

# MR-biosoft

## Guide d'utilisation

Théo Roncalli

Gustavo Magnaña López

Université Paris-Saclay

2021-2022

Le présent site web permet de réaliser l'analyse fonctionnelle de génome bactériens. Le site se veut interactif : l'utilisateur n'a pas besoin de suivre une formation ou d'avoir des connaissances en informatique pour utiliser les différentes fonctionnalités proposées. En raison de la contrainte temporelle, certaines fonctionnalités ne sont pas disponibles, en particulier la gestion des utilisateurs et par conséquent la partie annotation. Toutefois, des fonctionnalités supplémentaires et conviviales ont été proposées et sont décrites dans ce guide. De plus, l'ensemble des fonctionnalités ont été ultra-optimisés.

La page home (voir image ci-dessous) constitue la racine du site web : c'est ici que vous décidez quelles fonctionnalités vous souhaitez utiliser pour l'analyse fonctionnelle de génomes bactériens. La page home distingue deux parties majeures : l'analyse de séquences et l'annotation. L'analyse de séquence concerne la recherche de séquences (génomique, génique ou protéique) en fonction de certains critères décidés par l'utilisateur. Cette recherche de séquences peut porter sur des génomes, des gènes ou des protéines. La seconde partie concerne l'annotation composée de trois sections : la première concerne l'importation de données permettant ainsi à l'utilisateur de fournir davantage de séquences nucléotidiques et protéiques à la base de données, la seconde renvoie l'ensemble des séquences allouées à l'annotateur et la dernière renvoie un module pour demander aux validateurs de fournir des séquences à l'utilisateur (si celui-ci est annotateur). Comme évoqué précédemment, ces deux dernières fonctionnalités ne sont pas encore réalisables.

Remarquons également un champs de navigation en haut de la page. Ce champs de navigation est accessible depuis n'importe quelle page du site. Si vous cliquez sur *support*, vous atterrissez dans votre boîte mail électronique et pouvez envoyer un message électronique directement à l'administrateur. Ce module est à utiliser uniquement si vous souhaitez remonter une information jugée importante (bug rencontré par exemple). Si vous cliquez sur *home*, vous revenez à la page principale. Bon, cela n'a pas grand intérêt tant que nous sommes dans la page principale. Je vous propose donc d'essayer les différentes fonctionnalités proposées par le site, en commençant par l'analyse fonctionnelle génomique.

[Documentation](#) [Support](#) [Logout](#) [Home](#)

## Home

### The browser for bacterial genome annotation

Expanding platform that encompasses annotations and functional analysis of bacterial genomes



## Sequence Search & Annotation

### Sequence Search

#### Genome

Browser for retrieving bacterial genomes

#### Gene

Browser for gene sequences and functional annotations

#### Protein

Browser for protein sequences and functional annotations

### Annotation

#### Import

Upload FASTA files

#### Basket

Set of allocated entries

#### Request

Request sequences to annotate

Si vous cliquez sur *Genome*, vous accéderez à un formulaire pour rechercher des informations sur des génomes bactériens. Nous remarquons qu'il y a deux formulaires. Le premier permet de rechercher des informations sur un génome bactérien spécifique. Si vous connaissez à l'avance le génome

d'intérêt, il vous suffit d'entrer l'identifiant du chromosome (qui est unique) et d'appuyer sur le bouton *launch query*. Vous obtiendrez donc des informations sur le génome d'intérêt. Toutefois, il est possible que vous ne connaissiez pas à l'avance le génome d'intérêt ou même encore son identifiant, mais que vous ayez des informations sur le type de génome qui vous intéresse. Dans ce cas-là, vous pouvez remplir le second formulaire avec les informations dont vous disposez. Les différentes informations pouvant être fournies concernent l'espèce et la souche, le taille minimum et maximum du génome (en paire de bases) ou même encore un motif. Si vous recherchez donc tous les génomes qui contiennent un motif en particulier, il vous suffit uniquement d'entrer ce motif à l'endroit correspondant. De plus, vous pouvez bien entendu combiner les informations entrées. Afin de rendre ce formulaire conviviale, aucun problème de casse n'est possible lorsque vous entrez le nom de l'espèce ou de la souche. Également, il est permis de ne rentrer qu'une partie du nom de l'espèce considérée. Si par exemple, vous écrivez *coli*, vous obtiendrez tous les génomes des Escherichia Coli qui sont dans le base.

## Bacterial genome search

### The browser for bacterial genome annotation

Expanding platform that encompasses annotations and functional analysis of bacterial genomes



**Specific genome search**  
Chromosome

**Genome set search**  
**Taxonomy**  
Specie  Strain   
  
**Size Genome**  
Minimum (in bp)  Maximum (in bp)   
  
**Other characteristics**  
Motif

Ce formulaire, une fois remplie et soumis à la base, permet de renvoyer un tableau de résultats avec l'ensemble des génomes qui répondent aux critères que vous recherchez. Dans ce tableau de résultat, vous trouverez quelques informations, notamment le nom du chromosome, l'espèce, la souche et la longueur en paire de bases. Vous pouvez cliquer sur le nom du chromosome et ainsi récupérer diverses informations sur celui-ci. Un exemple vous est fourni ci-dessous pour le chromosome ASM584v2. Une partie est dédiée aux informations générales, une autre fournit des hyperliens vers des informations du génome d'intérêt produites par des sources externes (NCBI et EBI). La dernière partie renvoie la séquence génomique. Notons qu'une API est également disponible pour récupérer cette page web en écrivant dans la barre de recherche le lien suivant :

<nom serveur>/browse/genome?chromosome=<identifiant chromosome>

Bien sûr, cette API est utile uniquement si vous connaissez à l'avance l'identifiant du chromosome. Cela permet également d'envoyer la page web à un autre utilisateur afin que ce dernier obtienne directement la page d'information du génome d'intérêt.

## Entry ASM584v2 (Escherichia coli)

General Information		External link
Chromosome:	ASM584v2	<p>You can have complementary information with the following databanks:</p> <ul style="list-style-type: none"><li>National Center for Biotechnology Information (NCBI)</li><li>European Bioinformatics Institute (EBI)</li></ul>
Specie:	Escherichia coli	
Strain:	k12	
Length (bp):	4641652	

### Sequence

AGCTTTTCATTCTGACTGCAACGGGCAATATGCTCTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTTTATTTGACTTAGGTCACTAAATACCTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCCACCATACCACCATCA CCATTACCACAGGTAACGGTGC GGCTGACGCGTACAGGAAACACAGAAAAAGCCCGCACCTGACAGTGGCGGCTTTTTTTTCGACCAAAGGTAAACGAGGTAACAAC CATGCGAGTGTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTCGTGTGGCGATATCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTC CTCTCTGCCCCCGCCAAATACCAACCACTGGTGGCGATGATTGAAAAAACCATAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTAATTTTGCCGAA CTTTTGACGGGACTGCGCGCCCGCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGCTGATCAGGAATTTGCCCAAATAAAACATGTCTGTCATGGCATTAGTTT GTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTGCGGTGGCGAGAAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCAACAACGTTACT GTTATCGATCCGGTCAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCACCCTGCATATTGCTGAGTCCACCCGCGCTATTGCGGCAAGCCGATTCCGGCTGATCAC ATGGTGCTGATGGCAGTTTACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGTCTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCTGTTACGCG CCGATTGTCGAGATTGGACGAGCGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCGATGCGAGGTTGTTGAAGTCGATGCTTACCAGGAAGCGATGGA GCCTTTCCTACTTCGGCGCTAAAGTTCTTACCCCCGACCAATTACCCCATCGCCAGTTCAGATCCCTTGCCTGATTAATAAACCCGAAATCCTCAAGCACCGGTACG CTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACATGGCAATGTTACGCGTTTCTGGTCCGGGGATGAAGGGGATGGTCGG CATGGCGCGCGCGCTCTTTGCAGCGATGTCACGCGCCGCTATTTCGTGGTGTGCTGATTACGCAATCATCTTCCGAATACAGCATCAGTTTCTGCGTCCACAAAGCGACTG TGTGCGAGCTGAACGGGCAATGCAGGAAGAGTTCTACCTGGAACGAAAGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACGGCTGGCCATTATCTCGGTGGTAGG TGATGGTATGCGCACCTTTCGCTGGGATCTCGGCGAAATCTTTGCCGCACTGGCCCGCGCAATATCAACATTGTCGCCATTGCTCAGGGATCTTCTGAACGCTCAATCTC TGCTGTGTAATAACGATGATGCGACCACCTGGCGTGCGCGTTACTCATCAGATGCTGTTCATACCCGATCAGGTTATCGAAGTGTGTTGTGATTGGCGTCCGTTGGCGTTGG CGGTGCGCTGCTGGAGCACTGAAGCGTCAGCAAGCTGGCTGAAGAATAACATATCAGTCTACGTGTCTGCGGTGTTGCCAACTCGAAGGCTCTGCTACCAATGTA CATGGCCTTAATCTGGAAGAACTGGCAGGAAGAACTGGCGCAAGCAGGCGCTTTAATCTCGGGCGCTTAATTCGCTGTTGCGTGAAGAAAGTATCATCTGCTGAACCCGGT CATTTGTAAGTCTCACTTCAGCCAGGCGAGTGGCGGATCAATATGCCGACTCTCTGCGCGAAGGTTTCCACGTTGTCACGCCGAACAAAAGGCCAACCACTCTGCTGATGG ATTAATACCATCAGTTGCGTTATGCGCGGAAAAATCGCGCGTAAATCTCTATGACACCAACGTTGGGGCTGGATTACCGGTTATTGAGAACCTGCAAAATCTGCTCA ATGCAAGGTGATGAATTGATGAAGTTCTCCGGCATTCTTCTGGTTCGCTTTTATATCTTCGGCAAGTTAGACGAAGGCATGAGTTTCTCCGAGGCGACCACTGGCTGGCGG GGAATGGGTTATACCGAACCGGACCCGCGAGATGATCTTCTGGTATGGATGGCGCGTAACTATTGATTCTCGCTGCGTGAACCGGACGCTGAATGGAGCTGGCG GATATTGAAATTGAACCTGTGCTGCCCGCAGATTTAACGCCGAGGGTGATGTTGCCGCTTTTATGGCGAATCTGTCACTACGACGATCTCTTTCGCCGCGCGCTGGC GAAGGCCGCTGATGAAGGAAAGTTTTCGCTATGTTGGCAATATTGATGAAGATGGCGTCTGCCGCGTGAAGATTGCCGAAGTGGAATGGTAATGATCCGCTGTCAAAG TGAATAATGGCGAAACGCCCTGGCCTTCTATAGCCACTATTATCAGCCGCTGCCGTTGGTACTGCGCGGATATGGTGGCGGCAATGACGTTACAGCTGCCGCTGCTCTT

Maintenant que nous avons fait le tour concernant le génome, intéressons-nous à la recherche de séquences et d'informations sur les gènes et protéines. Pour cela, il faut appuyer sur *gene* ou *protein* sur la page d'accueil. Dans ce manuel, nous décrivons uniquement les fonctionnalités autour des gènes, car celles-ci sont très similaires entre les gènes et protéines, si ce n'est que les motifs et les séquences renvoyés sont des acides nucléiques ou acides aminés (et également que le blast n'est permis que pour les gènes, faute de temps). Comme pour le génome, nous atterrissons sur un formulaire (voir image ci-dessous). Néanmoins, les informations demandés sont différentes cette fois-ci. Dans ce formulaire, soit vous connaissez à l'avance l'accession number du gène d'intérêt et vous utilisez donc le premier formulaire, soit vous recherchez un ensemble de gènes et vous utilisez donc le second formulaire. Dans ce second formulaire, vous pouvez spécifier des informations sur la taxonomie (numéro d'identifiant du chromosome, l'espèce ou bien encore la souche). Vous pouvez également préciser la taille du gène, dans un certain intervalle, ou encore le nom, le symbole ou la fonction génique. Également, un motif peut être spécifié pour rechercher tous les gènes contenant la séquence que vous recherchez. De plus, vous pouvez spécifier le sens de lecture du gène.

### Specific gene search

Accession number

Launch Query

### Gene set search

#### Taxonomy

Chromosome
Specie
Strain

#### Length Gene

Minimum (in bp)
Maximum (in bp)

#### Annotation

Name
Symbol
Function

#### Other characteristics

Motif
Reading Frame

Launch Query

Supposons que vous recherchez tous les gènes synthétisant des kinases. Il vous suffit de taper *kinase* en dessous de *Function*. Vous obtiendrez tous les gènes dont la fonction contient le mot *kinase* (et même s’il y a d’autres mots avant ou après dans la fonction). La figure ci-dessous renvoie le résultat obtenu lorsque l’utilisateur demande les gènes synthétisant une kinase, ayant une taille minimum de 2400 paires de base et dont le sens de lecture soit direct. Ce tableau contient plusieurs entrées, pour lesquels nous avons l’accèsion number, l’identifiant du chromosome associé, l’espèce, la souche, la longueur du gène, le symbole génique et le fonction. Une fonctionnalité intéressante est que le tableau peut être trié par ordre croissant ou décroissant en appuyant sur le nom de colonne d’intérêt. De plus, vous pouvez cliquer sur l’accèsion number (resp. l’identifiant du chromosome) pour avoir plus d’informations sur le gène (resp. génome).

## Results

AC	Chromosome	Specie	Strain	Length (bp)	Symbol	Function
<a href="#">AAC73113</a>	<a href="#">ASM584v2</a>	Escherichia coli	k12	2463	thrA	Bifunctional aspartokinase/homoserine dehydrogenase 1
<a href="#">AAC75429</a>	<a href="#">ASM584v2</a>	Escherichia coli	k12	3594	evgS	hybrid sensory histidine kinase in two-component regulatory system with EvgA
<a href="#">AAC75828</a>	<a href="#">ASM584v2</a>	Escherichia coli	k12	2757	barA	hybrid sensory histidine kinase, in two-component regulatory system with UvrY
<a href="#">AAC76922</a>	<a href="#">ASM584v2</a>	Escherichia coli	k12	2433	metL	Bifunctional aspartokinase/homoserine dehydrogenase 2
<a href="#">AAG54302</a>	<a href="#">ASM666v1</a>	Escherichia coli	edl933	2463	thrA	aspartokinase I, homoserine dehydrogenase I
<a href="#">AAG57130</a>	<a href="#">ASM666v1</a>	Escherichia coli	edl933	2808	None	partial putative sensor kinase
<a href="#">AAG59141</a>	<a href="#">ASM666v1</a>	Escherichia coli	edl933	2433	metL	aspartokinase II and homoserine dehydrogenase II
<a href="#">AAN78503</a>	<a href="#">ASM744v1</a>	Escherichia coli	cft073	2526	thrA	Aspartokinase I
<a href="#">AAN81213</a>	<a href="#">ASM744v1</a>	Escherichia coli	cft073	2673	yojN	Putative sensor-like histidine kinase yojN

Cliquons sur la première entrée par exemple : AAC73113. Nous obtenons la fiche d’information ci-dessous. Nous avons un bloc pour les informations générales, un bloc pour les informations liées aux

annotations, un bloc contenant la séquence en acides nucléiques, un bloc pour des ressources externes (NCBI, ENA et UniProtKB) et un bloc pour lancer un blast. Les hyperliens vers les ressources externes renvoient directement vers les fiches d'information du gène d'intérêt (11C73113 dans notre cas) via l'utilisation des API de ces sites. Notons également que le présent site dispose d'une API pour les gènes (et protéines). Il suffit pour cela d'écrire :

<nom serveur>/browse/gene?gene=<accession number>

## Entry AAC73113 (Escherichia coli)

General Information		Annotation	
Accession number:	AAC73113	Gene name:	b0002
Specie:	Escherichia coli	Gene symbol:	thrA
Strain:	k12	Gene biotype:	protein_coding
Chromosome:	ASM584v2	Function:	Bifunctional aspartokinase/homoserine dehydrogenase 1
Length (bp):	2463		
Start position:	337		
End position:	2799		
Reading frame:	direct		

  

Sequence	External link
<p>ATGCGAGTGTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCT  GCGTGTGCGGATATTCTGAAAAGCAATGCCAGGCAGGGGAGGTGGCCAC  CGTCTCTCTGCCCCGCCAAATCACCAACCACTGTGTGGCGATGATTGAA  AAAACATTAGCGGGCAGGATGCTTTACCAATATCAGCGATGCCGAACGTAT  TTTTGCCGAACTTTGTGCGGACTCGCCGCCAGCCGCGGGTTCCCGCTG  GCGCAATTGAAAACTTTCGTGATCAGGAATTTGCCAAATAAACATGTCCT  GCATGGCATTAGTTTGTGCGGAGTGGCGGATAGCATCAACGCTGCGCTG  ATTTGCCGTGGCGAGAAATGTCGATCGCCATTATGGCCGCGTATTAGAAGC  GCGCGGTGACAACTTACTGTATCGATCGGTCGAAAACTGCTGGCAGTG  GGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCGCCGATTGC  GGCAAGCCGATTCCGGCTGATCAGATGGTCTGATGGCAGGTTTCACCGCC  GGTAATGAAAAAGCGAACTGGTGGTGTGTTGGACGCAACGGTTCCGACTACT  CTGCTGCGGTGCTGGCTGCTGTTTACGCGCGGATTGTTGCGAGATTGGAC  GGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCGATGCGAG  GTTGTTGAAGTCGATGCTCTACCGAAGCGATGGAGCTTCTACTTCGGCG  CTAAAGTTCTTACCCCGCACCAATTACCCCATCGCCAGTTCAGATCCCTT</p>	<p>You can have complementary information with the following databanks:</p> <ul style="list-style-type: none"> <li>• <a href="#">National Center for Biotechnology Information (NCBI)</a></li> <li>• <a href="#">European Nucleotide Archive (ENA)</a></li> <li>• <a href="#">UniProtKB</a></li> </ul>

  

Blast
<p>Database <input type="text" value="Nucleotide collection (nr)"/></p> <p>Optimization <input type="text" value="blastn"/></p> <p>Word Size <input type="text" value="10"/></p> <p><input type="button" value="Launch Blast"/></p>

Le dernier bloc est très intéressant puisqu'il est dédié au blast, fonctionnalité très utile pour l'annotation lorsque vous ne connaissez pas la fonction ni l'espèce du gène d'intérêt. Cette fonctionnalité est lancée sur les serveurs du NCBI mais les informations sont récupérées et traitées sur notre site MR-biosoft. Différentes options peuvent être spécifiées : la base de données sur laquelle est lancée blast, l'algorithme d'optimisation et la taille des mots minimum (par défaut, 10). Les bases de données fournies sont Nucleotide collection (nr), Reference RNA sequences (refseq\_rna), Patent sequences (patnt), et PDB nucleotide. Les deux algorithmes d'optimisation proposés sont blastn

(séquences similaires) et megablast (séquences fortement similaires). Supposons que nous voulons lancer blast avec la base de données refseq\_rna et l'algorithme blastn. Les résultats obtenus sont fournis ci-dessous. Pour chaque hit sont fournis le GenInfo Identifier (gi), l'identifiant gb/emb, le score d'alignement, le bits, la E-value, la couverture, le pourcentage d'identité et les gaps. Pour avoir davantage d'informations sur le hit (notamment connaître la taxonomie et la fonction du hit), vous pouvez cliquer sur le GenInfo Identifier qui vous redirigera automatiquement dans un nouvel onglet vers la fiche d'information de celui-ci (sur le NCBI).

## Blast Hits

gi	gb/emb	Score	Bits	$-\log_{10}(E \text{ value})$	Coverage	Identities	Gaps
<a href="#">2082271452</a>	XM_043072817.1	87.0	79.7328	9.3486	136	99	0
<a href="#">2082271450</a>	XM_043072818.1	87.0	79.7328	9.3486	136	99	0
<a href="#">1531824892</a>	XM_027304344.1	78.0	71.6177	7.1771	123	90	2
<a href="#">1527478562</a>	XM_027235780.1	78.0	71.6177	7.1771	123	90	2
<a href="#">1527473488</a>	XM_027238743.1	78.0	71.6177	7.1771	123	90	2
<a href="#">1527473486</a>	XM_027238742.1	78.0	71.6177	7.1771	123	90	2
<a href="#">545704168</a>	XM_005704351.1	73.0	67.1093	5.5485	79	62	0
<a href="#">1778663601</a>	XM_002623377.2	69.0	63.5025	4.4628	67	54	0
<a href="#">1778663599</a>	XM_002623376.2	69.0	63.5025	4.4628	67	54	0
<a href="#">1419011280</a>	XM_025572861.1	67.0	61.6992	3.9199	76	59	0
<a href="#">296810685</a>	XM_002845635.1	67.0	61.6992	3.9199	81	62	0
<a href="#">760448824</a>	XM_011403179.1	66.0	60.7975	3.9199	137	97	2
<a href="#">302831068</a>	XM_002947054.1	65.0	59.8958	3.3771	138	99	4
<a href="#">1484792821</a>	XM_026532665.1	64.0	58.9941	3.3771	107	77	0

Maintenant que nous avons décrit toute la partie axée autour de la recherche de séquences, nous pouvons décrire la dernière fonctionnalité proposée, qui est l'envoi de données sur le serveur. Si vous souhaitez bénéficier de cette fonctionnalité, vous pouvez appuyer sur *Import* dans la page d'accueil. Quatre champs sont proposés (voir image ci-dessous). Les deux premières fonctionnalités sont obligatoires : l'envoi d'un fichier que le serveur va analyser et le type de séquence que contient le fichier (génomique, gènes, protéines). Vous pouvez également préciser l'espèce et la souche. Dans le cas où vous ne les connaissez pas, vous pouvez laisser ces champs vides. Également, ces champs n'ont d'intérêt que lorsque le génome est envoyé. S'il s'agit de séquences géniques ou protéiques, le serveur va lui-même retrouver de quelle espèce et souche il s'agit, en se basant sur les données chromosomiques. Les fichiers doivent être rentrés dans un ordre précis pour fonctionner : le fichier contenant le génome, puis le fichier contenant les gènes et enfin celui contenant les protéines. Si vous ne respectez pas cet ordre, les séquences ne seront pas enregistrées dans la base de données. Lors de la survenue d'une erreur (condition d'intégrité non respectée par exemple), celle-ci est affichée à l'écran. Toutefois, il existe une erreur que nous n'arrivons pas à afficher à l'écran (faute de temps). Cette erreur apparaît lorsque vous envoyez une séquence protéique et que la séquence du chromosome et les séquences géniques n'ont pas été enregistrées au préalable. Toutefois, même si aucun message d'erreur n'apparaît, les séquences protéiques n'ont pas été envoyées dans la base. L'erreur est donc correctement gérée par le serveur. Également, le parsing des fichiers est ultra-optimisé via l'utilisation de la parallélisation computationnelle.

# File upload

File

**Parcourir...** Aucun fichier sélectionné.

Type

genome

Specie

Strain

Submit