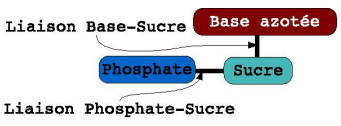
La localisation de la bio-information

- Il y a deux types de molécules supports de la bio-information :
  - les acides nucléiques (ADN et ARN)
  - les protéines
- Les relations entre ces molécules sont définies dans le dogme central de la biologie moléculaire : l'ADN se réplique, est transcrit en ARN qui est éventuellement traduit en protéine

Rappels de biologie

Un nucléotide

- Un nucléotide est composé de trois substances fondamentales :
  - un groupe phosphate
  - un sucre à 5 atomes de carbone, le désoxyribose
  - une base azotée



Une petite chronologie

de Pascaline au Human Genome Project

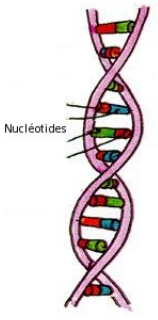
- 1985 : Idée de HGP
- 1989 : début de HGP
- 1989 : Internet
- 1995 : 1ère bactérie séquencée
- 1996 : 1er génome eucaryote
- 1998 : 1er génome d'organisme pluricellulaire
- 1998 : Celera Genomics
- 2000 : 1er génome de plante
- 2001 : publication de la séquence "brute"

Plan

Rappels de biologie

Acide DésoxyriboNucléique

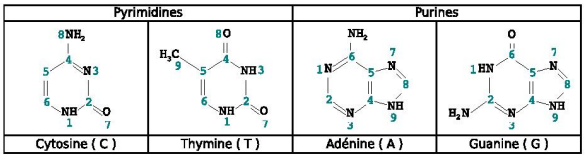
- L'ADN est une longue molécule présente chez tous les êtres vivants
- Sa structure est en forme d'une double hélice
- Chaque hélice est constituée d'une suite d'éléments appelés nucléotides



Rappels de biologie

Un nucléotide

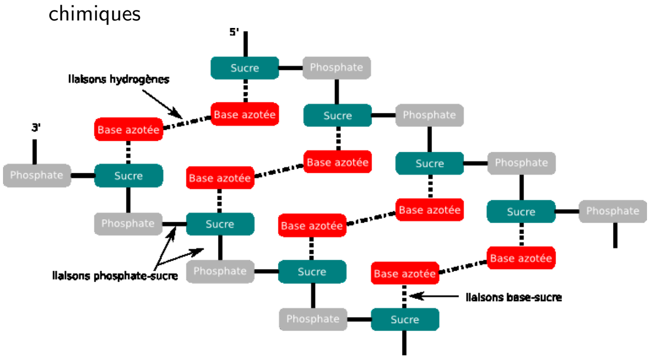
- Les bases azotées entrant dans la composition des nucléotides sont séparées en 2 familles : les purines et les pyrimidines
- La base azotée constitue la partie variable d'un nucléotide et détermine, par conséquent, sa nature



Rappels de biologie

Les liaisons

- La stabilité de la molécule d'ADN est due à des liaisons chimiques



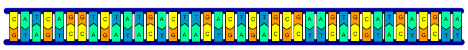
Rappels de biologie

La structure de l'ADN

- Les bases azotées sont au centre de la double hélice
- Ce sont les liaisons hydrogènes qui, par leur souplesse, permettent cette structure
- Le squelette de la structure correspond à deux brins (l'alternance des phosphates et sucres)
- Les bases azotées s'apparient comme suit : $A - T$ et $C - G$

Rappels de biologie

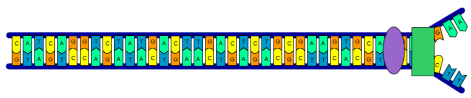
La réplication de l'ADN



- Au départ : une double hélice d'ADN

Rappels de biologie

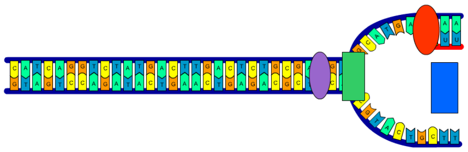
La réplication de l'ADN



- Sous l'action de l'ADN hélicase, l'ADN s'ouvre comme une fermeture éclaire

Rappels de biologie

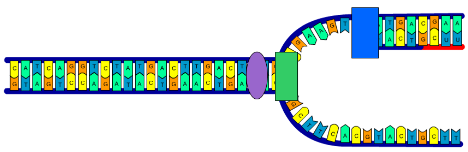
La réplication de l'ADN



- L'ADN hélicase tout en se déplaçant sur l'ADN sépare les deux brins. Des ADN dits polymérases assemblent des bases azotées par complémentarité tout au long du brin

Rappels de biologie

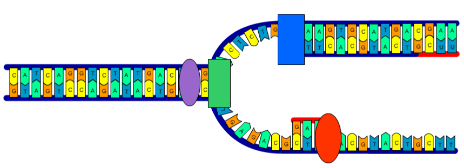
La réplication de l'ADN



- L'ADN hélicase tout en se déplaçant sur l'ADN sépare les deux brins. Des ADN dits polymérases assemblent des bases azotées par complémentarité tout au long du brin

Rappels de biologie

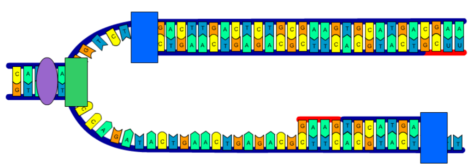
La réplication de l'ADN



- L'ADN hélicase tout en se déplaçant sur l'ADN sépare les deux brins. Des ADN dits polymérases assemblent des bases azotées par complémentarité tout au long du brin

Rappels de biologie

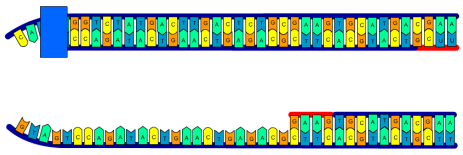
La réplication de l'ADN



- L'ADN hélicase tout en se déplaçant sur l'ADN sépare les deux brins. Des ADN dits polymérases assemblent des bases azotées par complémentarité tout au long du brin

Rappels de biologie

La réplication de l'ADN

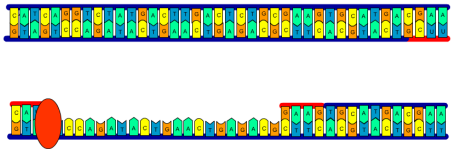


- L'ADN hélicase tout en se déplaçant sur l'ADN sépare les deux brins. Des ADN dits polymérases assemblent des bases azotées par complémentarité tout au long du brin

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Rappels de biologie

La réplication de l'ADN

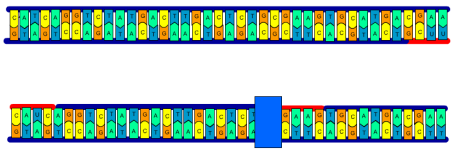


- Une fois les deux brins séparés, d'autres ADN polymérases complètent la synthèse des deux brins. On obtient alors deux molécules identiques d'ADN

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Rappels de biologie

La réplication de l'ADN



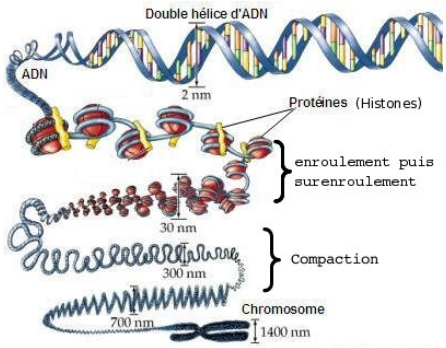
- Une fois les deux brins séparés, d'autres ADN polymérases complètent la synthèse des deux brins. On obtient alors deux molécules identiques d'ADN

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Rappels de biologie

Chromosome

- Le chromosome - support de l'information génétique



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Rappels de biologie

Acide RiboNucléique

- L'ARN est une molécule courte issue à partir de l'ADN
- Sa structure est simple brin et composée de nucléotides d'Adénine, Cytosine, Guanine et Uracile
- La stabilité de l'ARN est due aux mêmes liaisons que l'ADN
- L'ARN se replie dans l'espace par l'action de liaisons hydrogènes pouvant se former : A-U, C-G (principalement)
- La structure est importante car elle détermine en partie la fonction de l'ARN

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Rappels de biologie

Le rôle de l'ARN

- L'ARN est, entre autres, le messager de l'information génétique
- Il en existe différents types avec des rôles particuliers dans le processus de la synthèse de protéines
  - ARN messager
  - ARN de transfert
  - ARN ribosomal
  - ARN polymerase
  - ...

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

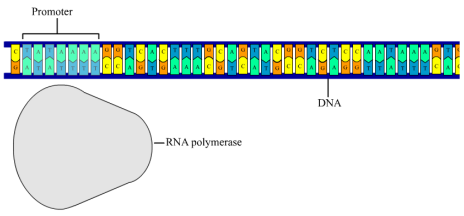
Rappels de biologie

La transcription

- Passage de l'information génétique de l'ADN à l'ARN
- Il s'effectue en trois étapes :
  1. L'initiation, qui se fait au niveau d'une région particulière (promoteur),
  2. La synthèse, qui nécessite l'ouverture de l'ADN,
  3. La terminaison, qui se fait au niveau d'une région particulière
- L'entité biologique à l'origine de la transcription est appelée ARN polymérase

Rappels de biologie

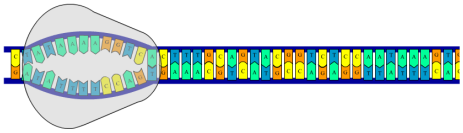
La transcription en images



- Au départ : une région promotrice, un ARN polymérase, de l'ADN

Rappels de biologie

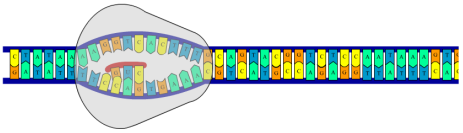
La transcription en images



- L'ARN polymérase se fixe au promoteur et créé une élongation des deux brins d'ADN

Rappels de biologie

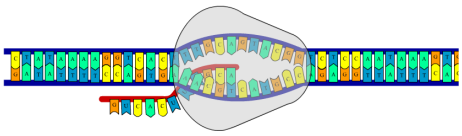
La transcription en images



- L'ARN polymérase tout en se déplaçant sur l'ADN, assemble des bases azotées par complémentarité avec les bases de la séquence du brin d'ADN

Rappels de biologie

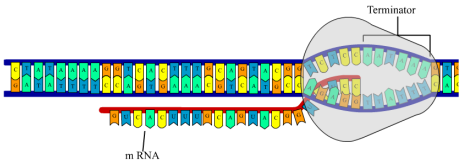
La transcription en images



- L'ARN polymérase tout en se déplaçant sur l'ADN, assemble des bases azotées par complémentarité avec les bases de la séquence du brin d'ADN

Rappels de biologie

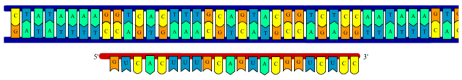
La transcription en images



- La transcription s'achève lorsque l'ARN polymérase rencontre la séquence de fin de transcription

Rappels de biologie

La transcription en images

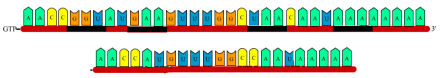


- L'ADN, l'ARN polymérase et l'ARN ainsi transcrit (appelé transcrit primaire) se séparent

Rappels de biologie

Le devenir du transcrit primaire

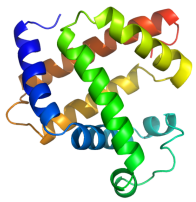
- Dans les organismes procaryotes (sans noyau), le transcrit primaire est opérationnel tout de suite
- En revanche, chez les eucaryotes, étant constitué de séquences codantes (exons) interrompues par des séquences non codantes (introns) ; il doit être débarrassé de ses introns - c'est l'épissage
- Dans les deux cas, l'ARN ainsi obtenu est appelé ARN messager ou ARNm



Rappels de biologie

Les protéines

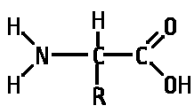
- Une protéine est une longue molécule constituée d'une suite d'éléments appelés acides aminés
- On estime à 100 000, le nombre de types de protéines fabriquées par l'être humain
- et à 15 000 types différents fabriqués en moyenne par cellule



Rappels de biologie

Un acide aminé

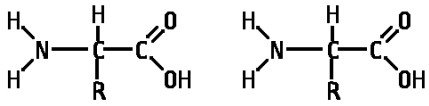
- Un acide aminé est composé d'un atome de carbone auquel sont liés :
  - un groupement amine (NH_2)
  - un groupement acide (COOH)
  - une portion variable d'un acide aminé à l'autre



Rappels de biologie

Les liaisons

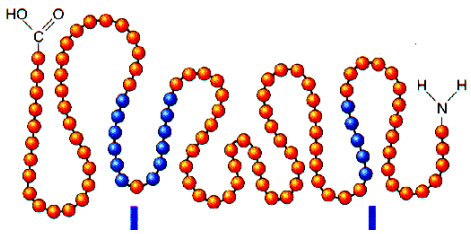
- D'un point de vue chimique, une protéine est un enchaînement d'acides aminés liés par des liaisons peptidiques
- Une liaison peptidique s'effectue entre le groupement acide et le groupement amine de deux acides aminés



Rappels de biologie

La structure d'une protéine

- Les protéines possèdent quatre niveaux de structures

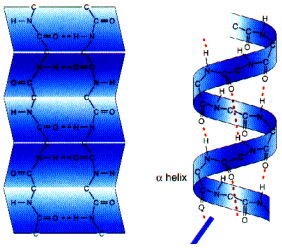


Structure primaire : la séquence d'a.a.

Rappels de biologie

La structure d'une protéine

- Les protéines possèdent quatre niveaux de structures

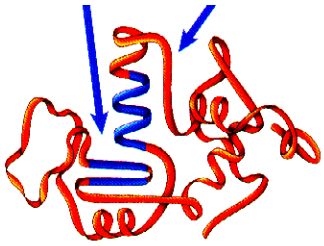


Structure secondaire : feuillet bêta et hélice alpha

Rappels de biologie

La structure d'une protéine

- Les protéines possèdent quatre niveaux de structures



Structure tertiaire : repliement en 3D

Rappels de biologie

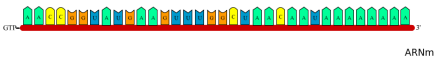
La structure d'une protéine

- Les protéines possèdent quatre niveaux de structures

Structure quaternaire : association de protéines ayant des affinités

Rappels de biologie

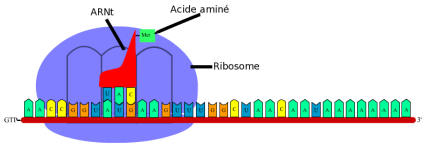
La traduction en images



- Au départ : un ARN messager

Rappels de biologie

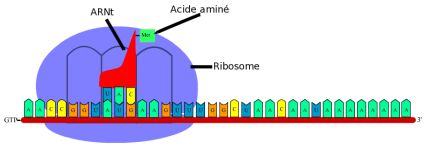
La traduction en images



- Un ribosome se fixe à l'ARNm
- Il va assembler une séquence d'acides aminés selon les *instructions* du code génétique : chaque codon (groupe de 3 bases) correspond à un acide aminé

Rappels de biologie

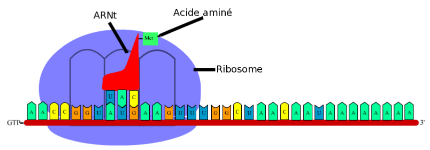
La traduction en images



- Le ribosome va parcourir le brin d'ARNm codon par codon
- Le codon AUG, appelé codon-initiateur, va permettre de commencer la traduction en formant l'acide aminé méthionine - qui se détachera plus tard de la chaîne polypeptidique

Rappels de biologie

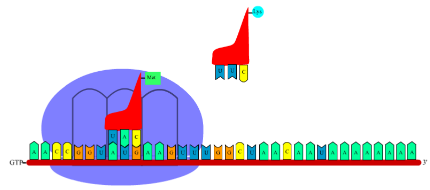
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

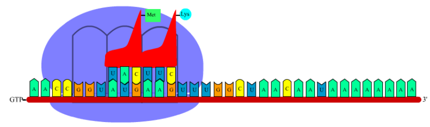
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

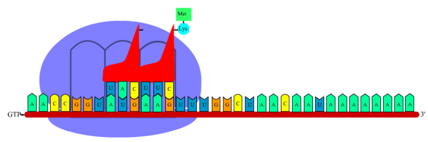
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

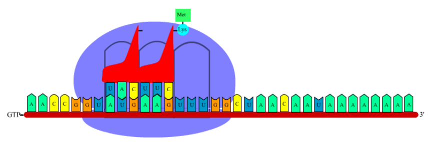
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

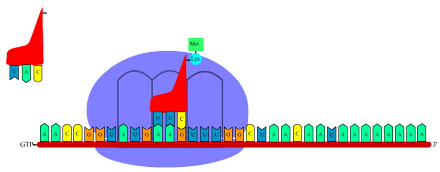
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

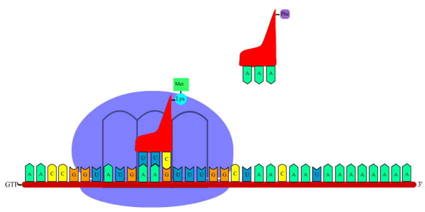
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

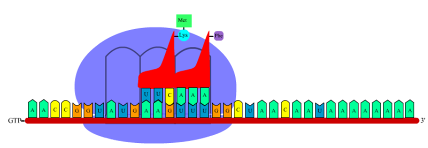
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

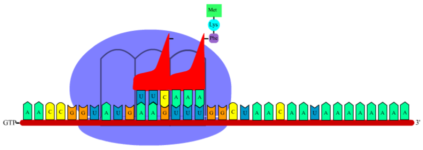
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

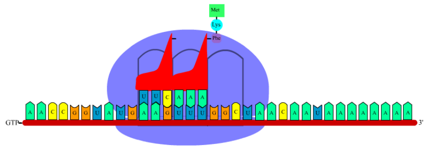
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

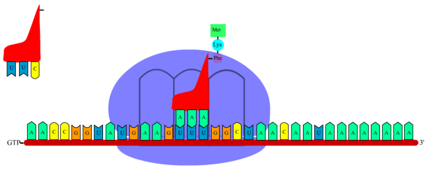
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

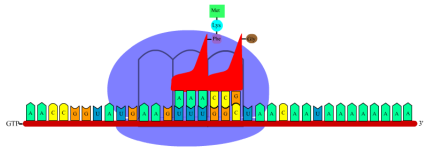
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

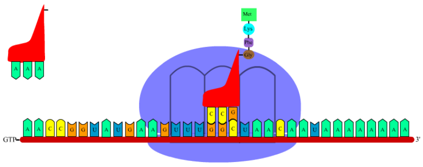
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

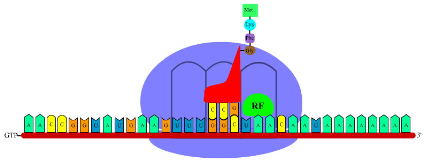
La traduction en images



- Le ribosome va par l'intermédiaire d'un ARN de transfert (ARNt) ajouter un acide aminé à la protéine en cours de fabrication selon le codon lu

Rappels de biologie

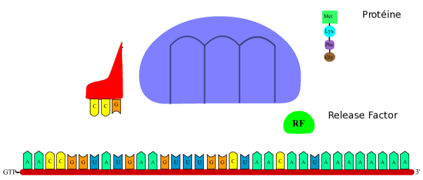
La traduction en images



- Une fois le codon-stop atteint, la protéine est complète

Rappels de biologie

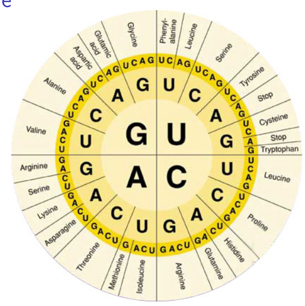
La traduction en images



- Le ribosome se détache de la protéine et du brin d'ARNm, et la protéine est libérée dans l'organisme

Rappels de biologie

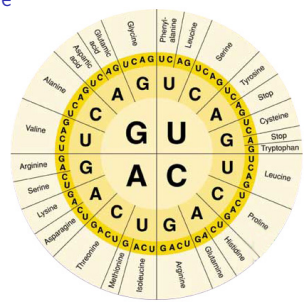
Le code génétique



- Les acides aminés sont produits en fonction du code génétique

Rappels de biologie

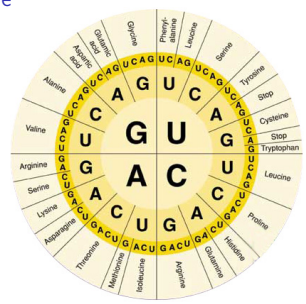
Le code génétique



- Il existe $4^3 = 64$ codons différents

Rappels de biologie

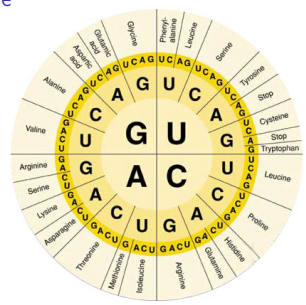
Le code génétique



- Parmi eux, trois sont des codons stop

Rappels de biologie

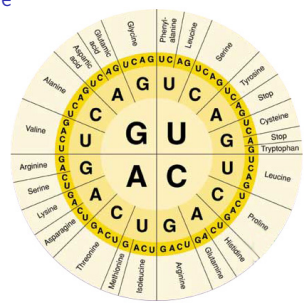
Le code génétique



- Les 61 restants ne correspondent qu'à 20 acides aminés

Rappels de biologie

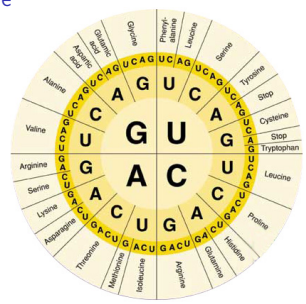
Le code génétique



- Le code génétique est donc redondant car en moyenne, un acide aminé est codé par trois codons

Rappels de biologie

Le code génétique



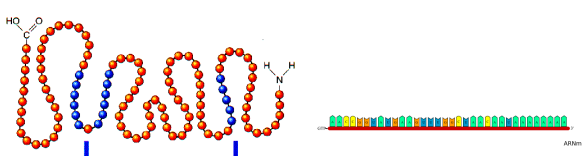
- Ainsi, en moyenne une mutation génétique sur trois affectant une séquence d'ADN codante n'entraîne aucune modification de la protéine traduite

Plan

Types de bio-informations

Les séquences

- Il existe principalement deux types de bio-information :
  - les séquences de nucléotides
  - les séquences d'acides aminés



Types de bio-informations

Les séquences

- Ces séquences sont dans les deux cas, des enchaînements d'unités élémentaires
  - ADN : 4 bases - A, C, G, T
  - ARN : 4 bases - A, C, G, U
  - Protéines : 20 acides aminés - Ala (A), Cys (C), Asp (D), Glu (E), Phe (F), Gly (G), His (H), Ile (I), Lys (K), Leu (L), Met (M), Asn (N), Pro (P), Gln (Q), Arg (R), Ser (S), Thr (T), Val (V), Trp (W), Tyr (Y)
- Elles possèdent deux extrémités distinctes et sont naturellement orientées
- Les séquences constituent l'un des principaux types de bio-information qu'analyse la bio-informatique

Types de bio-informations

Les séquences en informatique

- En conséquence les chaînes de nucléotides et d'acides aminés peuvent être représentées par une succession ordonnée et orientée d'unités élémentaires identifiées
- En informatique, elles sont donc représentées naturellement par des séquences de lettres
- Exemple de séquence nucléotidique

aattccggca tagaaactca aatcaaagag gaagaaacac cgattctcct tttctctctc
taaacaacta gatcagatct ctgagtttaa ggaagctttc agcctattcg ataaggatgg
cgatggttgc atcacaacca aggagcttgg aactgttatg cgatcattgg gacaaaaccc

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Types de bio-informations

Les séquences en informatique

- En conséquence les chaînes de nucléotides et d'acides aminés peuvent être représentées par une succession ordonnée et orientée d'unités élémentaires identifiées
- En informatique, elles sont donc représentées naturellement par des séquences de lettres
- Exemple de séquence polypeptidique

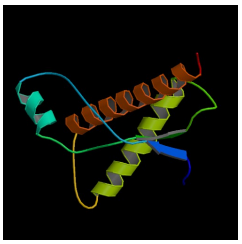
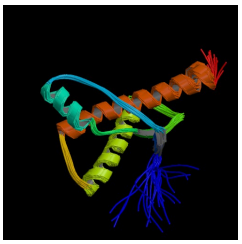
MADQLTDDQI SEFKEAFSLF DKDGDGCITT KELGTVMRSL
GQNPTEAELQ DMINEVDADG NGTIDFPEFL NLMARKMKDT
DSEEELKEAF RVFDKDQNGF ISAAELRHVM TNLGEKLTDE
EVDEMIREAD VDGDGQINYE EFVKVMMAK

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Types de bio-informations

Les autres types de bio-information

- Il en existe d'autres :
  - Les structures tridimensionnelles des protéines et des acides nucléiques

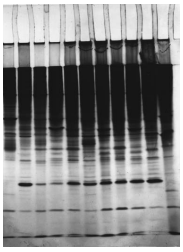



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Types de bio-informations

Les autres types de bio-information

- Il en existe d'autres :
  - Les données obtenues à partir d'électrophorèse sur gel (la séparation et la caractérisation des molécules - e.g. le séquençage de l'ADN)



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Types de bio-informations

Les autres types de bio-information

- Il en existe d'autres :
  - Les "Single Nucleotide Polymorphism" :
    - des variations d'une seule paire de bases du génome entre individus d'une même espèce
    - Ces variations sont très fréquentes (e.g. 1/1000 paire de bases dans le génome humain)
    - Les SNP peuvent se retrouver au sein de régions codantes de gènes (exon), de régions non-codantes de gènes (intron), ou de régions intergéniques, entre les gènes

GAGTCTGCCTAATAGTCCAAATCAT[C/T]ACAGGTCTTTTCTTAGCCATACACT

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Types de bio-informations

Les autres types de bio-information

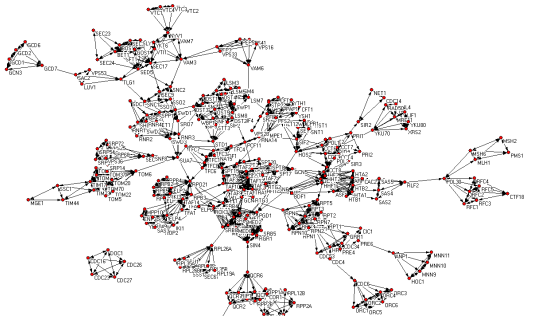
- Il en existe d'autres :
  - La taxonomie (classification) des organismes

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Types de bio-informations

Les autres types de bio-information

- Il en existe d'autres :
  - Les réseaux d'interactions qu'établissent les molécules biologiques (réseau d'interactions des protéines d'une levure)



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Types de bio-informations

Les autres types de bio-information

- Il en existe d'autres :
  - Les données bibliographiques (diffusion des résultats de la recherche par les articles)

1: Van Dijck J.
Cloning humans, cloning literature: genetics and the i
New Genet Soc. 1999;18(1):9-22.
PMID: 17256208 [PubMed - in process]

2: Suk J, Bruce A, Gertz R, Warkup C, Whitelaw CB, Braun A, Oram C
Dolly for dinner? Assessing commercial and regulato
Nat Biotechnol. 2007 Jan;25(1):47-53. No abstract available.
PMID: 17211395 [PubMed - in process]

3: Yu P.
Molecular chemical structure of barley proteins reveal
FTIR microspectroscopy: Comparison of barley variet

Plan

Le séquençage

Obtention de la bio-information

- Un exemple : le séquençage de l'ADN
- Le séquençage consiste à déterminer la succession des nucléotides formant un fragment d'ADN donné
- Actuellement, la plupart des séquençages d'ADN sont réalisés par la méthode de Sanger
- Cette technique, développée par Frederick Sanger vers 1977, est une méthode par synthèse enzymatique sélective

Le séquençage

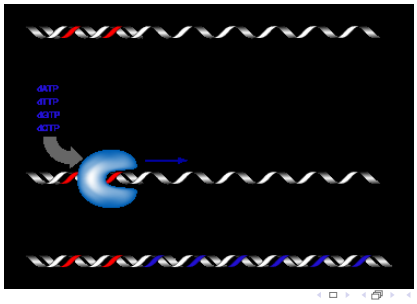
La méthode Sanger

- Initialement, la méthode de Sanger nécessitait de disposer d'un ADN simple brin
- C'est pour cette raison que le premier organisme biologique dont le génome a été séquencé en 1977 est le virus bactériophage $\varphi X174$ (dont le génome est constitué d'ADN simple brin)
- Le principe de cette méthode consiste à synthétiser toutes les copies partielles intermédiaires possibles de la molécule d'ADN

Le séquençage

La méthode Sanger

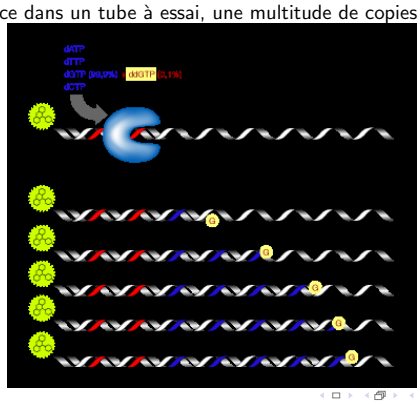
- Tout comme lors de la réplication, l'ADN polymérase utilise les nucléotides libres dans les parages pour synthétiser le complémentaire du brin modèle



Le séquençage

La méthode Sanger

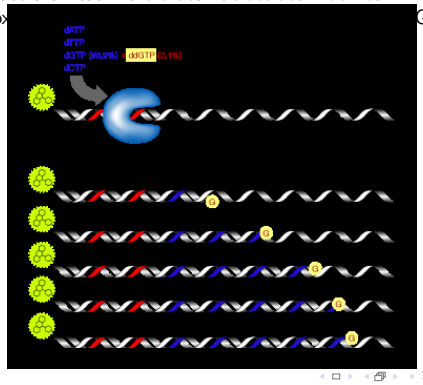
- On place dans un tube à essai, une multitude de copies du brin d'ADN



Le séquençage

La méthode Sanger

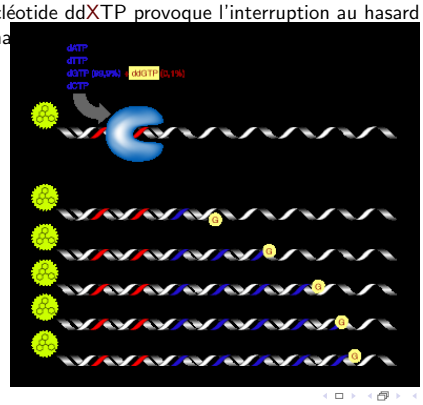
- On ajoute à un taux faible des nucléotides modifiés (didéox GTP



Le séquençage

La méthode Sanger

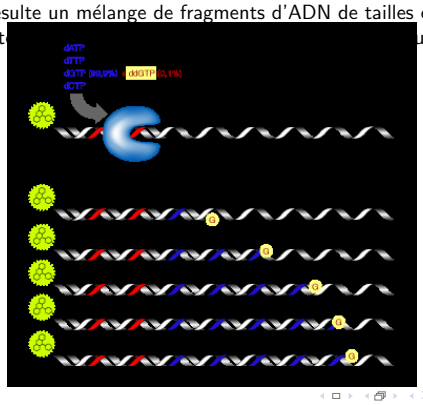
- Le nucléotide ddXTP provoque l'interruption au hasard mais systéma



Le séquençage

La méthode Sanger

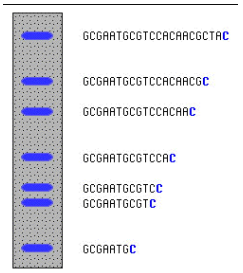
- Il en résulte un mélange de fragments d'ADN de tailles croissantes, qui se t uence



Le séquençage

La méthode Sanger

- On sépare alors les copies selon leur taille par une migration électrophorétique dans un gel poreux
- Ces gels permettent de séparer deux intermédiaires consécutifs qui ont une différence de taille d'un seul nucléotide

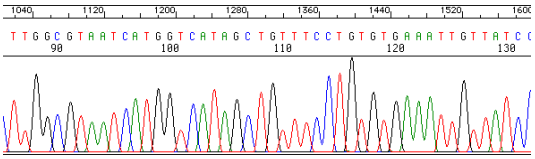


En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Le séquençage

La méthode Sanger

- Pour le séquençage complet d'un même fragment d'ADN, on répète cette réaction quatre fois en parallèle, avec les quatre didésoxyribonucléotides différents

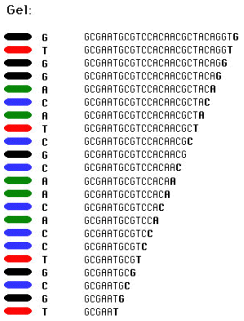


En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Le séquençage

La méthode Sanger

- Pour le séquençage complet d'un même fragment d'ADN, on répète cette réaction quatre fois en parallèle, avec les quatre didésoxyribonucléotides différents
- Il devient possible de reconstituer la succession des nucléotides tout au long de la séquence



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Le séquençage

La méthode Sanger

- Le séquençage est une technique automatisée - e.g. le séquenceur "MegaBACE®" (société Amersham)
- A l'heure actuelle seulement une séquence de 500 à 1000 bases à l'heure peut être séquencée

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Le séquençage

La méthode Sanger

- Il est donc impossible de lire "d'une traite" la séquence des immenses molécules d'ADN. Il faut donc fragmenter toute grande séquence et produire un volume de sous-séquences plusieurs fois supérieur à la taille de cette dernière
- Ces lectures redondantes permettent de raccorder les séquences les unes aux autres et de s'assurer de la qualité du résultat de chaque lecture. Reste alors à les assembler !

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Plan

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Le stockage

Les banques de données

- La bio-information est stockée sous forme de fichiers
- Ces fichiers sont regroupés dans des banques de données
- Ces bases de données sont réparties en deux catégories :
  - les BD généralistes - correspondant à une collecte de données la plus exhaustive possible et offrant un ensemble plutôt hétérogène d'informations
  - les BD spécialisées - correspondant à des données plus homogènes établies autour d'une thématique et offrant une valeur ajoutée

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Les banques de données

Les BD généralistes

- EMBL :
  - Banque européenne créée en 1980 et financée par l'European Molecular Biology Organisation
- Genbank :
  - Créée en 1982 par la société IntelliGenetics et diffusée maintenant par le National Center for Biotechnology Information
- DDBJ (DNA Data Bank of Japan) :
  - Créée en 1986 et diffusée par le National Institute of Genetics au Japon
- Swissprot :
  - Élaborée à l'Université de Genève à partir de 1986

Les banques de données

Les BD spécialisées

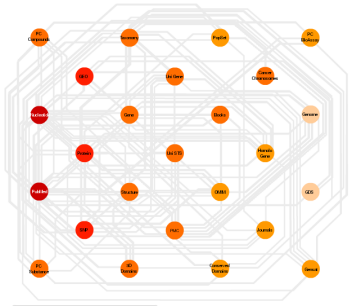
- SWISS-2DPAGE :
  - Cette base de données contient un grand nombre de données sur les gels d'électrophorèse bidimensionnel
- TRANSFAC :
  - Une base de motifs nucléiques recensant les séquences des différents motifs pour lesquels une activité biologique a été identifiée
- PROSITE :
  - Une base spécialisée de motifs protéiques pouvant être considérée comme un dictionnaire des motifs protéiques ayant une signification biologique

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Le stockage

Les banques de données

- Les multiples liens entre les groupes de données dans les banques généralistes sont d'une complexité étonnante
- Exemple de GenBank

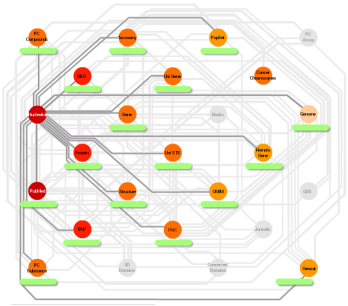


En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Le stockage

Les banques de données

- Les multiples liens entre les groupes de données dans les banques généralistes sont d'une complexité étonnante
- Exemple de GenBank



En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Les formats de stockage

Les banques de données

- Les séquences sont stockées sous forme de fichiers texte
- Le format correspond à l'ensemble des règles (contraintes) de présentation auxquelles sont soumises la ou les séquences dans un fichier donné
- Le format permet :
  - une mise en forme automatisée
  - le stockage homogène de l'information
  - le traitement informatique ultérieur de l'information
- e.g. format FASTA

¿HAHU Hemoglobin alpha chain - Human
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS
HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNF
KLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Algorithmes et de programmes

Les programmes

- La bio-informatique utilise des programmes écrits dans des langages qui peuvent lui être spécifiques
- La recherche de motifs dans les séquences est bien traitée par les algorithmes d'analyse de texte ("combinatorial pattern matching")

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Algorithmes et de programmes

Exemple de programmes

- Basic Local Alignment Search Tool - un algorithme qui permet de trouver les régions similaires dans un ensemble de séquences de nucléotides ou d'acides aminés
- Ce programme permet de calculer significativement les pourcentages de similitude entre les séquences en les comparant avec des banques de données
- BLAST est utilisé pour trouver des relations fonctionnelles ou évolutives entre les séquences et peut aider à identifier les membres d'une même famille de gènes

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Algorithmes et de programmes

Les algorithmes

- Mais la séquence seule ne suffit pas pour déterminer la fonction de certaines macromolécules comme les ARN
- Il faut tenir compte de leur structure tridimensionnelle
- Dans ce cas, l'analyse bio-informatique nécessite de nouvelles méthodes :
  - le développement de structures de données et d'algorithmes (arbres et tableaux de suffixes, automates...)
  - la construction automatique (inférence) de structures d'ARN à partir d'alignements de séquences d'ARN ou d'un ensemble de séquences de même type et issues de différents organismes

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Algorithmes et de programmes

Les algorithmes

- Il est difficile de définir en informatique les objets que manipulent les biologistes
- Depuis des années, les bio-informaticiens utilisent des concepts aussi divers que :
  - le recuit simulé
  - les chaînes de Markov
  - les statistiques bayésiennes
  - les réseaux de neurones

Plan

En quelques mots La bio-information L'informatique Du support à l'ordinateur Les apports de l'informatique

Des questions sur "Introduction à l'informatique génomique" ?

Informatique Génomique - Master 1

Guillaume Blin

IGM-LabInfo UMR 8049,
Bureau 4B066
Université de Marne La Vallée
gblin@univ-mlv.fr
http://igm.univ-mlv.fr/~gblin

2007-08

G. Blin - Informatique Génomique - Master 1 Cours n° 1 - Introduction