Introduction à la Bioinformatique

2023

Ezechiel Bionimian TIBIRI, Ph.D

Biologie moléculaire – Bioinformatique

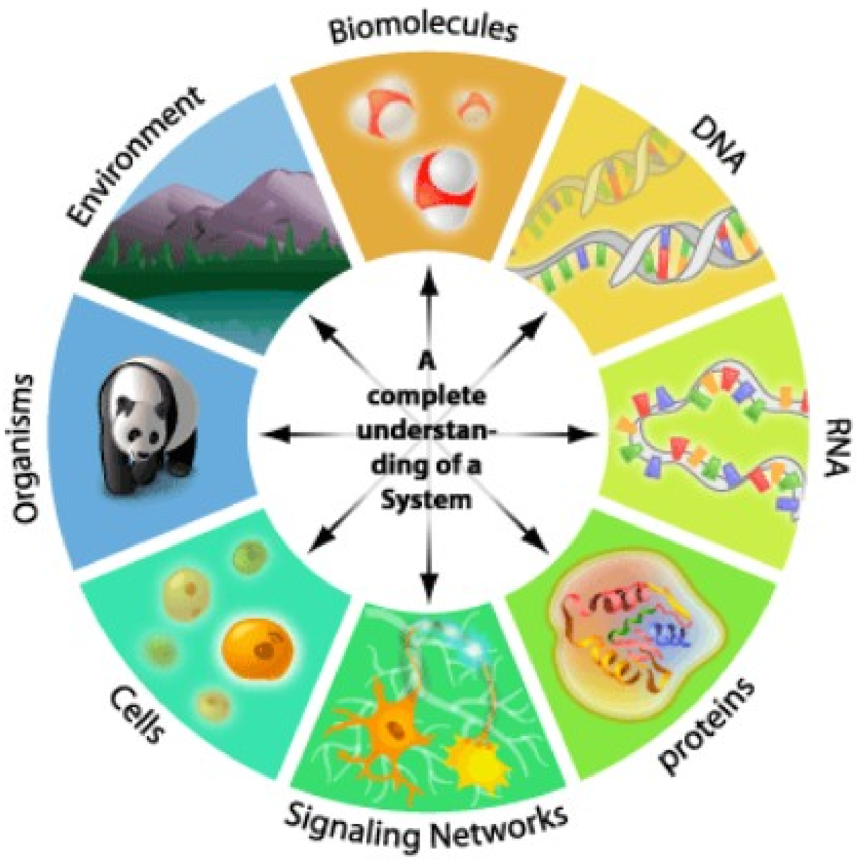
Introduction aux bases de données et aux ressources

Recherche de séquences

Objectifs du cours

- Comprendre la structure et la disposition des ressources de données du NCBI et de l’EBI
- Comprendre la différence entre les bases de données, les outils et les repositories.
- Recherche de données dans des bases de données spécifiques à l'aide de numéros d'accès, de noms de gènes, etc.
- Utiliser les ressources NCBI et EBI

Data


Biomolecules
DNA
RNA
proteins
Signaling Networks
Cells
Organisms
Environment
A complete understanding of a System


BIG
DATA
INFORMATION
STORAGE
LARGE
ANALYSIS
RESEARCH
SEARCH
PETABYTES
MANAGEMENT
SOFTWARE
SETS
MILLION
DEVELOPMENT
TECHNOLOGY
SCIENCE
VOLUME
PROCESS
INTERNET
LEARNING
data
big
approaches
analytics
networks
looking
clinical
predict
Security

Introduction

- Plusieurs bases de données et ressources en ligne
- Besoin de savoir laquelle :
  - ✓ Quelles sont les bases de données et les ressources existantes
  - ✓ Quels sont les outils disponibles pour exploiter ces ressources ?
  - ✓ Quels sont les outils disponibles pour rechercher dans les ressources?

Bases de données biologiques

Bases de données biologiques

- Bases de données biologiques sont :

  - ✓ Publique ou privée
  - ✓ Protéine, nucléotide, structure, littérature, annotation…
  - ✓ Généralisée ou spécialisée
  - ✓ Centré sur la séquence (aa ou nt) ou le génome

Bases de données biologiques

Quelques noms de banques de données:

- **Séquences en acides nucléiques** (DNA et mRNA); EMBL, GenBank, DDBJ
- **Séquences en acides aminés (**protéines); Swiss-Prot, wwPDB

- **Références bibliographiques**; PubMed

- **Informations générales sur les gènes et/ou les maladies**; EntrezGene, OMIM, HMGD

- **informations sur la structure tridimensionnelle des protéines ou de l'ADN**; PDB

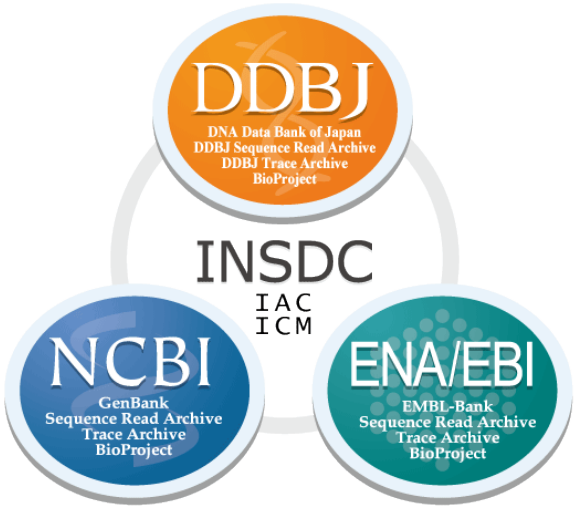
- Il existe aussi des banques spécialisées, comme Newt, qui donne des informations sur la classification des espèces

Bases de données primaires

- International Nucleotide Sequence Database Collaboration (INSDC)

- Données de séquences génomiques stockées dans 3 bases de données publiques

- Chacun a son propre numéro d'accès et ses propres outils



Bases de données secondaires

- Des bases de données spécialisées construites à partir de données de séquences primaires

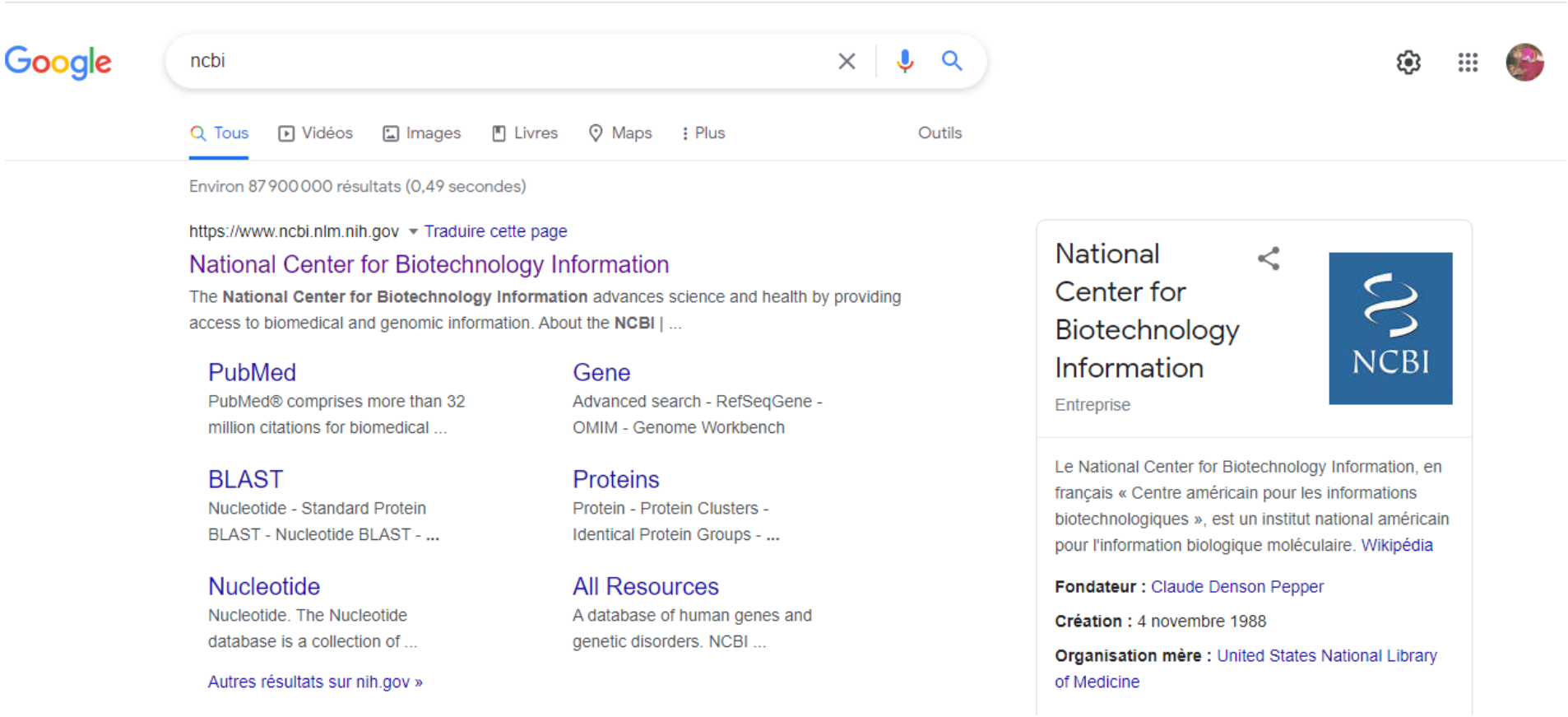
- Fournissent plusieurs ressources et annotations différentes

Ressources bioinformatiques les plus populaires

- National Centre for Biotechnology Information (NCBI)

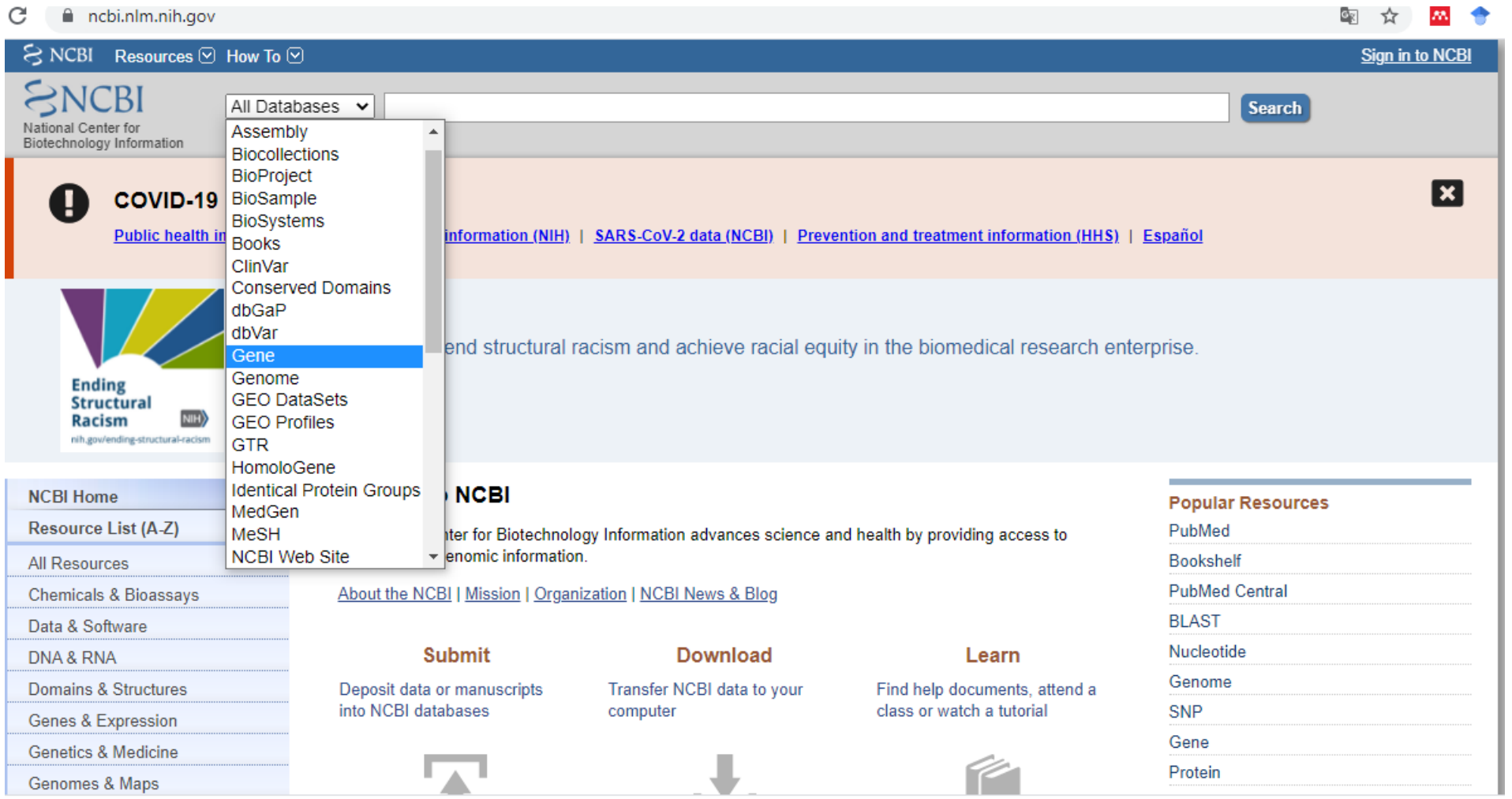
- European Bioinformatics Institute (EMBL-EBI)

Recherche dans les bases de données: NCBI



Recherche dans les bases de données: NCBI

Recherche dans les bases de données: NCBI



Bases de données de NCBI

- NCBI comprend plus de 30 bases de données
- la littérature : PubMed Central (PMC), Bookshelf et PubReader
- La santé: ClinVar, dbGaP, dbMHC, the Genetic Testing Registry, HIV-1/Human Protein Interaction Database et MedGen
- Les génomes: BioProject, Assembly, Genome, BioSample, dbSNP, dbVar, Nucleotide, Probe et RefSeq.
- Les génes: Gene, Gene Expression Omnibus (GEO), HomoloGene, PopSet, Refseq et UniGene.
- Les proteines: Protein, the Conserved Domain Database (CDD), COBALT, Conserved Domain Architecture Retrieval Tool (CDART), the Molecular Modeling Database (MMDB), Refseqp et Protein Clusters.
- Les produits chimiques: Biosystems et PubChem

EMBL - EBI

- Maintenir la gamme la plus complète au monde de bases de données moléculaires librement accessibles et actualisées

- Proposer des formations en ligne et en direct pour l'utilisation de leurs ressources.

- https://www.ebi.ac.uk/training

EMBL - EBI

The EMBL-EBI website has been redesigned. Please send us feedback about this page.

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

Find a gene, protein or chemical | All | Search

Example searches: blast keratin bfl1 | About EBI Search

Find data resources | Submit data | Explore our research | Train with us

Latest news

Organisations should embrace open science faster – interview with Prof. Dame Janet Thornton

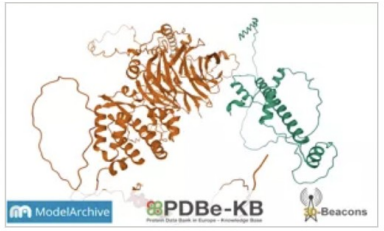
17 May 2022

Europe PMC: Harnessing the power of text mining to accelerate life sciences research

12 May 2022

2.4 billion sequences now available in the latest MGnify protein database release

11 May 2022



Predicted complexes from ModelArchive now on PDBe-KB pages

6 May 2022

EMBL - EBI

Services

Overview | A to Z | Data submission | Research infrastructure development programme | Support

The European Bioinformatics Institute (EMBL-EBI) maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically.

— You can read more about our services in the journal Nucleic Acids Research

Tools & Data Resources

Search all tools & data resources

Tools

Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Web API | Multiple sequence alignment

InterProScan

InterProScan searches sequences against InterPro's predictive protein signatures.

Web API | Protein feature detection | Sequence motif recognition

BLAST [protein]

Fast local similarity search tool for protein sequence databases.

Web API | Sequence similarity search

BLAST [nucleotide]

Fast local similarity search tool for nucleotide sequence databases.

Data resources

Ensembl

Genome browser, API and database, providing access to reference genome annotation

Web API | EMBL-EBI Terms of use

UniProt

A comprehensive resource for protein sequence and functional annotation.

Web API | CC-BY

PDBe

The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.

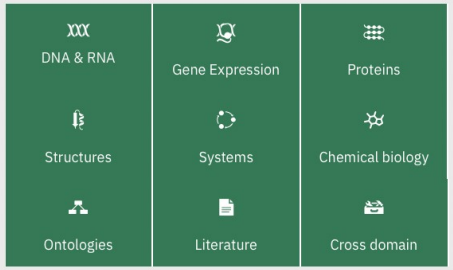
Web API | CC0

Europe PMC

A database to search the worldwide life sciences literature

Web API | EMBL-EBI Terms of use

Browse by type



Programmatic access

EMBL-EBI web services allow you to query our large biological data resources programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications. The Web Services technology we use are built on open standards to ensure client and server software from various sources will work well together.

Browse EMBL-EBI web services

Principles of service provision

Open

Our data and tools are freely available, without restriction. The only exception is potentially identifiable human genetic information, for which access depends

Bases de données spécialisées

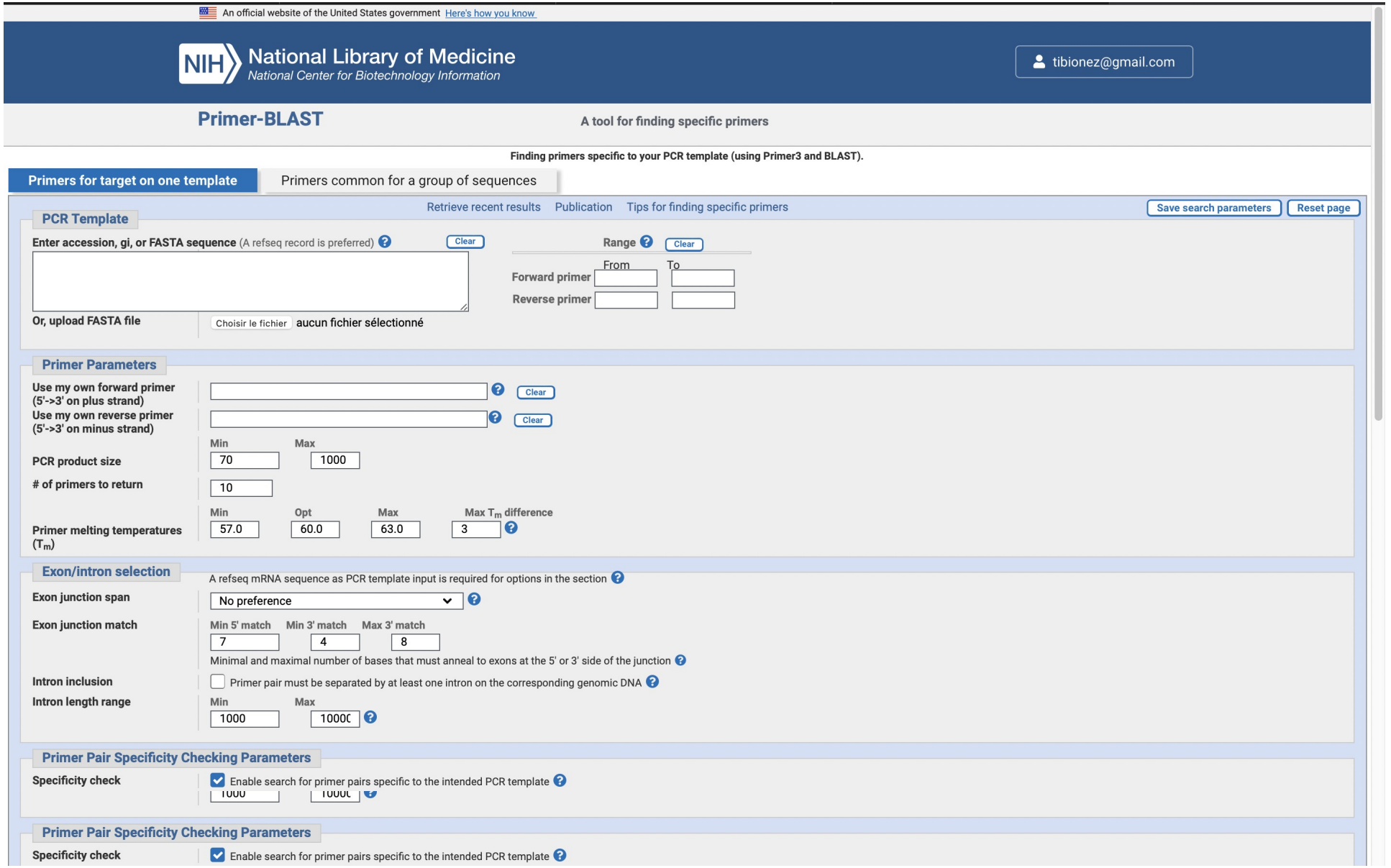
- Il existe un grand nombre de bases de données spécialisées

- La plupart des séquences sont également dans la banque GenBank/EMBL
- Peut contenir des génomes entiers
- Peut contenir des ressources spécialisées
- Contient des outils spécifiques pour l'exploitation des données

Bases de données spécialisées

- Plasmodium https://plasmodb.org/plasmo/app
- Les collections spécialisées de Sanger https://www.sanger.ac.uk
- Base de données sur les hépatites https://hcv.lanl.gov/content/sequence/HCV/news/old_news.html
- Base de données de recherche sur la grippe influenza https://www.fludb.org/brc/home.spg?decorator=influenza

Design d’amorce utilisant Primer Blast



Take home

- Une grande quantité de données existent
- Les bases de données primaires stockent les données brutes des séquences
- Les bases de données secondaires fournissent des informations sur l'annotation des données de séquence.
- Il est important de savoir comment et où les données sont stockées
- NCBI et EBI sont les deux ressources les plus populaires pour obtenir des données biologiques.