République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE ABBES LAGHROUR - KHENCHELA

FACULTE DES SCIENCES DE LA NATURE ET DE LA VIE

DEPARTEMENT DE BIOLOGIE

MEMOIRE

Présenté pour l'obtention du diplôme de

MASTER ACADEMIQUE

FILIERE : Sciences Biologies

OPTION :Biotechnologie et Amélioration des Plantes

Thème

La recherche d'une protéine (LEA) en utilisant une base de données spécialisée (TAIR) et une banque de données généraliste (PDB).

Présenté par :

ARAAR SAIDA

LATRECHE HABIBA

Soutenu le 01/06/2016

Jury de soutenance :

Président : Mr. ZeraibAzzeddine (M.C.B) Université Abbès Laghrour Khenchela.

Promoteur :Dr.Khabthan Abdelhamid(M.C.B)Université Abbès Laghrour Khenchela.

Examinateur : Mr.RahalKhaled (M.A.B) Université Abbès Laghrour Khenchela.

Promotion : Mai 2016

Remerciements

En préambule à ce mémoire nous remerciant ALLAH qui nous a aidé et nous a donné la patience et le courage durant ces longues années d'étude.

Nos remerciements s'adressent d'abord au Dr. Zeraib docteur à l'université Abbes Laghrour, Khenchela. Vous nous faites un grand honneur en présidant ce jury. On vous prie de bien vouloir recevoir le témoignage de notre profond respect.

Nous adressons nos sincères remerciements àMr. Rahale , enseignant à l'université Abbès Laghrour, khenchela. On vous remercie de nous avoir honorées par votre présence en tant qu'examinateur et pour avoir accepté d'évaluer ce mémoire.

Nous remercions très vivement notre encadreur, Dr.Khabeten Abdelhamid, Vous nous avez guidées dans la réalisation de ce travail malgré vos multiples occupations. Infatigable, toujours disponible, aimable, ces qualités vous ont valu l'estime de tous les étudiants et forcent l'admiration de tous. Apprendre à vos côtés a été un réel plaisir. J'ai gardé de bons souvenirs de vos enseignements avec clarté et précision, Que Dieu vous récompense

Je tiens à exprimer ma gratitude au Pr. Hmidachi Abdelhfid professeur à l'université de Khenchela pour, Je suis particulièrement honorée de bénéficier de ses critiques.

Nos remerciements vont aussi au corps professoral et administratif de la Faculté des Science de la nature et de la vie, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée.

Enfin, un immense merci à, nos camarades de promotion, nos amies et nos collègues de travail pour les encouragements et soutiens inaltérables, sans qui ce travail de thèse n'aurait pas été possible.

Latreche. H et Araâr. S

Dédicace

A mon cher père

Tes conseils m'ont suivi et m'ont permis d'atteindre le bout du chemin. Sois fier de moi aujourd'hui et vois à travers ce travail mon amour sincère et ma gratitude profonde.

A ma chère mère

Ma douce et tendre mère. Quoique je fasse, je ne pourrais te rendre ce que tu as fait pour moi. Si je suis arrivée là, c'est bien grâce à toi. Que dieu te donne longue vie et te protège.

A ma très chère grande mère

Tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement tu n'as jamais cessé de m'encourager et de prier pour moi.

A mes très chers frères

Bachir, Abderrahman, Abdelghani

Hichem, son épouse Nessrine

A mes très chères sœurs :

Adila, son mari Laarbi, et leurs enfants (Ilef, Sid ali).

Sabrina, son mari Amine, et leurs enfants (Assila, Amani).

Je vous souhaite un avenir plein de joie, de bonheur, de réussite et de sérénité.

A mes chères-amies

Saida, Nadhira, Rokia, Salema, Mareim, Khaoula, Djamila, Loubna, manel, Asema, HalimaVers lesquelles j'ai un grand respect.

A toute ma famille.

A toutes mes camarades de promotion.

Et plus particulièrement à fahimaKarkar, Tu as été pour moi durant ces années passées

LATRECHE HABIBA

Dédicace

Je remercie dieu tout puissant de m'avoir donné la patience et le courage afin d'achever toutes les années d'études

Avec joie et plaisir, fierté et respect. je dédie ce mémoire

- *A ma mère : pour son amour et son soutien chaleureux dont elle m'a entouré*
- *A mon père : pour son courage dont il m'a comblé, durant mes études*
- *A ma grand-mère*
- *A mes frères : Amor et Salim*
- *A mes sœurs : Nassima et leurs enfants (Hidaya *Ranya * Hibat arahman *Hanin)*Hourya et leurs enfants (Lina *Lamis*Mahdi *Akram) *Naima*
- *A tous qui porte le nom de la famille ARAAR et REZKALLAH*
- *A mes amis : Habiba *Amina *Nadhira * Amira *Nasira*

Saida ARAAR

Table des matières

Liste des tableaux

Liste des figures :

Liste des abréviations :

ARPA: Advanced researchprojectsagency.

BDD : Banques des données.

CLR : Common languageruntine.

CNRS : Centre national de la santé et de la recherche médical.

DDL: Data definition language.

DPM: Data protection manager.

EBI : European bioinformaticsInstitu.

EMBO : European moleculary biologie organisation.

EST : Expressedsurveybsequence.

GSS : Genome surveysequence.

IMA : Institu du monde de base.

INSERM : Institu national de la santé et de la recherche médical.

JIPIO : Japan international protein information database.

MEDLINE: Medical litérature analysis and retrieval système online.

MIPS: Martinsried institute for protein sequencer.

MGI: Mouse génome information.

OCLC: Online computer library center.

OPAC: Online public access catalog.

SIB: Swiss institute of bioinformatics.

SGBD : Système de géstion de base des données.

SQL : tructuredquery langage.

TFD : Des bases de facteurs de transcription.

URL: Uniform resource locator.

XML: Extensible markup language.

YAC: Yeastartificial chromosome.

Introduction

INTRODCTION

Les bases de données biologiques sont des bibliothèques répertoriant des informations sur les sciences de la vie collectées grâce à des expériences scientifiques, à la littérature publiée, aux technologies expérimentales à haut débit, et aux analyses informatiques.

Les bases de données spécifique sont des outils importants pour les scientifiques car elles leur permettent de comprendre et expliquer de nombreux phénomènes biologiques allant de la structure des biomolécules et leurs interactions à l'ensemble du métabolisme des organismes, et même l'évolution des espèces, et pour certaines espèces, en particulier celles qui sont souvent employées pour la recherche, il existe des bases de données spécialisées.(**Berge et ricroch ,2011)**

Notre travail a pour objectif de connaitre la différence de la représentation de la protéine LAE (Laite Embryogenesis Abundant Proteins) entre deux type de bases des données. la première spécialisée ,il s'agit de la base des données spécifique TAIR (The Arabidopsis Information Ressource), ou l'*arabidopsis thaliana* est l'espèce modèle, et la deuxième est celle de banques des données généraliste, il s'agit la banque de donnée protéique BDP. **(huala,2001)**

Chapitre I

Généralité sur les banques des données

I. Présentation générale

La bio-informatique est une discipline émergente de la recherche qui se place à l'interface de la biologie et de l'informatique. Il y a différentes façons de la définir. Il est possible de classer les bio-informaticiens qui la pratiquent en trois groupes. Les premiers se définissent comme pratiquant une branche fondamentale de la biologie capable de prédire, par des moyens informatiques, les lois ou les comportements biologiques. Par opposition aux classiques manipulations in vivo ou in vitro pratiquées en laboratoire, on parlera alors d'expériences « in silico » (néologisme d'allure semi-latine, formé à partir de l'anglais silicone). Les partisans de cette définition entendent exercer une bioinformatique théorique semblable à ce que les Anglo-Saxons nomment Computational Biology, c'est-à-dire la fabrication de modèles par le calcul à partir de données biologiques disponibles **(Claverie, 2000).** À l'opposé, un grand nombre de biologistes complètent leurs travaux en laboratoire par des analyses sur ordinateur. Ces bio-informaticiens-là ne créent pas de programmes, mais utilisent ceux qui sont écrits par d'autres, soit sur leurs ordinateurs personnels, soit sur des serveurs publics maintenus par des équipes pluridisciplinaires **(Dessen, 1995).** Ce domaine de l'analyse de données biologiques par ordinateur a de plus en plus tendance à se dénommer « bioanalyse ». Entre ces deux extrêmes, il existe des coopérations entre les biologistes et les informaticiens pour créer de nouveaux programmes informatiques destinés à la biologie. Ces projets interdisciplinaires constituent le creuset où se forgent les outils de la bio-informatique de demain **(Caudron, 2016**).

Divers instruments mathématiques sont largement et depuis longtemps utilisés en biologie, soit pour l'analyse statistique, soit pour la détermination de la structure des molécules. Ces domaines d'application ne diffèrent pas des usages des mathématiques et de l'informatique…..(EBI).**(Jongeneel, 2000)**

Pour des besoins spécifiques, de nombreuses bases de données spécialisées ont été créés, certaines sont pérennes et continuent d'être développées et mises à jour, d'autres sont laissées à l'abandon et enfin d'autres ont disparu. On en dénombre à cette date un peu plus d'un millier, accessibles directement par le Web. La nature ainsi que la quantité d'informations sont très variable**(http:// WWW.universalis.fr/encyclpedie/biologie -la-bio-informatique).**

La biologie et la bio-informatique

Au cours de ces trente dernières années, la récolte de données en biologie a connu un boom quantitatif grâce notamment au développement de nouveaux moyens techniques servant à comprendre l'ADN et d'autres composants d'organismes vivants. Pour analyser ces données, plus nombreuses et plus complexes aussi, les scientifiques se sont tournés vers les nouvelles technologies de l'information. L'immense capacité de stockage et d'analyse des données qu'offre l'informatique leur a permis de gagner en puissance pour leurs recherches. Et la rencontre entre la biologie et l'informatique, c'est ce qu'on appelle la bio-informatique. Celle-ci couvre des disciplines des sciences de la vie telles que la génomique, la protéomique et la biologie des systèmes **(Carlos et al. ,2010).**

Comment ça marche les banques des données?

"La bio-informatique fournit des bases de données centrales, accessibles mondialement, qui permettent aux scientifiques de présenter, rechercher et analyser de l'information. Elle propose des logiciels d'analyse de données pour les études de données et les comparaisons et fournit des outils pour la modélisation, la visualisation, l'exploration et l'interprétation des données", selon une définition de l'Institut Suisse de Bio-informatique.

Ça sert à quoi ?

La bio-informatique sert donc à stocker, traiter et analyser de grandes quantités de données de biologie. Le but est de mieux comprendre et mieux connaître les phénomènes et processus biologiques. Grâce à ces nouvelles connaissances ainsi acquises, les chercheurs ont la possibilité de faire de nouvelles découvertes scientifiques. Des découvertes qui peuvent par exemple améliorer la qualité de vie de personnes malades grâce à la mise en place de nouveaux traitements médicaux plus efficaces. **(Claverie, 2000).**

A quelles questions répond la bio-informatique ?

"La bio-informatique nous aide à visualiser les structures invisibles tels que les protéines et d'en apprendre davantage sur leur travail et leur fonction. Cela conduit à comprendre les questions essentielles de la vie : Comment les organismes fonctionnent-ils? Comment la vie s'est-elle développée ? Comment peuvent se développer de nouveaux traitements contre des maladies telles que le cancer **? http://www.adbs.fr/banque-de-donnees-16252.htm).**

I.1.La définition des banques des données :

Les banques des données c'est l'Ensemble de données relatif à un domaine défini des connaissances, généralement organisé et structuré en base de données pour être offert aux utilisateurs. On distingue les banques de données bibliographiques (références de documents primaires, avec ou sans résumé), les banques de données iconographiques (images fixes ou animées ; à ne pas confondre avec le mode image), les banques de données textuelles (texte intégral complet ou partiel de documents primaires) ou de type GED (document complet au format original), les banques de données numériques (données chiffrées, plus ou moins structurées) et multimédias (documents comprenant des textes, des images et des sons) **(Atwood et al.,2012).**

Une base de données (en anglais : database) est un outil permettant de stocker et de retrouver l'intégralité de données brutes ou d'informations en rapport avec un thème ou une activité ; celles-ci peuvent être de natures différentes et plus ou moins reliées entre elles. Dans la très grande majorité des cas, ces informations sont très structurées, et la base est localisée dans un même lieu et sur un même support. Ce dernier est généralement informatisé**.(hubert.2007)**

La base de données est au centre des dispositifs informatiques de collecte, mise en forme, stockage, et utilisation d'informations. Le dispositif comporte un système de gestion de base de données (SGBD) : un logiciel moteur qui manipule la base de données et dirige l'accès à son contenu. De tels dispositifs — souvent appelés base de données — comportent également des logiciels applicatifs, et un ensemble de règles relatives à l'accès et l'utilisation des informations.(**Schoof, 2002)**

La manipulation de données est une des utilisations les plus courantes des ordinateurs. Les bases de données sont par exemple utilisées dans les secteurs de la finance, des assurances, des écoles, de l'épidémiologie, de l'administration publique (statistiques notamment) et des médias.

Lorsque plusieurs choses appelées bases de données sont constituées sous forme de collection, on parle alors d'une banque de données (en anglais : databank**).**

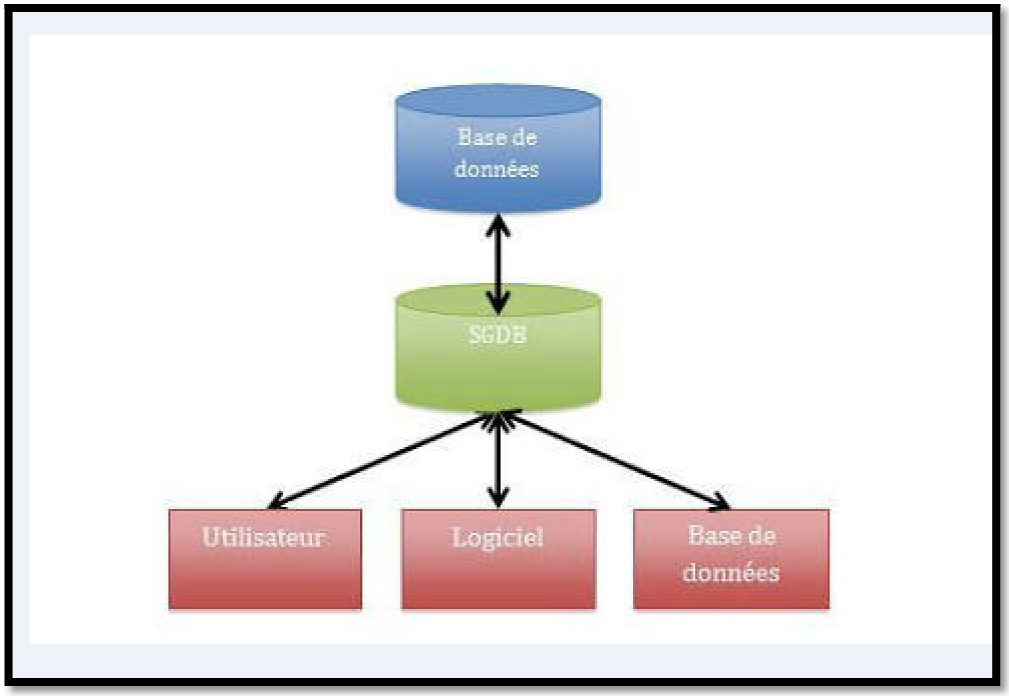


Figure 01 :SGBD un logiciel pour manipulation des données(**http://www.rcsb.org/pdb/)**

-I.2. L'évolution des banques des données dans la bio-informatique :

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 (820), 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.

1858 : Premier câble télégraphique transatlantique.

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 ducatabolisme 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".

1927 : Hermann Muller met au point l'induction artificielle de mutations par les rayons X.

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 (ElectronicNumericalIntegrator 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.

1951 : William Shotkley met au point le transistor.

- Le bureau de la statistique US reçoit le premier UNIVAC (UNIversalAutomatic Computer)

(1000 instructions/s) : 1er ordinateur commercialisé. Il utilise des bandes magnétiques en

Remplacement des cartes perforées. (UNIVAC Memories)

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".

1954 : Suicide d'Alan Turing : il croque une pomme remplie de cyanure, suite à une

inculpation pour "moeurs controversées".

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.

- Création de FORTRAN,premier langage procédural de haut niveau , par John Backus& al. D'IBM.

1959 : Annonces de l'IBM 1401 (tout transistor).

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

1961 : Marshall Nirenberg et J.HeinrichMatthaei déchiffrent le code génétique.

1962 : Atlas, Manchester University, premier ordinateur à mémoire virtuelle.

1964 : Annonce de l'IBM/360 : ordinateur de 3e génération.

CDC 6600 par Seymour Cray, premier supercomputer (9 MFLOPS : 9 millions d'opérations par seconde).

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.

- Théorie de l'horloge moléculaire (Zuckerkandl& Pauling).

- Atlas of Protein Sequences : première compilation de protéines (**Dayhoff, Georgetown,1965**).

- PDP8 (Programmed Data Processor) de DEC : 1er mini-ordinateur diffusé massivement (>50000 exemplaires).

1967: "Construction of Phylogenetic Trees"

- Début des circuits intégrés CMOS (voir aussi : Circuits intégrés logiques).

1968 : Annonce par Seymour Cray du CDC 7600 (40 MFLOPS : 40 millions d'opérations par seconde).

1969 : Premières interconnexions ARPANET (réseau).

1970 : Programme d'alignement global de séquences

- Ken Thompson & Dennis Ritchie développe UNIX aux Bell laboratories.

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)

- Annonce du Cray 1, crée par Seymour CRAY (cf. interview, 1996): 1er super-ordinateur à architecture vectorielle.

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). (**Graham et Van der ,1973**).

- 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.

1974 : Création d'un Comité sur l'ADN recombinant, présidé par Paul Berg (Université de Stanford, Californie), appelant la communauté scientifique à un moratoire sur les expériences de recombinaison génétique.

- Programme de prédiction de structures secondaires des protéines (**Chou &Fasman,1974**).

1975 : MITS Altair 8080 : 1er ordinateur personnel (commercialisé en kit).

- Conférence internationale d'Asilomar (Californie), organisée par Paul Berg et ses collègues sur le risque génétique.

- Mise au point de la technique "Southern blot"

1976 : Le Cray 1 atteint 138 MFLOPS (138 millions d'opérations par seconde).

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,1978**)

- Séquençage du premier génome à ADN, le bactériophage phiX174 (5386pb) **(Frederick Sanger,1978)**

- Annonce du VAX 11/780 : premier super-mini-ordinateur.

1979 : Début d'USENET, échanges d'émail et Newsgroups.

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.

Elle est aujourd'hui diffusée par l'EBI, Cambridge, GB)

1981 : IBM-PC (8088), 16-32kb : 1er IBM-PC (PC-DOS)

- **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, Los Alamos, US).

- Annonce d'Internet (TCP/IP).

1983 : Barbara McClintock découvre les éléments mobiles génétiques (transposons) chez les plantes.

- IBM-XT Disque dur (10 Mbytes = 10 Moctets).

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).

- Commercialisation du LISA et du premier Macintosh**1985** : ACNUC, un des premiers logiciels d'interrogation des banques, a été développé et est maintenu à Lyon.

- Programme Fasta (**Pearson- Lipman,1984**) : recherche rapide d'alignements locaux dans une banque.

- Publication du 1er article relatant l'utilisation de la PCR.

- L'idée de décrypter les trois milliards de bases du génome humain naît pour la 1ère fois à l'Imperial Cancer Research (ICR) de Londres.

- Annonce du Cray 2 à un GIPS.

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.

- Le DOE propose de créer des centres du génome pour s'atteler au séquençage du génome humain

- Clonage du gène responsable de la myopathie de Duchenne

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

- Mise au point d'un nouveau vecteur : le YAC, 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 (**Altschul et al**,**1990**) : recherche rapide d'alignements locaux dans une banque.

- Premier essai de thérapie génique.

- Création du 1er Généthon.

- Tim Bernes-Lee développe le prototype du WEB.

1991 : Programme Grail (**Mural et al**,**1991**) : localisation de gènes.

1992 : Fondation du Centre de recherche SANGER par le Welcome Trust et le British

Medical Research Concil (Cambridge, UK). C'est le centre le plus productif des instituts public de séquençage : il réalise la moitié de la "production" mondiale.

- Publication de la 2e carte génétique du génome humain, établie par le Généthon à partir de 814 fragments génomiques (marqueurs choisis : microsatellites - résolution : 4,4 cM).

1993 : Etzold et Argos créent SRS, logiciel d'interrogation multibanques accessible sur le web

1994 : Publication de la 4e carte génétique du génome humain, établie par le Généthon à partir de 2066 fragments génomiques (marqueurs choisis : microsatellites - résolution : 2,9 cM).

- Succédant au navigateursLynx et NCSA, Netscape Navigator est disponible.

1995: Séquençage de la 1ère bactérie, Haemophilusinfluenzae(1,83 Mb) (**Fleischmann,1995).**

Séquenceur à capillaire qui a conduit à augmenter les performances des laboratoires d'un facteur dix entre 1995 et la fin de 1997, et d'un nouveau facteur dix à la fin du siècle.

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

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

2000 : Séquençage du 1er génome de plante, *Arabidopsis Thaliana*

- ASCI White (RS/6000) : IBM construit le premier superordinateur qui dépasse les 10 TERAFLOPS (dix mille milliards d'opérations par seconde).

Années 2000 : Epigénétique : développement de technologies d'analyse des modifications de l'ADN et des histones.

Accès aux revues et journaux scientifiques : développement de l'open access". Montée en puissance de la biologie synthétique.

Détermination de structures de systèmes biologiques de plus en plus complexes (ribosomes, spliceosome, virus, ...) - cryo-microscopie électronique et autres techniques ("femtosecond pulses / X-ray free-electron laser")

2001 : Annonce du décryptage presque complet du génome humain. (Février) : les travaux de la compagnie américaine privée CeleraGenomics et du projet public international Génome Humain (HGP pour Human Research Project) sont sur les sites Internet des deux revues Science et Nature.

2007 – 2008 : Avènement des nouvelles technologies de séquençage à très haut débit, dites de "seconde génération".Prise de conscience du phénomène "big data" (pas seulement en biologie) qui devient peu à peu une discipline scientifique**.**

Février 2015 plus de 1.143.000.000.000 nucléotides :

Plus de 18.900 génomes eucaryotes et procaryotes séquencés et des milliers en projet (GenomesOnLine).

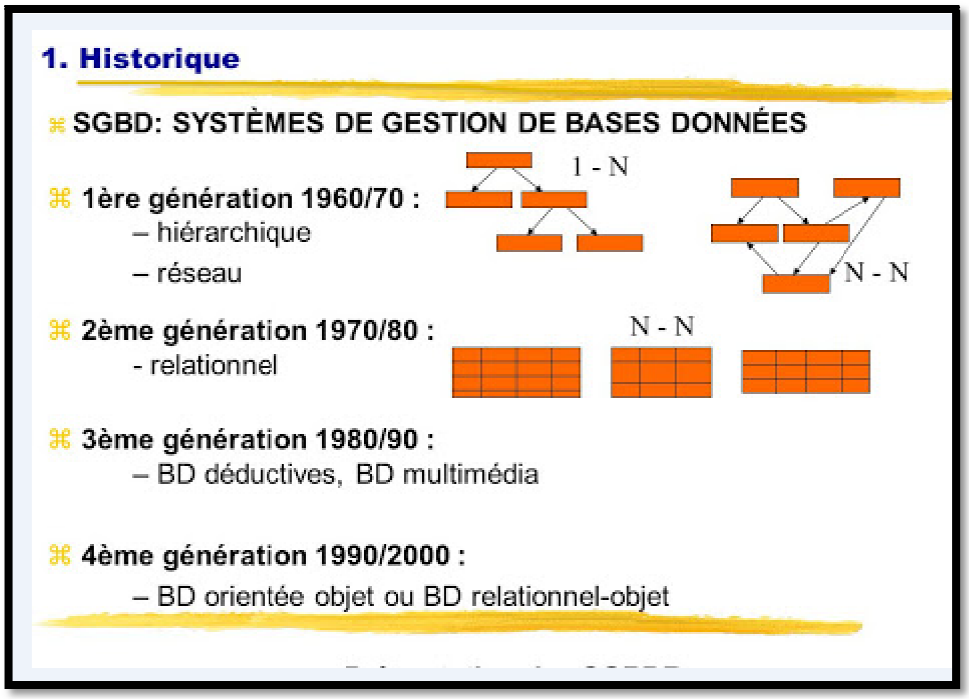


Figure 02 : L'évolution des banques des données **(www-igm.slidephayer.fr)**

II. La création des banques des données

Pour créer une base de données conforme à vos besoins métier, vous devez savoir comment concevoir, créer et gérer chacun des composants mentionnés ci-dessus. Vous pourrez ainsi assurer un fonctionnement optimal de votre base de données (**Krishmakumas et al., 2015**).

Tableau I : La fonction optimale des bases des données **(NCBI)**

Rubrique	**Description**
Bases des données	Explique comment utiliser des bases de données pour représenter des données, les gérer et y accéder. Ces tâches comprennent la conception, la mise en œuvre, et la maintenance des bases de données.
Serveurs de bases de données fédérés	Décrit les instructions et considérations de conception pour l'implémentation d'un niveau de base de données fédéré.
Tables	Explique comment utiliser des tables pour stocker des lignes de données et définir les relations entre plusieurs tables.
Index	Décrit comment utiliser des index pour augmenter la rapidité d'accès aux données dans la table.
Tables partitionnées	Décrit de quelle manière le partitionnement peut simplifier la gestion

et index	et l'évolutivité des tables et index volumineux.
Vues	Décrit l'utilisation des vues pour fournir un autre moyen de recherche des données dans une ou plusieurs tables.
Procédures stockées	Décrit la manière dont ces programmes Transact-SQL centralisent les règles d'entreprise, les tâches et les processus dans le serveur.
Déclencheurs DML	Décrit l'utilisation des déclencheurs DML en tant que types de pro cédures stockées spéciaux exécutés uniquement lors de la modification des données d'une table.
Déclencheurs DDL	Décrit l'utilisation des déclencheurs DDL en tant que déclencheurs spéciaux qui s'exécutent en réponse à des instructions DDL.
Déclencheurs de connexion	Décrit les déclencheurs de connexion qui sont activés en réponse à l'événement LOGON.
Notifications d'événement	Décrit les notifications d'événement en tant qu'objets de base de données spéciaux pouvant envoyer des informations relatives aux événements du serveur et de la base de données à un Service Broker.
Fonctionsdéfinies par l'utilisateur	Décrit l'utilisation des fonctions pour centraliser les tâches et processus dans le serveur.
Assemblys	Décrit l'utilisation d'assemblys dans SQL Server afin de déployer des fonctions, des procédures stockées, des déclencheurs, des agrégats définis par l'utilisateur et des types définis par l'utilisateur écrits dans l'un des langages de code managé hébergés par le CLR Microsoft .NET Framework, et non écrits en Transact-SQL.
Synonyms	Décrit l'utilisation d'un synonyme pour référencer un objet de base. Un synonyme est un autre nom désignant un objet contenu dans un schéma.

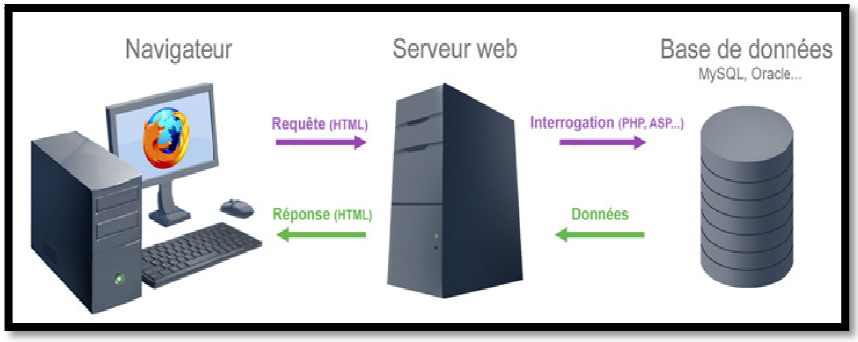


Figure 03 : La création des bases des données**.(www-igm.univ-m/v.fr)**

III. Les banques et les bases des données en biologie

Les fichiers contenant l'information biologique sous la forme de séquences est l'élément central autour duquel les banques de données se sont constituées à l'origine.

On peut distingue:

- **les bases de données généralistes** : elles correspondent à une collecte des données la plus exhaustive possible et qui offrent un ensemble plutôt hétérogène d'informations
- **les bases de données spécialisées** : elles correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée

Les techniques récentes de biologie moléculaire génèrent une quantité massive de données qui ne sont pas gérables par les techniques de publication traditionnelle (**Hesper et Hgeweg, 1970**).

En génomique, on distingue souvent de manière un peu arbitraire les banques de données généralistes, pour désigner les sites, qui gèrent et archivent les collections de données primaires c'est-à-dire expérimentales et globales non focalisées sur un champ d'application particulier sous forme d'un fichier texte structuré, et les bases de données spécialisées, pour nommer des ressources, qui gèrent des données davantage dédiées à un type d'organismes ou à une thématique donnée, le plus souvent à travers un logiciel dédié de type système de gestion de base de données (SG BD). Même si cette distinction fait encore partie du langage courant nous-mêmes distinguerons les « banques généralistes » et les« bases spécialisées », il est important de noter que de nombreux cas intermédiaires existent, et nous parlerons dans cet ouvrage de banques et bases de données en biologie pour désigner tout ensemble de données biologiques stockées, organisées, structurées et accessibles à un ensemble d'utilisateurs**.(Samson,2002)**

Cette fiche présente les principales banques et bases de données en génomique, sans prétention d'exhaustivité, en essayant de guider l'utilisateur clans ses recherches d'informations. Nous aborderons dans des fiches séparées les banques généralistes (appelées aussi banques primaires) et les principales bases spécialisées selon trois axes majeurs : les ressources dédiées aux génomes complets les bases dédiées aux expériences à grande échelle et pour finir les bases de motifs et d'éléments mobiles. Enfin, nous présenterons quelques outils permettant d'interroger et d'interfacer ces banques et bases de données, car ils ont pris une importance considérable en biologie. **(ricroch,2011)**

III.1. Banques généraliste

On appelle banques généralistes, ou banques primaires, les ressources qui collectent, gèrent, archivent et mettent à disposition de la communauté scientifique un ensemble de données primaires, c'est-à-dire obtenues expérimentalement.

Classiquement, on considère comme banques primaires les banques généralistes de séquences nucléiques et protéiques bien que la plupart des séquences protéiques ne soient pas obtenues expérimentalement, mais à partir des données des séquences nucléiquesainsi que les banques qui gèrent les structures tridimensionnelles des protéines **(colbert,2001)**

- Exemples de grandes bases de données généralistes

EMBL - EBI : Banque européenne créée en 1980 et financée par l'EMBO (European MolecularyBiology Organisation). Elle est aujourd'hui diffusée par l'EBI ("European Bioinformatics Institute", Cambridge - UK).

Genbank - NCBI : Créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI ("National Center for Biotechnology Information", Bethesda - Marylandet).

DDBJ ("DNA Data Bank of Japan") : Créée en 1986 et diffusée par le NIG ("National Institute of Genetics", Japon).

Swissprot& TrEMBL : Elle a été constituée à l'Université de Genève à partir de 1986. Elle est maintenant développée par le SIB (Swiss Institute of Bioinformatics) et l'EBI. Elle regroupe (entre autres) des séquences annotées de la PIR-NBRF ainsi que les séquences codantes traduites de l'EMBL (TrEMBL).

Ces banques s'échangent systématiquement leur contenu depuis 1987 et adoptent un système de conventions communes (The DDBJ/EMBL/GenBank Feature Table Definition).

PIR-NBRF ("Protein Information Ressource") : banque de protéines créée sous l'influence du NBRF ("Natio0nal Biomedical Research Foundation") à Washington. Elle diffuse maintenant des données issues du MIPS, de la base Japonnaise JIPID et des données propres de la NBRF.

UniProt ("Universal Protein Resource") : c'est la base de données des protéines : ExPASyProteomics Server. Consortium [EBI - SIB - PIR]

GOLD ("GenomesOnLine Database") : base de données qui recensse les milliers de génomes séquencés ou en voie de séquençage **(Colin Ritchie et Design,2008)**.

TableauII : Intégration des données **(http//www.rcsb/org/pdb)**

Banque de données	Nombre d'entrées	Taille de la base (Go)	Nombre d'objets bio	Durée d'import
PROSITE	1,5 K	0,8	108 K	6 min
SWISSPROT	100 K	2,9	2,4 M	5h30
SPTREMBL	660 K	13	8,4 M	20h33
EMBL	17 M	261	122 M	25j
PRODOM	305 K	3,1	2,5 M	3h50
PFAM	85 K	1,9	1,6 M	10h04
BLOCKS	12 K	0,6	690 K	1h40
ENZYME	4 K	0,2	42 K	5 min
RHDB	133 K	1,9	1,34 M	1h58

-Banques nucléiques

Il existe trois banques nucléiques internationales : GenBank, la banque américaine gérée par le National Center for Biotechnology Information (NCBI), l'European Molecular BiologyLaboratorydatabank (EMbL), la banque européenne maintenue à l'European Bioinformatics Institute (EBI), et enfin la DNA Database of Japan (DDBJ), la banque japonaise. Ces trois banques gèrent l'ensemble des séquences nucléiques et leurs annotations : elles coopèrent et échangent quotidiennement leurs données afin de garantir une cohérence maximale dans la mise à disposition des séquences de la communauté scientifique. Ainsi, même si chacune de ces banques présente quelques petites spécificités, la philosophie de structuration des données y est semblable et leur contenu en séquences nucléiques est

strictement identique. De plus, les entrées nucléiques sont organisées dans les trois banques en « division », selon deux types de critères :

- le groupe taxonomique d'origine de la séquence : bactéries, vertébrés, plantes, virus, etc.
- le type de molécule séquencée : EST, GSS, etc.

Chaque entrée, ou enregistrement, correspond à une séquence nucléique primaire disponible dans un format de fichier texte plat propriétaire : les données sont décrites dans un format texte où les lignes correspondent à des associations mot-clé/valeurs dans un format propre à chaque banque ; on parle de format Gen Bank, format EMBL, etc. Dans tous les cas, le format est très similaire et l'entrée est structurée en quatre parties. La première partie correspond à un en-tête contenant des informations générales sur la séquence : identifiant unique, numéro d'accession, définition, mot-clé, taxonomie de l'organisme dont la séquence provient. La deuxième partie décrit la ou les références bibliographiques associées à la séquence. La troisième partie, essentielle, décrit les annotations biologiques associées à la séquence sous forme standardisée : on parle de features qualifiers (les caractéristiques des annotations). Pour cette partie, la feature table est le document de référence qui définit le format d'annotation commun aux trois banques depuis 19901. Enfin, La quatrième partie contient la séquence nucléique elle-même**.(Targets,2011).**

Des logiciels dédiés permettent aux utilisateurs de soumettre les données en respectant les standards définis pour les annotations (Webin pour EMBL, Banklt ou Sequin pour GenBank).

Banques protéiques

Les entrées des banques protéiques sont structurées suivant des principes similaires et mettent à disposition des utilisateurs l'ensemble des protéines connues ainsi que les annotations biologiques associées.

Pendant de nombreuses années, trois banques protéiques ont coexisté de manière indépendante, avec leurs objectifs propres en termes de couverture — c'est-à-dire d'exhaustivité et d'annotations :

*la banque de données européenne Swiss-Prot, qui se caractérise par une excellente qualité d'annotation des données grâce à la contribution d'experts au détriment de l'exhaustivité ;

* La banque TrEMBL, qui contient l'ensemble des séquences protéiques conceptuelles obtenues par traduction automatique des séquences codantes contenues dans EMBL, avec des annotations automatiques non vérifiées, mais avec l'objectif d'obtenir une couverture maximale. De même, la banque Gen Pept correspond à la traduction automatique de l'ensemble des séquences annotées comme codantes (CDS) dans GenBank **(van,2009).**

* La banque américaine Protein Information Resource (PIR), à la National Biomedical Research Foundation (NBRF), qui dans les années 1960 fut historiquement la première banque de protéines développée. Sa particularité consiste à proposer une classification des séquences protéiques en familles, en fonction de leur degré de similarité. L'avantage est, d'une part, de limiter le degré de redondance de la banque et, d'autre part, de travailler à la standardisation de l'annotation des protéines.**(Baptiste, 2007).**

En 2002, à l'initiative d'un consortium international incluant l'EMBL-EBI, le SIB et la PIR, ces trois banques se sont regroupées pour donner naissance à l'Universel Protéine ressource (UniProt). UniProt propose ainsi, en profitant de leur complémentarité, un accès unifié à l'ensemble des informations contenues dans les trois banques primaires, notamment en ce qui concerne la qualité des annotations et la couverture de chacune des banques**.(Berardini et al,2003)**

Ces ressources primaires sont d'une importance cruciale en biologie. Elles posent cependant deux types de problème importants : d'une part, la qualité extrêmement variable des séquences et des annotations associées, sans qu'aucune procédure de traçabilité ni de contrôle permette d'évaluer cette qualité, et, d'autre part, le niveau de redondance important des donnés primaires, sans qu'actuellement aucune procédure permette d'identifier facilement l'origine de cette redondance (polymorphisme, erreurs de séquences, etc.). **.(Targets,2011).**

Enfin, dans le domaine des structures de protéines, la Protéine Data Bank (PDB) est une source d'informations qui fait référence ; elle archive et diffuse l'ensemble des données disponibles sur les structures cristallographiques des protéines. Notez que la PDB contient aussi quelques structures nucléotidiques, comme celles d'ARN de transfert

Les banques de données publiques sont très souvent le point de départ pour réaliser une analyse ou une caractérisation la plus exhaustive possible des séquences d'un organisme donné ou d'une famille protéique particulière. Il est cependant essentiel de garder à l'esprit que des vérifications expérimentales seront en général indispensables pour confirmer ou infirmer des résultats d'analyses obtenues in silice.(**Wisman et ohlogge,2000)**

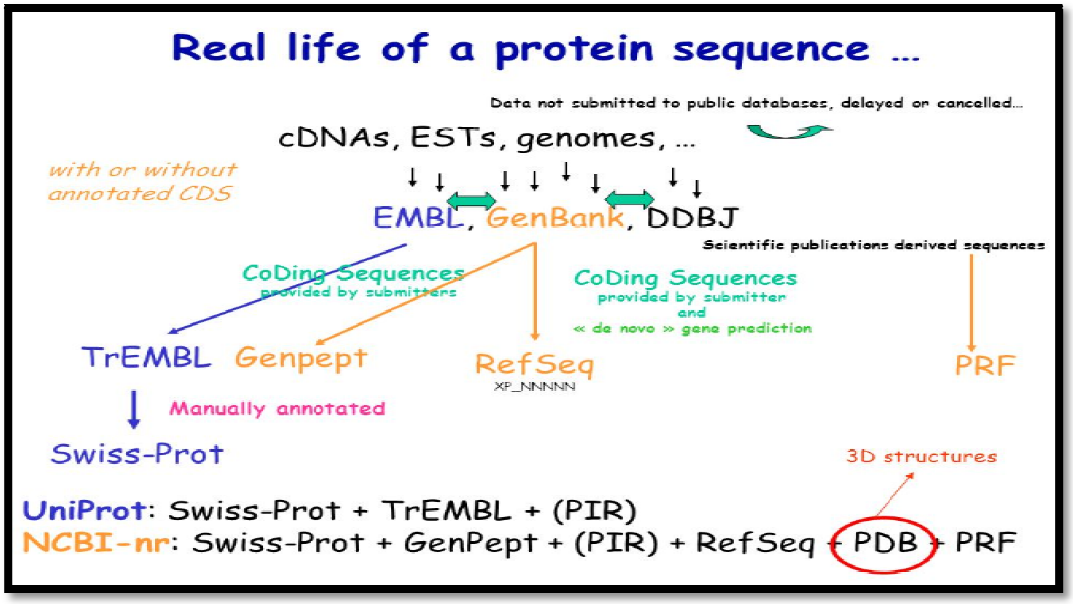


Figure 04 : exemple de banques protéiques. **(NCBI,national centre for biotechnology information)**

III.2. Les banques spécialisées

Pour des besoins spécifiques liés à l'activité d'un groupe de personnes, ou encore par compilations bibliographiques, de nombreuses bases de données spécifiques ont été créées au sein des laboratoires. Certaines sont inconnues ou mal connues et attendent qu'on les exploite davantage.

Les bases de données spécialisées sont d'intérêt divers et la masse des données qu'elles contiennent peut varier d'une base à une autre. Ces bases correspondent à des améliorations ou à des regroupements par rapport aux données issues des bases généralistes.**(Baptiste, 2007).**

Exemples de banques spécialisées :

Late Embryogenesis Abundant Proteins database (LEAPdb - **G. Hunault& E. Jaspard**) : cette base de données contient un grand nombre d'informations sur les proéines LEA impliqués dans la tolérance à de nombreux stress, notament la déshydratation et le froid. Pour l'instant, on les a mises en évidence principalement chez les plantes.

smallHeatShock Proteins database (sHSPdb - **G. Hunault& E. Jaspard**) : cette base de données contient un grand nombre d'informations structurales sur les cystéines de plus de 400

protéines cristallisées. Elle a aussi pour but de servir à la mise au point d'un logicel de prédiction des cystéines impliquées dans la formation de pont disulfure.

RESID Database : Base de données sur les acides aminés peu fréquents (sous-partie de la base de données PIR) **(Baptiste, 2007).**

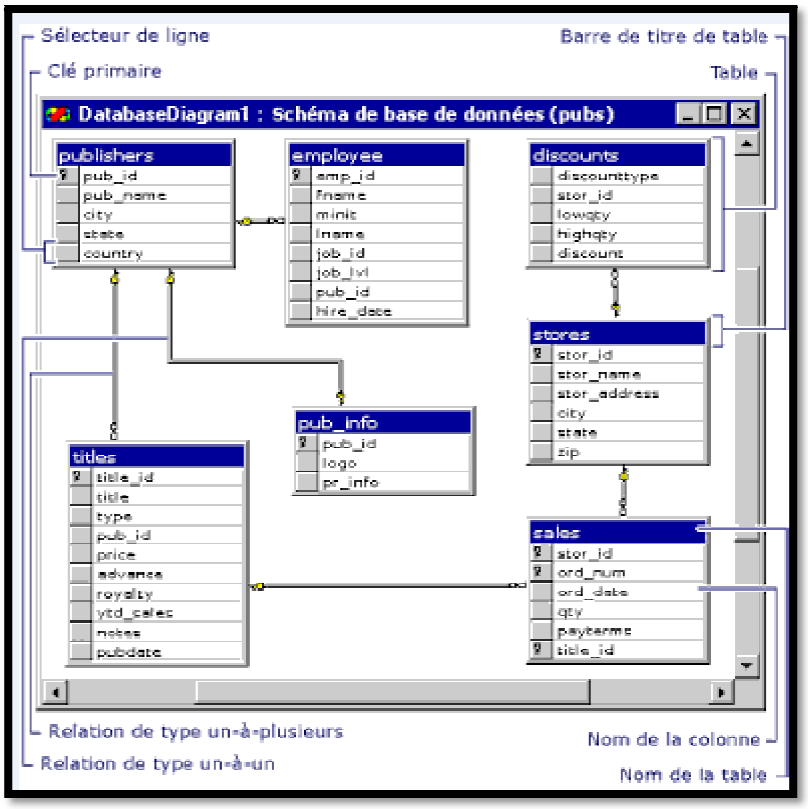


Figure 05 : La structure des bases des données**.(http//www.arabidopsis.org)**

***Les bases de motifs**

L'utilisation de bases spécialisées comme les bases de motifs est devenue un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéines inconnues ou savoir à quelle famille appartient une séquence non encore caractérisée.

a. Les bases de motifs nucléiques

La plupart de ces bases consiste à recenser dans des catalogues les séquences des différents motifs pour lesquels une activité biologique a été identifiée. Certains motifs sont simples et non ambigus, d'autres correspondent à des activités biologiques plus complexes et engendrent donc des séquences moins précises. Pour ces derniers types de motifs, des compilations ont été établies pour donner des listes annotées de motifs qui peuvent être communs à plusieurs séquences**.(Koornneef, 2010)**

Il existe principalement deux bases de motifs nucléiques qui sont régulièrement actualisées et qui correspondent à un travail de synthèse bibliographique : il s'agit des TFD **(Ghosh, 1993)** et TRANSFAC**(Knüppel et al, 1994).**

b. Les bases spécialisées de motifs protéiques

La base PROSITE (ExPASyProteomics Server) peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique.

Elle est établie en regroupant, quand cela est possible, les protéines contenues dans Swiss-Prot par famille comme par exemple les kinases ou les protéases. On recherche ensuite, au sein de ces groupes, des motifs consensus susceptibles de les caractériser spécifiquement. **(Caudron,2013)**

La conception de la base PROSITE repose sur quatre critères essentiels :

- collecter le plus possible de motifs significatifs
- avoir des motifs hautement spécifiques pour caractériser au mieux une famille de protéines
- donner une documentation complète sur chacun des motifs répertoriés
- faire une révision périodique des motifs pour s'assurer de leur validité par rapport aux dernières expérimentations.

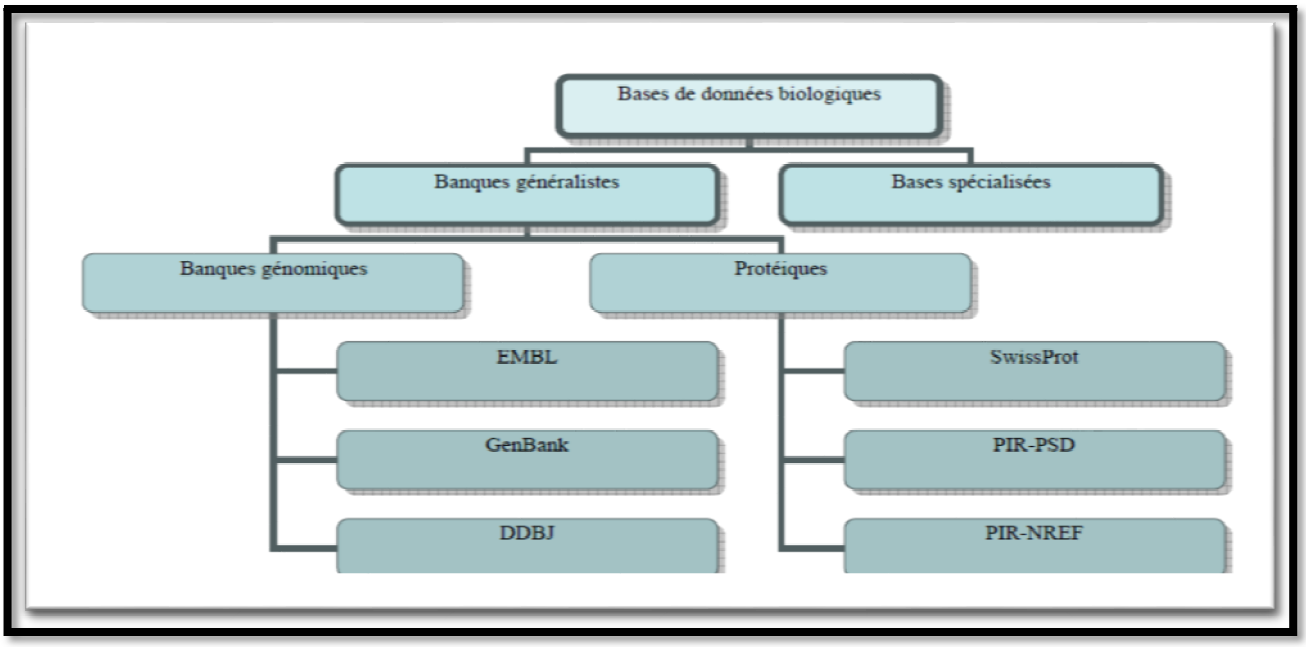


Figure 06 :les types des bases des données biologiques**.(NCBI)**

b-1- Base de données dédiées aux expériences à grandes échelle

Parallèlement au développement des banques généralistes. Un certain nombre de bases de données dédises aux génomes complets est développées. Ces ressources suivent deux évolutions majeures : d'une part une volonté d'intégration maximale de toutes les informations disponibles sur Les génomes séquencés — aussi bien au niveau des séquences nucléiques génomiques. Transcrites (ARN), traduites (protéines) que des annotations associées et, d'autre part. Une évolution marquée vers la génomique comparée et, dans certains cas. La phylogénomique est en effet une approche récente de phylogénie, qui a pour objectif de reconstruire l'histoire évolutive des gènes et des espèces à partir d'un large échantillon de données génomiques. Les enjeux sont importants et de la disponibilité de nombreux génomes complets proches d'un point de vue évolutif dans de nombreux groupes taxonomiques — tant chez les procaryotes que chez les eucaryotes — fait que les ressources dédiées à la phylogénomique sont actuellement en plein essor.**(Schaich,2011)**

b-2- Ressources généralistes

Une des ressources les plus anciennes dédiées aux génomes complets procaryotes et eucaryotes est la base Référence Séquence (RefSeq) du NCBI. À notre connaissance. C'est aussi à ce jour la seule ressource exhaustive. RefSeq a pour objectif de mettre à disposition de la communauté scientifique l'ensemble des séquences génomiques non redondantes, réannotés de manière homogène et sous des formats standards, Son principal Inconvénient est le manque de traçabilité des annotations automatiques et manuelles et la perte, parfois dommageable de l'annotation originale.**(Corey,1998)**

Depuis 2002, l'EBI mettait lui aussi à disposition une base de données contenant des génomes complets réannotés pour une large sélection d'organismes : la base Genome Revlews. L'ensemble des génomes complets des bactéries, des archées ainsi qu'un petit lot de génomes complets eucaryotes (la levure Saccharomyce, cerviceae et la plante modèle Arabidopsis Thaliana) était contenu dans cette base.

L'annotation originale y avait été enrichie grâce à l'ajout de nombreux types de features/qualifiers ainsi que l'intégration systématique d'un grand nombre de références croisées avec d'autres bases de données. En 2010, EnsemblGenomes a pris le relais de Genome Reviews afin de couvrir l'ensemble des génomes complets disponibles de non-vertébrés. Cinq instances différentes de cette base sont accessibles pour parcourir les

différentes branches du vivant, EnsemblBacteria, Ensembl Plants, EnsemblProtists, EnsemblFungi, et EnsemblMetazoa.(**Baptiste, 2007).**

b-3- Ressources pour les procaryotes

Pour les procaryotes, les deux bases de données de génomes complets les plus couramment utilisées sont la section"" MicrobialGenomces ""de la base RefSeq du NCBI et sa concurrente plus récente, la partie Procaryotes de la base Ensemble Genomes. Il est aussi à noter que plusieurs centres dc séquençage mettent à disposition des bases de données contenant les génomes qu'ils ont séquencés avant même qu'ils soient intégrés dans les banques publiques (parfois sous forme de brouillon, ou draft): voir notamment tes sites du ComputationalBiology and FunctionalGenomicsLaboratory' (Compbio) aux bats-Unis et du Sanger Institute en Grande-Bretagne.(**Hirayama et Shinozaki,2010)**

Enfin, d'autres ressources ont vu le jour plus récemment, en particulier des bases plus spécialisées sur l'annotation et la comparaison de génomes. Il s'agit par exemple d'ASAP, une base pour l'annotation collaborative et comparative des entérobactéries pathogènes, ou encore de la collection xBASE, un ensemble de bases dédiées à la comparaison de génomes bactériens proches, ou enfin de La base MOSAIC, qui met à disposition l'ensemble des régions conservées et des régions variables dans les espèces bactériennes pour lesquelles plusieurs souches sont séquencées.

b-4- Ressources pour Les animaux

Une des principales ressources de données pour les génomes des eucaryotes supérieurs est le projet Ensembl, issu d'une collaboration entre l'EBI et le Sanger Institute et dédié à l'annotation automatique des génomes de métazoaires. Ce projet fournit un environnement intégré de bases de données et d'interfaces graphiques pour annoter et comparer les grandes séquences chromosomiques à partir de l'ensemble des données disponibles.

D'autres bases constituent des références importantes pour les organismes modèles eucaryotes. Parmi les plus connues, nous pouvons citer trois exemples: la base FlyBase, pour l'annotation et l'analyse fonctionnelle des génomes de Drosophila, le MGI, qui fournit un environnement intégré pour l'annotation, la génomique fonctionnelle et la génomique comparée du génome de la souris, et la base de données de I'UCSC Genome Browser, qui permet l'analyse comparée des génomes de vertébrés**. (hanada et zhang, 2007)**

Parmi les autres bases d'intérêt sur les autres génomes eucaryotes modèles, citons deux derniers exemples : la base Worm Base, développée au Cold SpringHarborLaboratory pour intégrer les informations disponibles sur le nématode, et La base A CaernobabditiselegansDataBase (AceDB), développée en 1989 pour la gestion et l'annotation du génome du nématode modèle C. elegans et maintenant applicable à tout autre organisme procaryote ou eucaryote.(**http//www.rcsb.org)**

b-5- Ressources pour les plantes

Il n'existe pas de ressource unique dans le domaine végétal, et plusieurs bases sont développées en parallèle autour d'espèces d'intérêt. Les bases les mieux avancées à ce jour concernent ainsi le plus souvent les deux plantes modèles Arabidopsis Thaliana et Oryzasativa, pour lesquelles les données sont à la fois les plus anciennes et les plus complètes. Ainsi, la base relationnelle The Arabidopsis Information Resource (TAIR) centralise la plupart des informations disponibles sur Arabidopsis: données du programme de séquençage systématique, cartes génétiques et physiques, clones, marqueurs, etc. Parmi les autres ressources existantes, nous ne citerons ici que trois exemples importants: la base PLAGdb++, qui intègre les données génomiques de Arabidopsis, du riz, du peuplier et de la vigne, la base Gramene, référence internationale pour les céréales, et les bases MIPS plants databases (MIPS PlantsDB), qui incluent plusieurs bases dédiées à l'analyse fonctionnelle de génomes végétaux d'intérêt: par exemple, la base MIPS Arabidopsis Thaliana database (MAtDB), ou encore la base MIPS Oryzasativa database (MOsDB).**(samson,2002)**

b-6- Ressources pour les champignons

Pour les levures, qui abritent le premier génome eucaryote séquencé, Saccharomyces cerevisiae, une des bases de données qui fait référence est la Saccharormyces Genome Database (SGD), une base centrée sur la biologie moléculaire et la génétique de la levure de boulanger S. cercvisiae. Beaucoup d'autres ressources mettent à disposition l'ensemble des données disponibles pour cet organisme modèle ; nous n'en citerons ici que deux exemples: le consortium français Génolevures, qui inclut l'ensemble de ressources génomiques et protéomiques disponibles sur les génomes de levure séquencés, et la MIPS ComprehensiveYeast Genome Database (CYGD), une base de connaissances dédiée au génome de S. cerevisiae. Enfin, parmi les autres bases consacrées à d'autres organismes de levures, les bases CandidaDB et Candida Genome Database font référence pour le pathogène fongique Candida albicans.

Pour les autres génomes fongiques, les données génomiques commencent juste à s'accumuler massivement, et plusieurs ressources se sont développées récemment. Citons par exemple les bases de données e-Fungi et FUNYBASE pour l'analyse comparative des génomes fongiques complètement séquencés **(DenisTagu,2007).**

Chapitre II

La base de données TAIR

I/La description et la répartition d'*Arabidopsis thaliana*

Arabidopsis Thaliana est une espèce de plantes appartenant à la famille des Brassicacées. Elle est appelée Arabette des dames, Arabette de Thalius, Arabidopsis de Thalius, Arabette rameuse ou encore Fausse arabette (**Moller, 2011**).

*Découverte et origine du nom :

La plante a été décrite pour la première en 1577 dans les montagnes du Harz par Johannes Thal (1542-1583), un médecin de Nordhausen, Thüringen, Allemagne, qui a appelé Pilosella siliquosa. En 1753, Carl Linnaeus renommé la plante Arabis Thaliana en l'honneur de Thal. En 1842, le botaniste allemand Gustav Heynhold érigé le nouveau genre Arabidopsis et placé la plante de ce genre. Le nom du genre, Arabidopsis, vient du grec, signifiant «ressemblant Arabis» (le genre dans lequel Linné avait initialement placé) (**Koorneef et Meinke., 2010**).

I/01/La Description et les caractéristiques générales

- Organes reproducteurs :

  * Couleur dominante des fleurs : blanc.

  * Période de floraison : avril-août.

  *****Inflorescence : racème simple.

  * Sexualité : gynodioïque.

  * Ordre de maturation : homogame.

  * Mode de pollinisation : entomogame.

- Graine :

  *Fruit : silique.

  *Mode de dissémination : anémochore.

I/02 /L'*Arabidopsis Thaliana* comme un organisme modèle

Cette plante est un organisme modèle pour la recherche génétique dans le monde végétal. En 2000, ce fut le premier génome végétal séquence. Les raisons de ce choix sont nombreuses :

- petite taille ; en laboratoire, on peut cultiver un millier de pieds sur un mètre carré.
- cycle de développement court, le cycle graine → plante → graine ne dure que deux mois.
- un plant produit environ 40 000 graines.

C'est un des plus petits génomes connus dans le monde végétal. Sa taille a initialement été estimée à 125 millions de paires de bases, réparties sur cinq paires de chromosomes contenant 33 323 gènes, dont 27 206 codant pour des protéines ; mais une étude datée de 2003 montre que la quantité d'ADN a été sous-estimée et qu'elle serait en réalité de 0,16 pictogramme par noyau cellulaire, soit environ 157 millions de paires de bases**(Hays et al, 2007**).

- absence d'intérêts économiques sur cette espèce, ce qui facilite la diffusion des informations entre laboratoires.

II/ T A I R base des données statistique

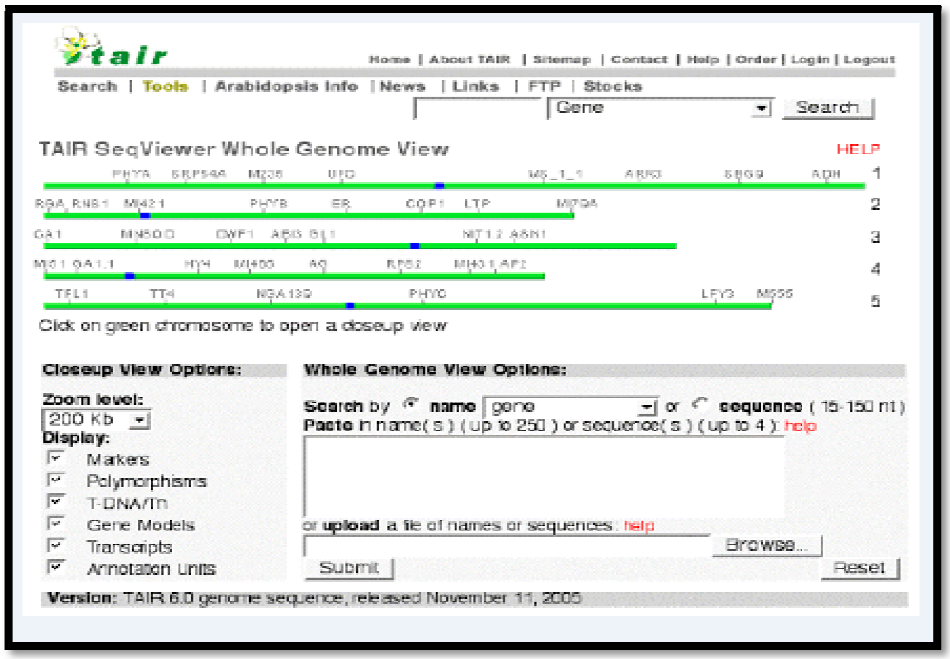


Figure 07 : La base de données TAIR.**(http// seqviewer.arabidopsis.org)**

The Arabidopsis Information Resource (TAIR) est une base de données bio-informatiques consacrée à l'organisme modèle *Arabidopsis Thaliana* (ou Arabette des dames) (**Schaich ,2011**).

La ressource d'information Arabidopsis (TAIR) est une base de données en ligne constamment mise à jour des données génétiques et de biologie moléculaire pour le modèle *Arabidopsis Thaliana* plante qui fournit une communauté mondiale de recherche avec un accès centralisé aux données pour plus de 30.000 gènes d'Arabidopsis. Lesbiocurators de TAIR extraire systématiquement, organiser et interconnexion des données expérimentales de la littérature ainsi que des prévisions de calcul, présentations communautaires et des ensembles de données à haut débit pour présenter une haute qualité et de l'image globale de la fonction du gène Arabidopsis. TAIR fournit des outils pour la visualisation et l'analyse des données, et permet la commande des semences et de l'ADN, les stocks de puces à protéines, et d'autres ressources expérimentales. TAIR engage activement avec ses utilisateurs qui apportent une expertise et des données qui augmentent le travail du personnel de la conservation. L'accent TAIR dans un vaste écosystème et de l'évolution des ressources en ligne pour la biologie végétale est sur le rôle très important d'extraire les résultats de la recherche sur la base expérimentale de la littérature et de rendre cette information accessible informatiquement. En réponse à la perte de la subvention du gouvernement, l'équipe TAIR a fondé une entité à but non lucratif, Phœnix Bioinformatics, dans le but de développer des modèles de financement durable pour les bases de données biologiques, en utilisant TAIR comme un cas de test. Phœnix a réussi la transition TAIR au financement par abonnement tout en gardant ses données relativement ouvert et accessible. **(Stanke et al, 2016).**

TAIR sert de la base de données communautaire pour les chercheurs d'Arabidopsis et comme une source d'information essentielle pour la biologie végétale et organisme modèle des communautés plus larges. TAIR contient des données génétiques et génomiques pour Arabidopsis Thaliana, une plante bien étudié qui sert comme une espèce de référence pour de nombreux aspects de la biologie végétale. Arabidopsis Thaliana a également servi comme un organisme de recherche hautement productive pour explorer de nombreux domaines de la biologie fondamentale, y compris la réparation de l'ADN, la photobiologie, la dégradation des protéines, l'horloge circadienne, méthylation de l'ADN **(Van Anken et al ,2009).**

L'utilisation de TAIR continue d'augmenter avec 45 000 visiteurs uniques par mois en 2010 sur la base des données d'utilisation recueillies à l'aide de Google Analytics et plus de

1,8 millions de visites dans l'année écoulée, soit une augmentation de 6% par rapport à l'année précédente. Visites proviennent de partout dans le monde avec la comptabilité en Asie pour 36%, les Amériques 31% et Europe 30%. Bien que l'enregistrement ne soit pas nécessaire à la visualisation des données à TAIR, les utilisateurs doivent enregistrer et connecter pour commander des semences et l'ADN des stocks de l'Arabidopsis Centre de ressources biologiques (CARL), saisir des commentaires sur les pages TAIR ou soumettre des données à TAIR via notre outil de soumission de données en ligne. Le nombre d'utilisateurs de TAIR enregistrés en Septembre 2011 a atteint 22 000, avec 9400 de ces documents ajoutés ou modifiés au cours des 5 dernières années, servant une estimation de l'ensemble le plus actif des utilisateurs (**Haas et al., 2010**).

II/01/Sources de données TAIR

Les sources principales pour des données dans TAIR incluent la curation manuelle de la littérature de recherches, canalisations informatiques pour annoter la structure et la fonction de gêne et tracer les objets ordonnancés sur le génome, importation des données de GenBank et des soumissions de la communauté de la recherche. La curation manuelle de littérature à TAIR est actuellement limitée aux articles de recherches d'Arabidopsis paraissant en ces journaux avec le facteur d'impact le plus élevé de journal dû au coût de grand temps de curation manuel. Approximativement 36% de la moyenne de chaque mois de 107 articles de recherches d'Arabidopsis contenant des données gène-connexes manuellement des plantes. Le processus de curation inclut l'annotation des gènes avec l'Ontology de gêne (fonction composant de processus et cellulaire) et limites d'Ontology des plantes (structure et étape développementale) avec des codes et des références appropriés d'évidence. En outre, des symboles de gêne, les allèles, les phénotypes et l'information de matériel génétique sont saisis de la littérature et une description de gène de texte libre récapitulant les dispositifs importants d'un gène se compose par des conservateurs. **(http://mips.gsf.de/proj/planet/araws/tAIGaSearch.html**

Des données informatiques sont produites par une série de canalisations automatisées. Les canalisations de structure de gêne mettent à jour des dispositifs de gêne tels que des exons et UTRs et ajoutent de nouveaux gènes basés sur la nouvelle évidence de transcription. Les canalisations fonctionnelles d'annotation assignent VONT des limites aux gènes basés sur la présence des domaines de protéine ou des ordres de signal et produisent d'une expression courte décrivant la fonction d'un gène. Traçant les canalisations assignent une position de

génome aux objets ordonnancés comprenant ESTs et cDNAs, T-DNA et insertions de transposon, marqueurs, SNPs, des canalisations d'importation de données etc. sont employées pour télécharger des données d'ordre de GenBank comprenant nouvel ESTs et cDNAs et ordres de flanquement de mutant d'insertion, et données de charge liées à la graine d'ABRC et aux stocks d'ADN. Les soumissions de données de la Communauté à TAIR incluent des familles de gènes, des données de fonction de gêne, de nouvelles gènes et mises à jour aux structures existantes de gêne, des phénotypes de mutant, des associés d'interaction, des modèles d'expression de gêne, SNPs, des marqueurs, des protocoles, des symboles de gêne, des données métaboliques de voie et des liens à d'autres ressources. Pour plusieurs de ces types de données, les formes Excel-basées de soumission de données sont disponibles sur le site Web. Des symboles de gêne sont manipulés par la soumission en ligne. La page de soumission de données **(Swarbreck et al, 2008**).

III. La génomique dans la base des données TAIR

III.1.Annotation de génome *Arabidopsis thaliana*

Bien que la séquence du génome d'Arabidopsis ait été achevée en 2000, il reste encore beaucoup à faire pour intégrer toutes les données expérimentales disponibles sur la structure et la fonction des gènes dans l'annotation du génome. Le génome d'Arabidopsis est le seul génome dicotylédone terminé et annotée à un niveau élevé, et (avec du riz) l'un des deux seulement fini, plutôt que de qualité brouillon, les séquences du génome de la plante. Parce que l'annotation de nouveaux génomes est dans une large mesure sur la base de l'annotation des génomes complets existants, l'amélioration de l'annotation du génome d'Arabidopsis sera directement aider l'annotation des futurs génomes de plantes**.(Coronel et Steven,2012)**

TAIR a assumé la responsabilité principale de la mise à jour de l'annotation Arabidopsis suivant l'Institute for Genomic Research (TIGR) Sortie du génome finale en 2004.

Depuis l'annotation originale du génome d'Arabidopsis dans lequel 25 498 gènes ont été rapportés, le nombre de gènes annotés a augmenté régulièrement en tant que nouveau séquençage et technologies sur les baies ont fourni des preuves pour de nombreux gènes préalablement annotées. En particulier, les nouvelles données de séquence déposée dans les bases de données de séquences nucléotidiques (EMBL / GenBank / DDBJ), ainsi que des soumissions des utilisateurs de la communauté de Arabidopsis, a donné lieu à l'ajout de plus

d'un millier de nouveaux gènes depuis la version finale TIGR en Janvier de 2004 **(Hanada et al, 2007).**

Un ensemble de données du génome de référence rigoureusement annotée est essentiel pour faire des inférences, produisant des modèles précis, et générer des hypothèses testables sur les fonctions des gènes dans Arabidopsis et d'autres génomes de plantes. L'une des utilisations les plus courantes de TAIR est d'extrapoler la fonction des gènes dans les espèces importantes en agriculture basé sur orthologie à des gènes d'Arabidopsis. Les rôles biologiques peuvent être inférés avec un plus grand degré de confiance si les éléments de preuve soutenant l'affirmation du gène de référence sont expérimentalement basés plutôt que le résultat d'une prédiction de calcul. Ainsi, l'activité principale des conservateurs TAIR intègre les données expérimentales de la littérature de recherche pour produire des annotations de haute qualité qui permettent la génomique fonctionnelle et comparative **(Lamesch et al. 2010**).

Les Conservateurs TAIR extraient et organisent une variété de données de la littérature examinée par des pairs (tableau II). Les données comprennent, mais ne sont pas limités à des informations de la fonction du gène capturé sous la forme de Gene Ontologie (GO) annotations.

L'information de l'expression du gène sous la forme de plante ontologie (PO) annotations symbole de gêne et les noms complets, allèles, phénotypes et informations de matériel génétique, et des publications. Ces données sont soigneusement organisées selon des normes rigoureuses et sont présentées de manière structurée qui reflète les relations biologiques réelles. Lorsque les membres de la communauté soumettent des données à TAIR, un conservateur examine la demande et effectue des vérifications de contrôle de qualité standard avant l'incorporation de ces données dans la ressource **(Cooper et al. 2013)**.

Tableau III. Littérature des Données Ajoutée à TAIR Depuis le 31 Août 2013 (données au 15 Juin, 2015).

Type de données	Nombre ajouté ou mis à jour
articles	7,428
les symboles des gènes	1,288
Les gènes liés aux articles	10,818
Articles liés aux gènes	4019
Articles utilisés pour GO expérimentalement pris en charge ou PO annotations	432 (conservateurs de TAIR à communauté) 215 (conservateurs de TAIR seulement)
Expérimentalement supporté GO et PO annotation).	2,870 (conservateurs de TAIR à communauté) 1.120 (conservateurs de TAIR seulement)
Les allèles de la littérature Phénotypes de la littérature	150 90

Toutes les soumissions communautaires sont examinées par un conservateur de TAIR avant l'incorporation dans la base de données. De temps en temps, des informations supplémentaires que les suppléments ou clarifie la soumission de l'utilisateur est ajouté au cours du processus d'examen.

Parce que la littérature Arabidopsis est vaste et les ressources de curation de TAIR sont limitées, nous avons établi un système de triage qui nous permet de concentrer nos littérature efforts de curation sur les articles avec des résultats d'impact roman ou élevé **(Liet al, 2012)**. Chaque mois, environ 350 nouveaux articles indexés par PubMed contiennent le mot «Arabidopsis» dans les rubriques titre, résumé. Parmi ceux-ci, environ 60% contiennent des informations sur un ou plusieurs gènes d'Arabidopsis. conservateurs TAIR utilisent une reconnaissance d'entités semi-automatique et une méthode de liaison pour associer des gènes

à des articles de recherche **(Yoo et al, 2006).** Conservateur lire les résumés de tous les nouveaux articles, revue liens générés informatiquement, et valider, invalider, ou manuellement ajouter de nouveaux liens entre les gènes et articles. En conséquence, la littérature corpus de croissance est précisément associée à des gènes individuels et devient accessible à partir des pages de détail du gène correspondant. Un sous-ensemble de ces articles, tels que ceux qui contiennent des résultats expérimentaux sur la fonction du gène pour les gènes non définies auparavant, est en profondeur et les résultats obtenues expérimentalement sont ajoutés à TAIR de manière structurée.(**School,2002)**

III.02.Gène Fonction

Information de la fonction des gènes est capturé sous la forme d'annotations GO qui décrivent la fonction moléculaire, rôle biologique, et la localisation subcellulaire des produits géniques. Les ontologies sont eux-mêmes vocabulaires structurés de manière à représenter avec précision la biologie et sont continuellement revus et mis à jour contrôlée. conservateurs TAIR jouent un rôle important dans le maintien des ontologies et de veiller à ce que les termes de biologie végétale et les relations sont incorporées correctement. TAIR est la principale source de manuelles annotations GO pour Arabidopsis. Depuis 2001, TAIR a créé des dizaines de milliers d'annotations de la littérature. Dans l'ensemble actuel des annotations valides, la contribution de TAIR comprend 91,445 manuels GO annotations à 18.932 produits du gène Arabidopsis distinctes représentant 80,6% de l'ensemble sur la base expérimentale disponible publiquement GO annotations pour A. thaliana **(http: //www.ebi .ac.uk / GOA / arabidopsis_release, GOA Arabidopsis (version 118), publié le 27 mai 2015)**.

Les annotations sont curation selon un ensemble rigoureusement défini des normes qui ont été largement discutés et documentés par le consortium GO **(Hill et al, 2008)**. En plus de l'annotation faite par les conservateurs TAIR, nous intégrons manuelles annotations A. thaliana de UniProt KB, le Consortium GO, et les contributions de notre communauté de recherche de présenter une vision unifiée de la fonction du gène Arabidopsis. TAIR intègre également des annotations générées informatiquement à partir d'UniProt qui reposent sur la cartographie de domaine Inter Pro ainsi que des inférences de calcul en interne sur la base de caractéristiques de séquence. Les annotations de calcul sont particulièrement utiles dans le cas où aucune autre information n'est disponible à partir de la littérature de recherche. Informatiquement prédites géniques annotations de fonction sont retirés de l'affichage que de nouvelles annotations validées expérimentalement sont ajoutés, ou lorsque plus exactes et à

jour des prévisions de calcul deviennent disponibles. Les chercheurs peuvent filtrer les annotations basées sur des preuves pour sélectionner uniquement les annotations qui sont soutenus par l'expérience. **(Burge et al, 2012)**

III.03.L'expression du gène

Les chercheurs qui souhaitent exploiter des données d'expression du gène Arabidopsis dans TAIR peuvent accéder aux données à partir de différents types d'expériences allant de grande échelle, génome études à l'échelle d'expression de plantes entières, des tissus et des cellules individuelles à l'expression in situ des gènes uniques en utilisant des sondes ou des gènes rapporteurs . Nous intégrons des données d'expression des gènes en utilisant des annotations PO. Les termes PO décrivent des structures végétales de la plante entière vers les cellules individuelles et la croissance des plantes et des stades de développement « plante entière et parties de plantes » **(Jaiswal et al, 2005).** Comme avec les annotations GO, les annotations PO facilitent croisées espèces expérimentales et transversales comparaisons **(Cooper et al, 2013).** Dans le cadre de notre flux de travail de la littérature curation, nous annoter les modèles d'expression de gènes rapportés dans des publications. Cela ajoute souvent une granularité importante à ce qui peut être connu à partir d'études utilisant des puces à ADN ou les données de l'ARN-seq. Par exemple, un gène représenté à exprimer dans les fleurs dans l'outil At Gen Express peut être démontré que l'expression spécifique d'un tissu même, lorsqu'on l'examine par hybridation in situ. TAIR capture et affiche à la fois de ces résultats pour présenter un profil d'expression plus détaillée. Littérature curation ajoute granularité et rend les ressources expérimentales plus visibles en incluant des informations sur le type d'expérience soutenant l'annotation. Si une fusion de gêne est utilisée pour localiser l'expression que l'information fera partie de l'annotation de sorte qu'un chercheur peut déterminer rapidement la personne à contacter pour obtenir la construction. (**Schmid et al, 2005**)

III.04.organisation et la structure du génome

alors et maintenant l'annotation du génome de référence pour le *A. thaliana* Col-0 séquence du génome d'origine a été périodiquement mis à jour pour intégrer de nouvelles données expérimentales sur l'expression génique, placer à jour les structures de gènes, et d'ajouter des variantes d'épissage et les gènes nouvellement découverts. De 2005 à 2010, TAIR a été responsable de la mise à jour des deux structures de gènes et la fonction des gènes pour la libération du génome d'Arabidopsis standard, disponible à partir de RefSeq de NCBI

et de nombreuses autres ressources. En utilisant une combinaison de méthodes de calcul et examen manuel, TAIR a produit un total de cinq génomes communiqués par le plus récent (TAIR10) rendu public en Novembre plus récent 2010 **(Krishnakumar et al, 2015**).

Site (02). En 2013, le projet d'Araport nouvellement financés a pris la responsabilité de fournir le génome d'Arabidopsis supplémentaires rejets le génome de Araport presse seront mis à la disposition de TAIR ainsi que des données sur les sites d'insertion marqués, les caractéristiques de protéines comme la localisation subcellulaire prédite et les domaines, et les données de orthologie provenant d'autres organismes. Comme nous l'avons fait avec les versions précédentes générées par TAIR, nous allons aussi produire un ensemble de jeux de données de séquences personnalisées basées sur la libération du génome d'Araport pour nos outils d'analyse de séquence **(Cornette et al ,2010).**

III.05.Nomenclature des gènes

Chaque locus dans Arabidopsis est assigné un identifiant unique, appelé code locus IAG (IAG, Arabidopsis Genome Initiative) qui est constitué du préfixe A, suivi par l'identificateur de chromosome (5/1 ou M ou C), suivie par g pour le gène, puis un nombre à 5 chiffres unique (par exemple At2g46340). En règle générale, les auteurs d'articles de recherche font référence à un gène en utilisant un symbole de gêne plutôt que l'identifiant de locus AGI. Cependant, les symboles de gènes ne sont pas des identificateurs uniques pour les gènes d'Arabidopsis; dans de nombreux cas, un symbole de gêne se réfère à deux ou plusieurs gènes ou le même gène a été donné deux ou plusieurs symboles de gêne par différents groupes de recherche. Cela crée des difficultés pour les chercheurs et les conservateurs qui veulent rechercher ou ajouter de nouvelles informations à un gène spécifique. conservateurs TAIR dépensent parfois des efforts considérables pour résoudre ces problèmes de nomenclature et de lien correctement gènes à des publications ou d'autres informations. Dans les cas où l'auteur ne comprend pas explicitement l'unique, AGI identifiant de locus dans un article, les conservateurs doivent souvent entreprendre un travail de détective pour déduire quel identifiant doit être associé au nouveau symbole du gène et de la publication **(Hiraymas et Shinozaki, 2010).**

TAIR maintient également un registre communautaire pour le symbole de gêne nomenclature qui a été mis en place pour réduire le chevauchement des symboles de gènes et donc le rendre plus facile d'identifier sans ambiguïté des gènes dans les articles. Les chercheurs peuvent utiliser ce registre pour assurer que le symbole qu'ils souhaitent utiliser

pour un nouveau gène n'a pas déjà été utilisé pour un gène différent, et réserver pour une publication ultérieure. Les chercheurs qui ont découvert un nouveau gène qui manque un identifiant de locus AGI dans la version la plus récente du génome devrait communiquer avec le projet Araport à contact d'avoir un affecté. L'identificateur nouvellement affecté devrait être inclus dans les résumés et les publications traitant le nouveau gène. **(https://www.araport.org/).**

Partie expérimentale

Matériel et méthodes

I- Matériel et méthodes

I.1. Matériel

Notre travail a pour objectif de rechercher la différence dans la fourniture de données entre les bases de données spécialisée et des banques de données généralistes et pour ce bute en a choisi la base de données spécialisée (TAIR). outil pour présenter la façon de fourniture de données et de mettre en évidence les caractéristiques sur lesquelles pour recueillir et organiser l'information, en se fondant sur le modèle choisi, la protéines (LEA).

Pour voir la façon de présentation des informations dans les banques de données généralistes, nous avons utilisé une banque de données généraliste (BDP) .est un modèle parfait pour la compréhension de la forme globale de l'information au sein de la banque de données généraliste.

I.2. Méthode

Toute les recherches scientifiques qui vise à étudier et déterminer la différence entre deux composants pour être pris en charge principalement sur une méthode de comparaison , une méthode qui permet de montrer des caractéristiques différentes de façon précise, et utilisant cette méthode , nous avons été en mesure d'identifier et de comprendre les points de différence et la compatibilité également entre chacun de la base de données spécialisée (TAIR) et de la banque de données généraliste (BDP)

I.3. Présentation générale de la protéine LEA

Découvertes chez les plantes, les protéines LEA «Late Embryogenesis Abundantproteins» recouvrent un grand nombre de protéines très variées et sont principalement associées à la tolérance au stress hydrique résultant de la dessiccation ou d'un choc thermique du au froid. De nombreuses fonctions ont été associées aux LEA, mais leur rôle précis n'est pas encore connu. Ces protéines assurant des rôles aussi importants que la protection des structures cellulaires, il semble judicieux de tenter de mieux les comprendre et de mieux les caractériser.

Des travaux récents chez l'***Arabidopsis thaliana*** utilisant des mutants KO du gène ATEM6 (un gêné codant une LEA) ont montré que les graines présentaient une déshydratation et une maturation précoce**)**. L'une des fonctions liée à cette caractéristique

structurale serait donc de réguler la perte en eau en intervenant pendant la maturation de l'embryon**(Manfree, Lanni 2006**.)

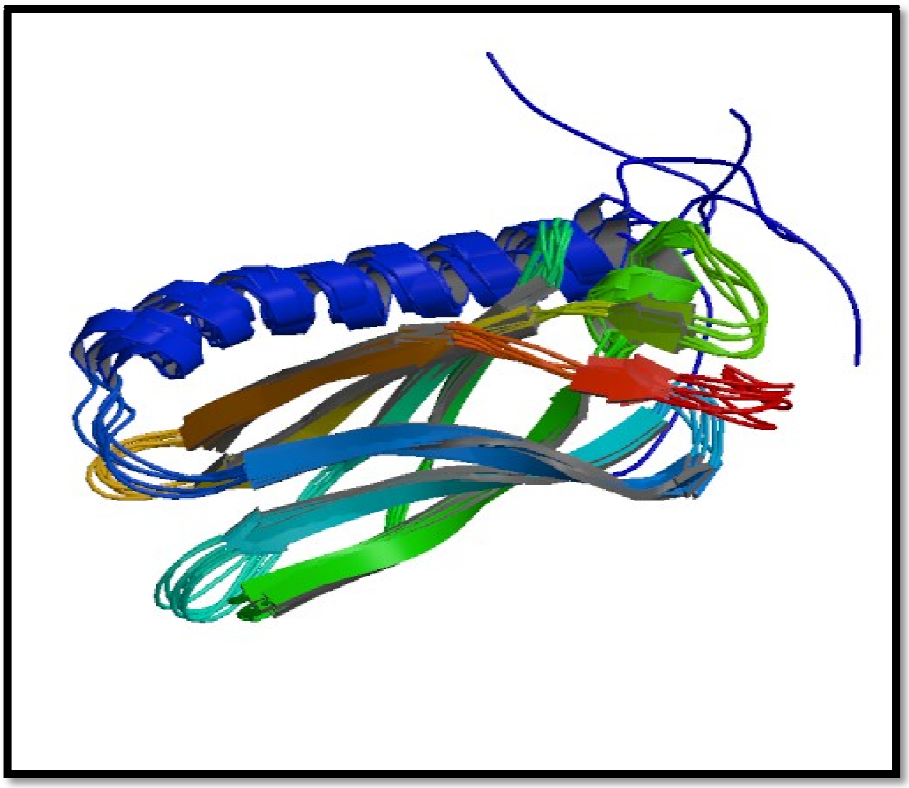


Figure 08 :La structure secondaire de la protéine LEA**.(www.uniprot.org/uniprot/082355)**

Résultats et discussion

Résultats

I. Les étapes de la recherche dans la base de données TAIR

On trouve dans la bio-informatique des bases de donnée spécifiques à des espèces végétal, telle que l'*Arabidopsis thaliana*, ce qui est montré par sa nomenclature (the Arabidopsis information ressource), cette plante est considérée comme une plante modèle de la physiologie végétale qui est notre spécialité universitaire, et dans cette recherche on suit les étapes suivantes :

- Etape 01 : La précision de la base de données TAIR.
- Etape 02 : L'utilisation du code At2g46140.1 spécifique à la protéine cible LEA.

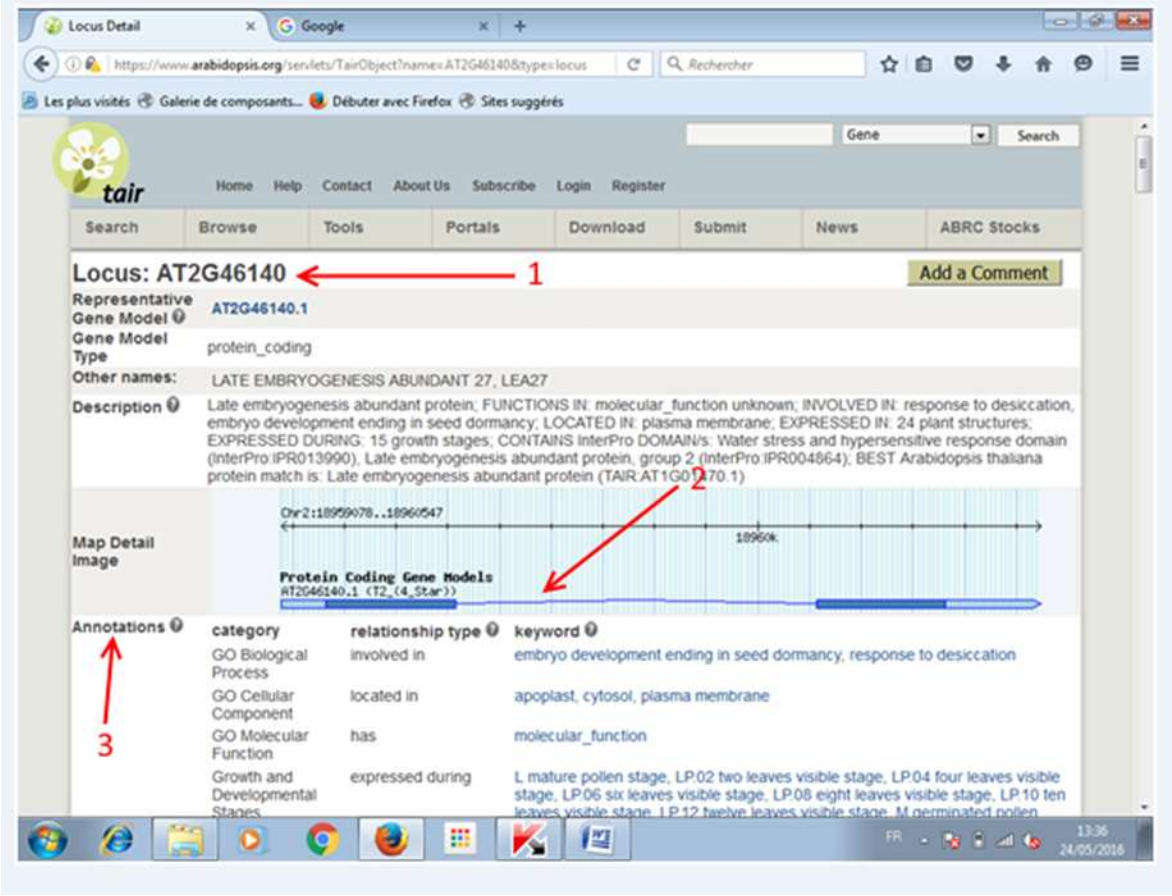


Figure 09 : Représentation du menu de la bas de données TAIR (**http//www.arabidopsis.org)**

«la base de données TAIR contient un menu varié -1: le code At2g46140.1 spécifique à la protéine LEA, 2: carte «mup» représentée le gène modèle qui codant la protéine LEA, 3: annotation de LEA».

- Etape 03 : L'étape précédente va nous donner une page WEB qui présente un menu varié.

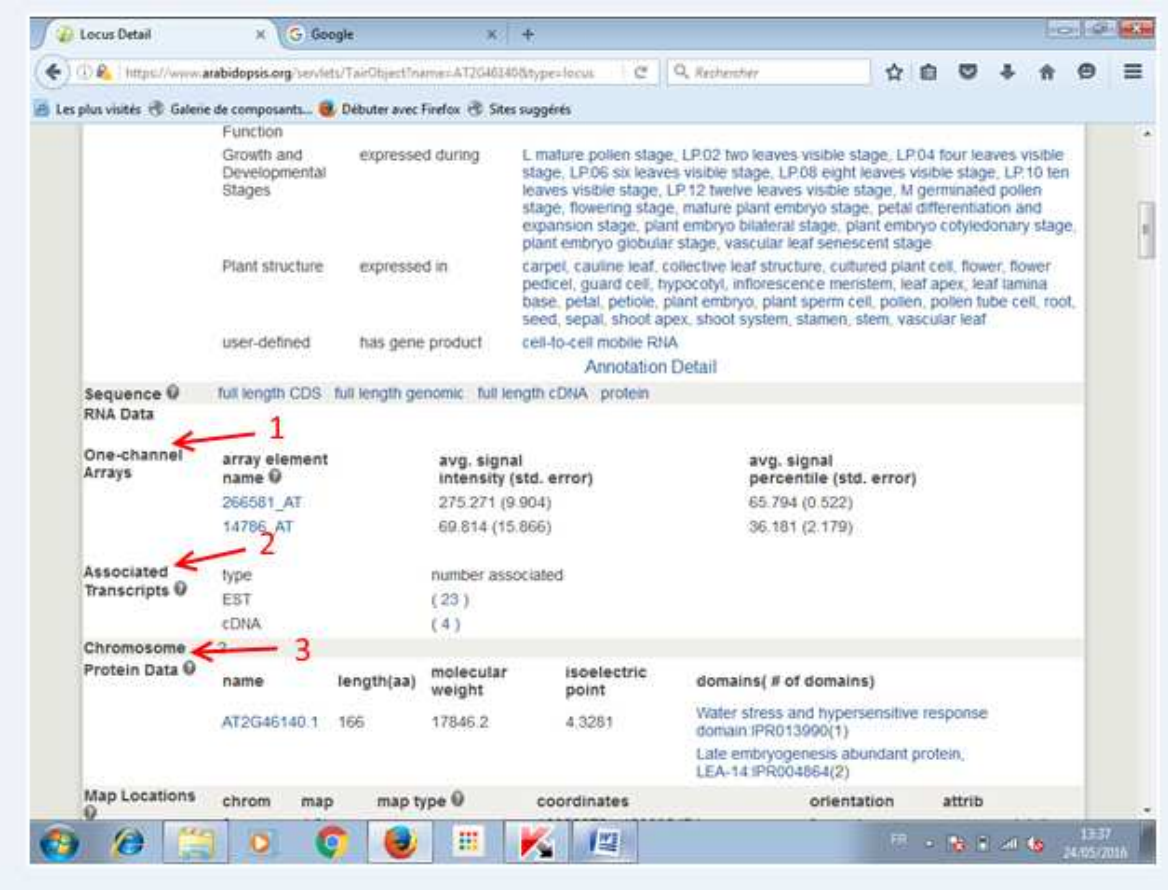


Figure 10 : La base de données TAIR représentant les détails d'annotation (**http//www.arabidopsis.org)**

(1 : nom d'élément de matrice, 2 : les nombre associé des types EST et ADNc, 3 : le nombre de chromosome).

- Etape 04 : A chaque fois on va ouvrir les pages du menu avec un enregistrement des pages par leur ordre présenté dans le menu.

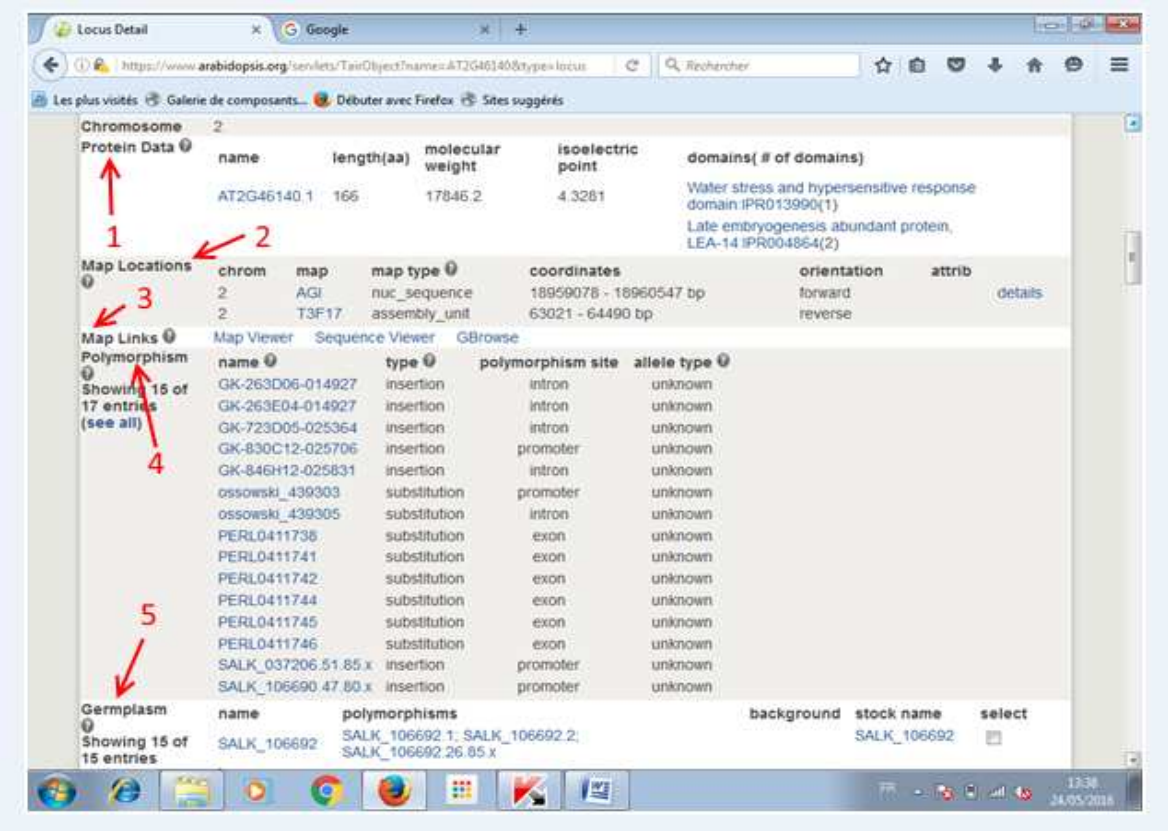


Figure 11 : Les caractéristiques de la protéine LEA (**http//www.arabidopsis.org)**

(1 : tableaux pour montrer les caractéristiques de LEA, 2 : endroit de carte représentée le gène modèle qui codant la protéine LEA, 3 : le lien de carte, 4 : le polymorphisme

II.1 La base de données TAIR représentée la protéine LEA

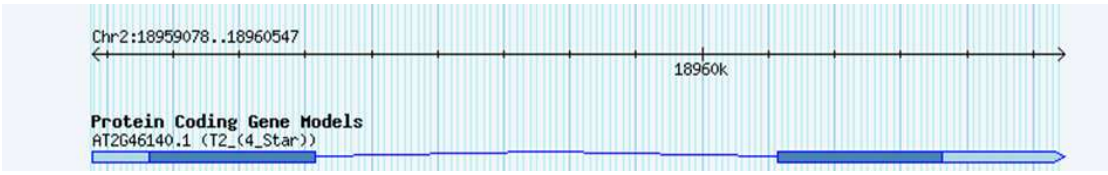


Figure 12 : Carte représenté le gène modèle qui codent la protéine LEA(**http//www.arabidopsis.org)**

*Description : Protéine abondante d'embryogenèse en retard.

* fonctions : inconnu de fonction moléculaire.

* impliqué dans : réponse à la dessiccation, fin de développement d'embryon dans la dormance de graine.

*situé dans : membrane de plasma.

*exprimé en : 24 structures de plante.

*exprimé pendant : 15 étapes de croissance.

*contient INTERPRO domain/s : effort de l'eau et domaine hypersensible de réponse (INTERPRO : ipr013990), protéine abondante d'embryogenèse en retard, groupe 2 (INTERPRO : ipr004864).

*la ma meilleure allumette de protéine de thaliana d'Arabidopsis est : Protéine abondante d'embryogenèse en retard (TAIR : AT1G01470.1

Détail d'annotation :

Séquence d'ARN : protéine intégral génomique d'ADNc.

Rangées d'Un-canal :

Nom d'élément de matrice :

266581_AT.

14786_AT.

avg. Intensité de signal (erreur de norme) :

275.271 (9.904).

69.814 (15.866).

avg. Percentile de signal (erreur de norme)

65.794 (0.522).

36.181 (2.179).

Chromosome : 2

Données de protéine :

Tableau IV : Les caractéristiques de la protéine LEA

Nom	longueur	poids moléculaire	point isoélectrique	Domaines
AT2G46140.1	166	17846.2	4.3281	Effort de l'eau et domaine hypersensible de réponse : IPR013990 (1) Protéine abondante d'embryogenèse en retard, LEA-14 : IPR004864 (2)

Endroits de carte :

Tableau V : Endroit de carte représenté le gène modèle qui codant la protéine LEA

Chromosome	Map	type de carte	Coordonnées	Orientation
2	Agl	carte nuc_sequene	18959078 - 18960547 bp	vers l'avant
2	T3F17	assembly_nit	63021- 64490 bp	inverse

Tableaux VI: Polymorphisme de la protéine LEA

Polymorphisme	Nom	Type	emplacement de polymorphisme	type d'allèle
Montrant 15 de 17 entrées (voir tous)	GK-263D06-014927	Insertion	Intron	inconnu
	GK-263E04-014927	Insertion	Intron	inconnu
	GK-723D05-025364	Insertion	Intron	inconnu
	GK-830C12-025706	Insertion	instigateur	inconnu
	GK-846H12-025831	Insertion	Intron	inconnu
	ossowski_439303	Substitution	instigateur	inconnu
	ossowski_439305	Substitution	intron	inconnu
	PERL0411738	Substitution	exon	inconnu
	PERL0411741	Substitution	exon	inconnu
	PERL0411742	Substitution	exon	inconnu
	PERL0411744	Substitution	exon	inconnu
	PERL0411745	Substitution	exon	inconnu
	PERL0411746	Substitution	exon	inconnu
	SALK_037206.51.85.x	Insertion	instigateur	inconnu
	SALK_106690.47.80.x	Insertion	instigateur	inconnu

III. Les étapes de la recherche dans la base de données PDB

Dans cette recherche on va préciser les protéines pour ce- là on va choisir une base de données protéique PDB (protéine Data Bank) pour faire une comparaison correcte entre les deux bases de données, la recherche comprend les étapes suivantes :

- Etape 01 : La précision de la base de données PDB.

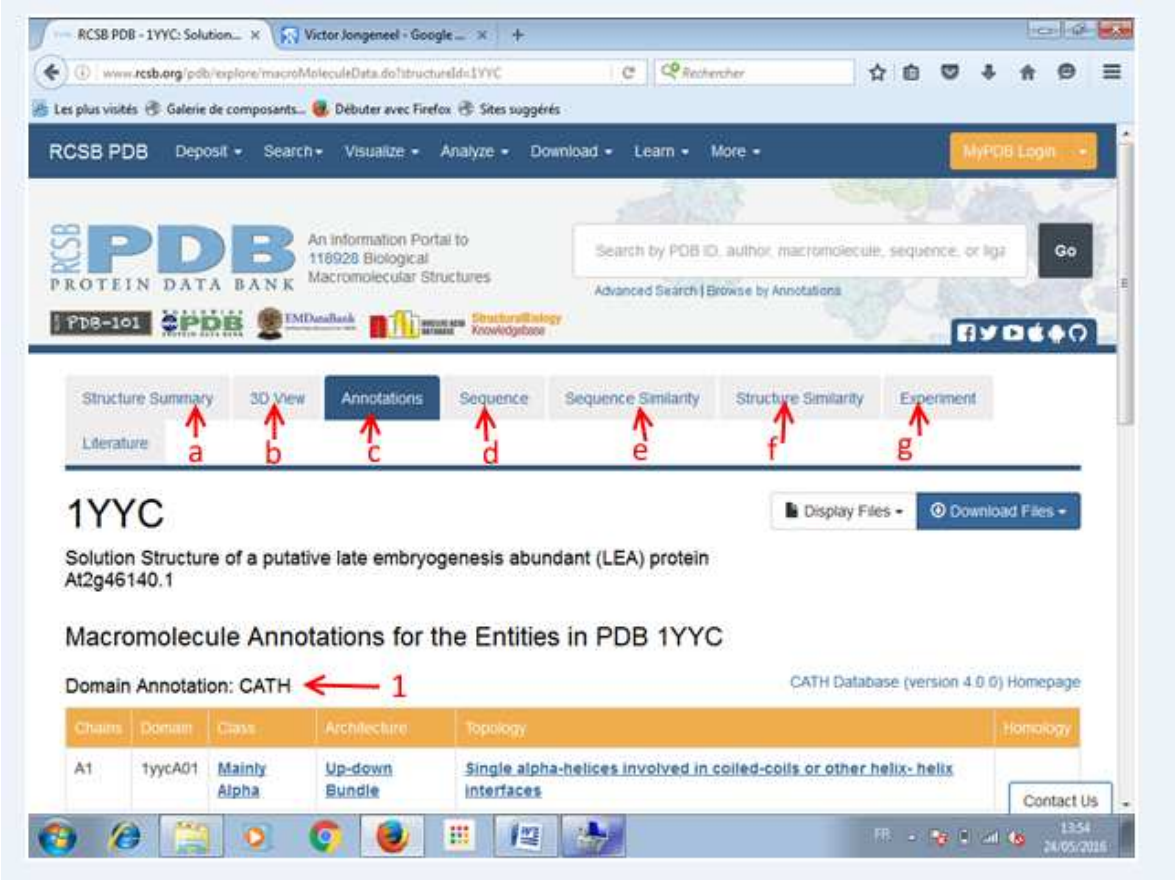


Figure 13 : La banque de donn**ée** PDB contient un menue varie pour présentée LEA **(WWW.rcsb.org/pdb)**

(a : structure summary, b : 3D view, c : annotation, d : séquence , e : séquence similarity,f :structure similarity,g : experiment/ 1 :annotation de LEA).

- Etape 02 : L'utilisation du code At2g46140.1 spécifique à la protéine cible LEA.

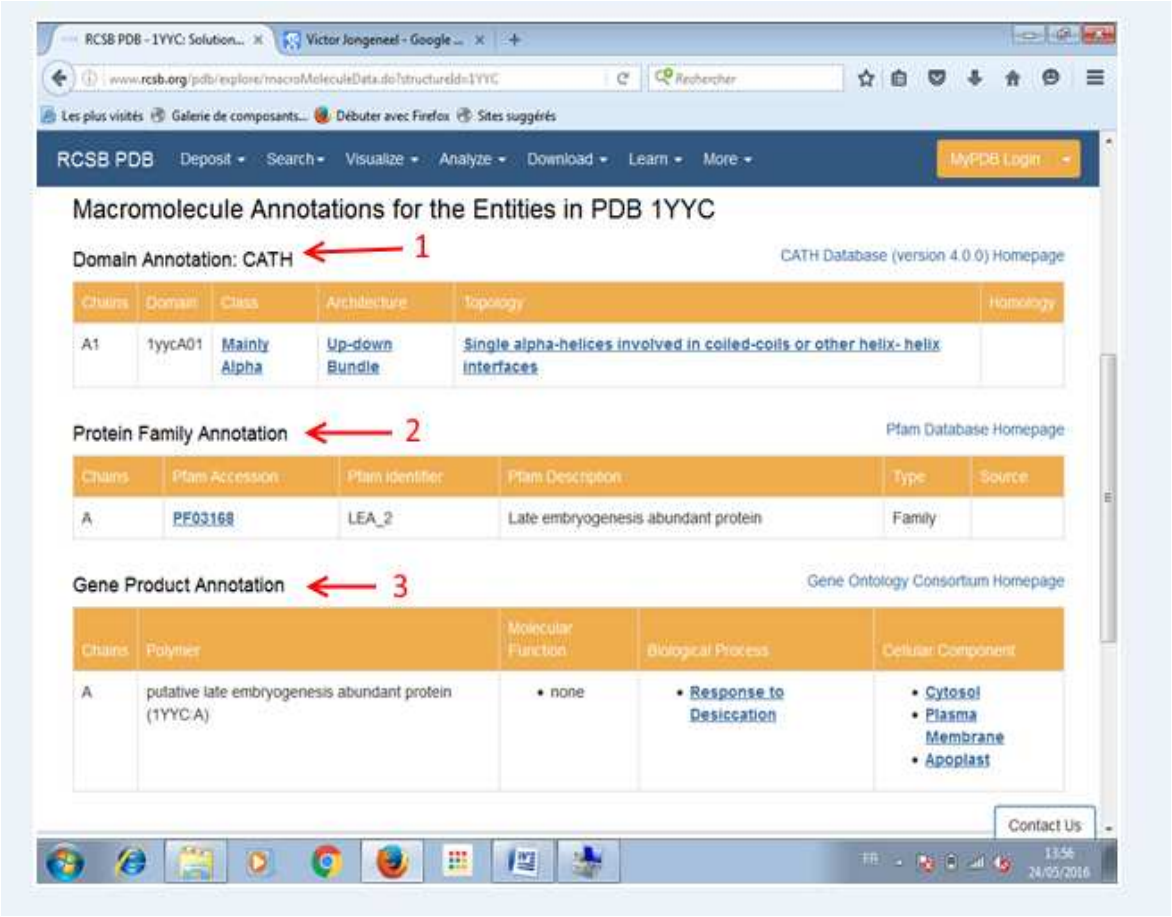


Figure 14 : annotation de macromolécule pour les entités dans la PDB 1yyc **(WWW.rcsb.org/pdb)**

(1 : l'annotation de la protéine LEA, 2 : l'annotation de la famille de protéine, 3 : l'annotation de produit de gêne).

- Etape 03 : Après l'étape passée on trouve une page WEB qui présente un menu varié (structure, 3D view, annotation….)

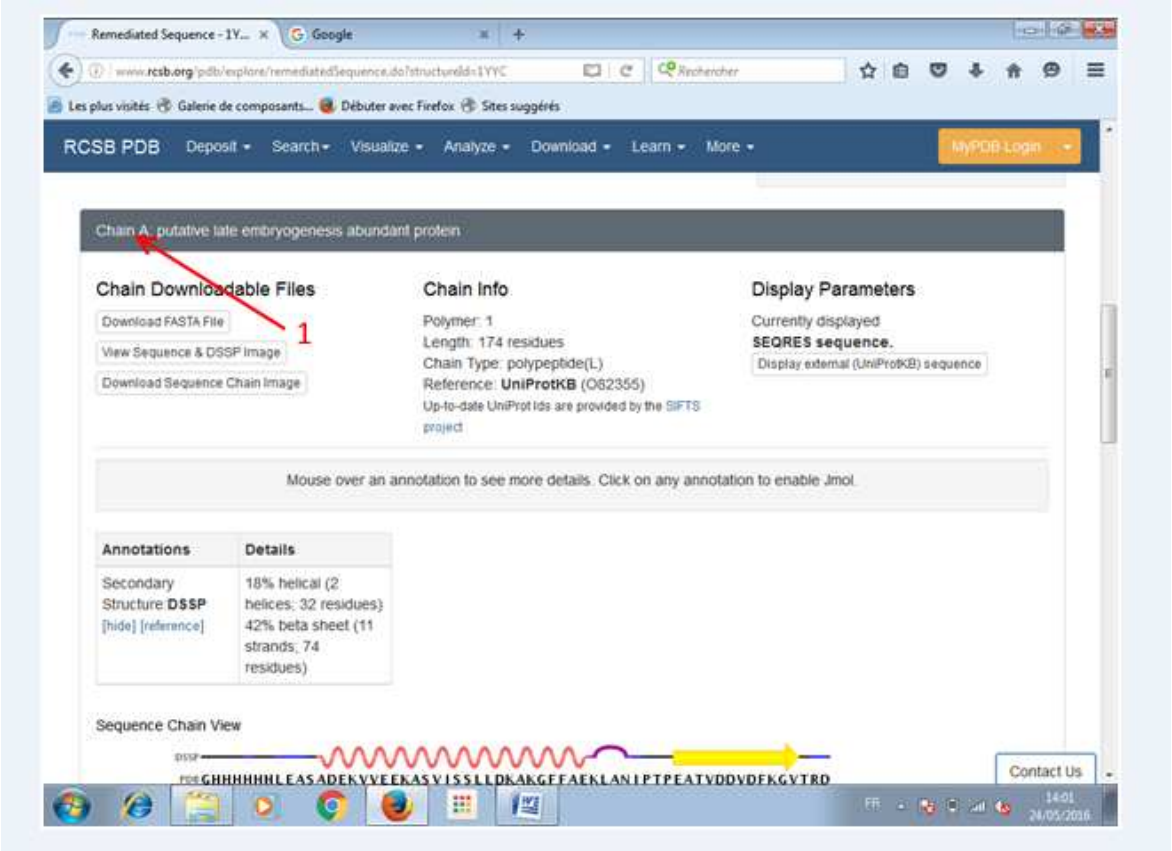


Figure 15 : La banque de donnée PDB pour montrer la structure secondaire de LEA **(WWW.rcsb.org/pdb)**

(1 : information de chaine A).

- Etape 04 : A chaque fois on va ouvrir les pages de menue et enregistrer les pages par leur ordre.

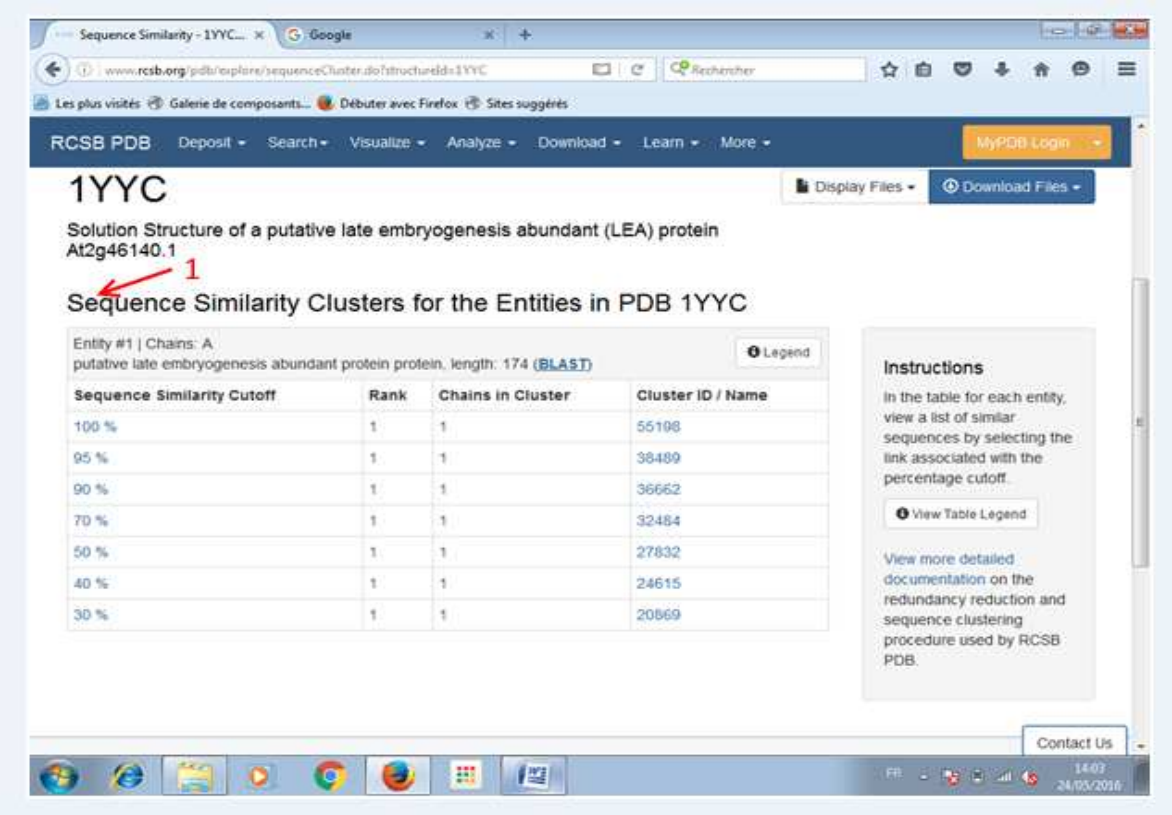


Figure 16 : La banque de donnée PDB représentée la structure similarité de LEA.

III.1 La base de données PDB représentée la protéine LEA ;

A) ANNOTATION

1YYC

Structure de solution d'une protéine abondante At2g46140.1 d'embryogenèse (LEA) en retard putative

Tableaux VII : Description de domaine 1yycA01 de la chaine A1.

Chaine	Domaine	Classe	Architecture	Topologie
A1	1yycA01	Principalement alpha	Vers le haut de vers le bas empaqueter	Choisir les alpha-spirales impliquées dans les enroulements enroulés ou d'autres interfaces de spirale

Tableaux VIII: Détermination de la fonction de la chaine A.

Chaines	Polymère	Fonction moléculaire	Processus biologique	Composant cellulaire
A	protéine abondante d'embryogenèse enretard putative (1yyc : A)	Aucun	Réponse à la dessiccation	-Cytosol -Membrane de plasma -Apoplast

B) Similarité de structure

Options d'affichage

*Chaînes uniques actuellement de visionnement seulement.

*Rapports d'ordre et de structure. Permettre à Jmol de regarder des annotations dans 3D.

*Réduction de redondance et groupement d'ordre. Regarder les résultats de groupement pour 1YYC.

*Enchaîner A : protéine abondante d'embryogenèse en retard putative.

Information de chaîne

Polymère : 1

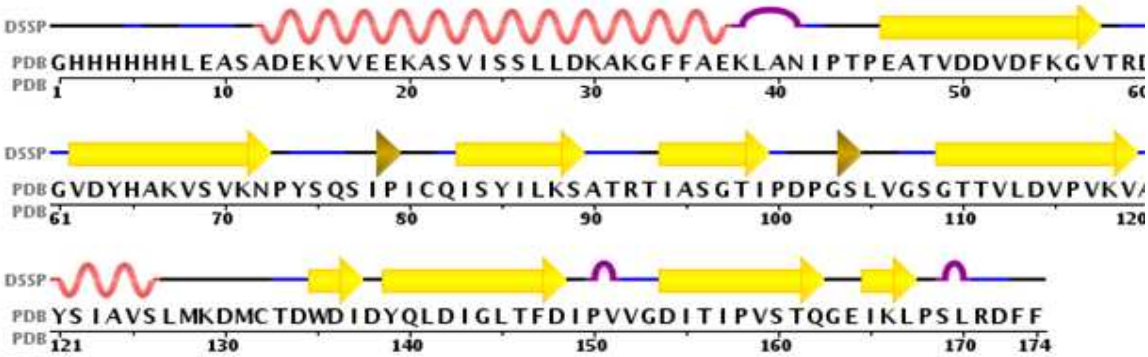
Longueur : 174 résidus

Type à chaînes : polypeptide (L)

Référence : UniProtKB (O82355).

Annotation	Détails
Structure secondaire : DSSP	18% hélicoïdal (2 spirales ; 32 bêta feuille des résidus) 42% (11 rives ; 74 résidus)

Vue à chaînes d'ordre :



Structure secondaire de DSSP

Dictionnaire de structure secondaire de protéine : reconnaissance des structures des dispositifs hydrogène-collés et géométriques.

C) Structure summury

DOI: 10.2210/pdb1yyc/pdb

•Classification : fonction inconnue de génomique structurale

•Organisation : Thaliana d'Arabidopsis

•Système d'expression : synthèse sans cellule

Instantané expérimental de données :

* Méthode : Solution RMN.

* Conformer a calculé : 100.

*Conformer a soumis : 20.

*Critères de sélection : Fonction de cible.

Littérature :

1yyc

Structure de solution d'une protéine abondante At2g46140.1 d'embryogenèse (LEA) en retard putative.

Macromolécules

*Poids de structure totale : 18859.50

Tableaux IX : La protéine abondante d'Arabidopsis thaliana.

Entités de macromolécule				
Molécule	Chaînes	longueur	organisation	détail
protéine abondante d'embryogenèse en retard putative	A	174	Arabidopsis thaliana	Noms de gêne : At2g46140 T3F17.21

Validation de structure

Le rapport de validation n'est pas disponible pour cette entrée RMN

D) Expérience

*Données expérimentales RMN de solution.

*Détails expérimentaux.

Tableau X : Données expérimentales RMN de solution

États d'échantillon

Prélever le contenu	BRI-Tris de 10 millimètres, 100 millimètres de Na Cl, 10 millimètres DTT, pH 7.0, 90% H2O, 10% D2O
Dissolvant	90% H2O,10% D2O
Concentration ionique	n/a
Ph	7.0
Pression	1atm
Température	298.2
Expérience	D_13C-separated_NOESY, 3D_15N-separated_NOESY, HNCACB, HN (CO) CACB, HCCHTOCSY, CCONH

Information de spectrometer		
Fabricant	Modèle	Intensité de champ
Varian	INOVA	600.0
Bruker	DMX	500.0

Amelioration RMN

Méthode	recuit simulé, dynamique moléculaire
Information RMN ensemble	
Critères de sélection de Conformer	fonction de cible
Conformers a calculé le nombre total soumis par Conformers	100
total de nombre	20

Discussion

➢ La discussion :

D'après les résultats obtenus entre la base des données TAIR (the Arabidopsis information ressource) et la base des données PDB (Protéine data base) il parait que la protéine LEA présentée dans la base des données TAIR d'une façon différente à la façon de la présentation qui se trouve dans la PDB.

Les bases des données sont soit généralistes ou spécialisées, les banques des données généralistes offrant un ensemble hétérogène d'information pour but d'éviter les redondances, et les bases des données spécialisées regrèvement plus homogènes établies auteur d'une thématique ou d'une méthode spécifique de production des données, (**www.mathon.com/cours TP bioinformatique).**

Nous constatons que la recherche scientifique exiger l'exploitation des bases des données riches, et ce que nous trouvons lors de l'utilisation de la base de données spécifique TAIR , une base des données en biologie constamment mise à jour des données génétiques et de biologie moléculaire pour le modèle *Arabidopsis Thaliana.*

TAIR, contrôle, traite, conserve, maintient et donne accès a l'ensemble des informations génomiques, fonctionnelles, méthodologiques et bibliographiques concernant *Arabidopsis thaliana* **(Samson ,2002) .**

Les structures dans la PDB sont essentiellement déterminées par cristallographie aux rayons X ou par spectroscopie RMN.

Protéine Data Bank ou PDB est une collection mondiale de données sur la structure tridimensionnelle, de macromolécules biologique (protéine), essentiellement et acide nucléique, ces structures sont déterminées par cristallographie aux rayons X et spectroscopie RMN, ces données expérimentales du monde entier et appartiennent au domaine public (**Bernstein et al 2003) .**

Les données expérimentale dans la PDB et la base des données TAIR sont déposées par des biologistes et des biochimistes du monde entier et appartiennent au domaine public.

Par l'utilisation de protéine LEA comme modèle et déterminer la forme présentée dans TAIR et PDB, nous pouvons confirmer l'existence comptabilité entre les bases des données, ou ils ont présenté la protéine par deux façon différente **(Swareck,wiliks,2008)** .

Conclusion

➢ Conclusion :

La base de données TAIR est considéré comme un contenu scientifique riche de grande avoir un impact positif sur la recherche scientifique en fournissant des informations riche et varié par des Informations circulant dans le service de domaine scientifique et il nous permet de recommander l'utilisation de la base de données comme un outil efficace pour faciliter l'accès à des informations spécifiques :

- La base de données **TAIR** pour la plante modèle Arabidopsis Thaliana. Les données fournies par TAIR incluent l'ordre complet de génome avec la structure de gêne, l'information sur le produit de gêne, l'expression de gêne, les stocks d'ADN, les cartes de génome, et les informations sur la communauté de la recherche d'Arabidopsis.
- La banque de données **PDB** est une collection mondiale de données sur la structure tridimensionnelle (ou structure 3D) de macromolécules biologiques : protéines, essentiellement, et acides nucléiques.
- L'utilisation de la protéine LEA comme modèle, elle a permis de déterminée de la déférence dans la fourniture des données entre la base de données spécialisée TAIR et la banque de données généraliste PDB
- La comptabilité entre la base de données **TAIR** et la base de données **PDB** c'est-à-dire chaque base contient une présentation spécifique de la protéine LEA (Late Embryogenesis Abundunt Protéine)

Références bibliographiques

Listes des références bibliographiques

1. Attwood T.K., Gisel A., Eriksson N-E. and Bongcam-Rudloff E., « *Concepts, HistoricalMilestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective* » [archive], Bioinformatics - Trends and Methodologies, InTech, 2011 (consulté le 8 janvier 2012 .
2. Berge JB ,Ricroch AE , (2011) evaluation of genetically engineered crops using transcriptomic, proteomic and metabolomic profiling technique plant physiol 055 : 152-1761.
3. Bernard CAUDRON, « Biologie la bio-informatique » Encyclopaediauniversalis en ligne.consulte le 18 mai 2016.URL.
4. Carlos Coronel, Steven Morris ET Peter Rob, *Database Systems: Design, Implementation, and Management*, Cengage Learning - 2012, (ISBN 9781111969608) .
5. Colin Ritchie, *Database Principles and Design*, Cengage Learning EMEA - 2008, (ISBN 9781844805402) .
6. Cornette*et al*. (2010) "*Identification of Anhydrobiosis-related Genes from an Expressed Sequence Tag Database in the Cryptobiotic Midge Polypedilumvanderplanki (Diptera; Chironomidae)*" *J. Biol. Chem.* 285, 35889 – 35899.
7. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.
8. Hanada K, Zhang X, Borevitz JO, Li WH, Shiu SH. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. Genome Res. 2007; 17:632–640.
9. Hays JB. *Arabidopsis thaliana*, a versatile model system for study of eukaryotic genome-maintenance functions. DNA Repair. 2002; 1:579–600. [PubMed].
10. Hesper ET P Hogeweg, « *Bioinformatica: eenwerkconcept* », *Kameleon*, vol. 1, n^{o} 6, 1970, p. 28–29.
11. Hirayama & Shinozaki (2010) "Research on plant abiotic stress responses in the post-genome era: past, present and future" *The Plant Journal* 61, 1041 – 1052.
12. Jean-Baptiste Waldner, *Nano-informatique et Intelligence Ambiante - Inventer l'Ordinateur du XXIe Siècle*, London, Hermes Science (réimpr. 2007), 121 p. (ISBN 2-7462-1516-0.

13. Koornneef M, Meinke D. The development of *Arabidopsis* as a model plant. Plant J. 2010; 61:909–921. [PubMed.
14. Lamesch Philippe and romanaMadipu Lauren M, Brinkac Jennifer Harrow Laurns G. Wilming Ulrike Bohme,Hannick; Meeting report: a workshop on best practices in genome annotation 2010.
15. Li yuling, Karen yook; todd W. Harris TamberlynBieri, Ab.
16. NCBI Centre américain pour les informations biotechnologiques.
17. Principe d'utilisation des outils/Denis Tagu,Jean-Loup Risler,Coord.
18. Schlaich NL. *Arabidopsis thaliana* – the model plant to study host-pathogen interactions. Curr. Drug Targets. 2011; 12:955–966. [PubMed].
19. Site des Étudiants de la filière de Bio-informatique ET Bio Statistiques d'Orsay [archive] (Description, Ressources, Wiki, et forum sur la bio-informatique).
20. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006; 34: W435–W439. [PMC free article] [PubMed].
21. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 2008; 36:D1009–D1014. [PMC free article].
22. Van Auken K, Jaffery J, Chan J, Müller HM, Sternberg PW. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. BMC Bioinformatics. 2009; 10:228. [PMC free article] [PubMed].
23. Xu XM, Møller SG. The value of *Arabidopsis* research in understanding human disease states. Curr. Opin. Biotechnol. 2011; 22:300–307. [PubMed].

Sites web

http://www.adps.fr/banque-de-donnees-16252.html.

http://www.araport.org

http://www.ebi.ac.uk/GOA/arabidopsis_release, GOA Arabidopsis (version 118).

http://www.universalis.fr/encyclopedie/biologie

Résumé :

Le but de cette étude est de prouver qu'il y a une comptabilité dans l'exploitation des données concernant la façon de présentation des informations biologiques sur un produit biologique quelconque (LEA).

Les résultats obtenus trouvent cette expression on utilisant la base de données spécifique (TAIR) destiné une espèce modèle Arabidopsis Thaliana dans le domaine végétale, et une banque de donnée générale des protéines ce qui me preuve conforment que pour une recherche scientifique à profonde, il faut faire recours des bases plus spécialisé.

Abstract:

The aim of this study is to prove that there is an accounting in the use of data about how the presentation of biological information on any biological product (LEA).

The results found that expression is using the specific database (TAIR) for a model species Arabidopsis thaliana in the plant area, and a general database of proteins which show me conform for a deep scientific research, it is necessary make use of the specialized databases

الملخص:

الهدف من هذه الدراسة هو تحديد التكامل في طرق تقديم المعطيات التي تتضمنها كل من قاعدة المعطيات المتخصصةTAIR وقاعدة المعطيات العامة BDP .

من اجل التأكد من وجود طرق مختلفة لتقديم المعارف استعملنا لهذا الغرض البروتين LEAالذي قدمته قاعدة المعطيات TAIRمركزتا على بنيته العامة . وفي المقابل قدمه بنك المعطيات BDP بشكل يبرز أساسا وضيفته و تخصصه.

- تبين النتائج السابقة وجود تكامل بين المعارف المقدمة من طرف قواعد المعطيات المتخصصة و قواعد المعطيات العامة.