Détecter la bactérie E.Coli dans l'eau

Eléonore d'Agostino et Yannick Widmer

3.5.2016

Contents

1 Introduction

Notre but dans ce projet est de pouvoir concevoir un test facile d'usage permettant à une personne moyenne de tester la présence de la bactérie E.Coli dans de l'eau.

2 Contexte

Plus précisément, nous cherchons à trouver des épitopes sur la protéine **OmpF Porin** de la bactérie E.Coli, qui seraient capables de détecter E.Coli avec 100% de taux de réussite quel que soit ses variations, mais sans détecter d'autres bactéries accidentellement. Ceci fait, notre épitope pourra être utilisé pour génerer un anticorps, qui sera ensuite utilisé dans le test ELISA.

2.1 Escherichia coli

Escherichia coli, communément appelé E.Coli, est une bactérie normalement trouvée dans le système digestif de divers animaux. Ceci en fait un indicateur potentiel pour tester des échantillons pour de la matière fécale. Donc si E.Coli est présent dans de l'eau, c'est un avertissement comme quoi il ferait mieux de ne pas la consommer, sous risque d'intoxication alimentaire.

Additionnellement, E.Coli est une bactérie ayant été énormément utilisée en laboratoire, ce qui fait que plusieurs dizaines de séquences génomiques complètes sont disponible pour analyse. [1]

Nous allons nous pencher sur la protéine OmpF Porin de E.Coli, qui est une protéine faisant partie de sa membrane externe, et permettant la diffusion de petites molécules polaires. OmpF Porin nous donnera 1807 nucléotides avec lesquels travailler, dont 1086 font partie de sa structure primaire. [2]

2.2 Epitopes et Anticorps

Un épitope, ou déterminant antigénique, est une partie d'un antigène, à laquelle un anticorps se lie. Cette liaison est le principe que nous allons utiliser pour faire fonctionner le test. Notre but est de trouver un épitope qui soit partagé par toutes les souches de E.Coli sans être présent dans d'autres bactéries.

Un antigène est une molécule capable de forcer le système immunitaire à produire des anticorps pour la contrer.

Un anticorps est une protéine pouvant se lier à un antigène spécifique. En conditions normales, les anticorps sont utilisés par le système immunitaire pour identifier et neutraliser des pathogènes (bactéries, virus, etc). Dans notre cas, nous allons utiliser les anticorps associés aux épitopes que nous avons choisi pour faire fonctionner le test de détection de E.Coli.

2.3 ELISA

ELISA, acronyme de "Enzyme-Linked ImmunoSorbent Assay", est un test qui utilise des anticorps et des enzymes pour identifier une substance. Il existe plusieurs variations de ELISA, *Direct*, *Sandwich*, et *Competitive*. Nous allons nous intéresser à la variation Sandwich, qui utilise un anticorps de capture, à la différence du Direct qui utilise le principe d'adsorption. [3]

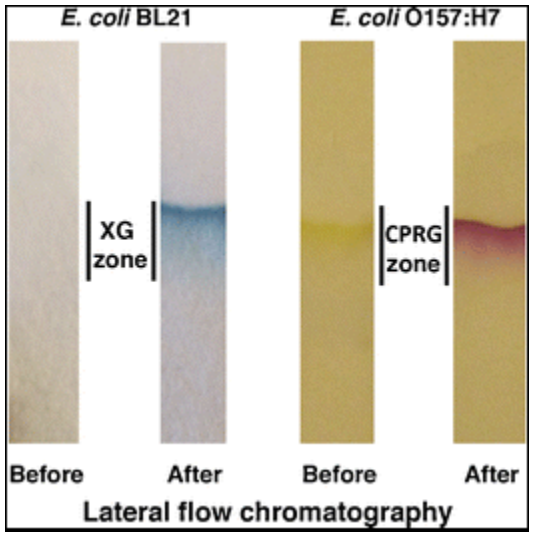
En gros, un test Sandwich ELISA utilise un anticorps de capture sur une surface (dans notre cas, une coque contenant du papier, similaire à un test de grossesse) pour que des antigènes y adhérent. Ensuite, l'anticorps de détection (celui que l'on aura préparé à partir d'un épitope de E.Coli) y est ajouté, et il se lie à l'antigène associé s'il est présent. Ensuite, un dernier anticorps lié à un enzyme est ajouté, puis un subtrate, qui est converti en une couleur.

Ceci fait que le test change de couleur si E.Coli est présent. Typiquement, il y aura aussi une partie qui détecte n'importe-quel antigène, pour vérifier que le test à fonctionné. Si aucun antigène n'est détecté par le test, c'est que les anticorps de capture ne ne sont pas liés à des antigènes, et qu'on ne peut rien affirmer sur la présence ou non d'E.Coli.

3 Etat de l'art

En premier lieu, nous allons d'abord nous intéresser à savoir si des tests existent déjà pour détecter la bactérie E.Coli. Voici une méthode pour détecter la bactérie :

Il existe un test avec une bande papier de (0.5 * 8 cm). Ce papier contient soit le 5-bromo-4-chloro-3-indolyl-B-D-glucuronide sel de sodium (zone XG ci-dessous en image)et/ou le chlorophénol rouge B-galactopyranoside (zone CPRG ci-dessous en image). La bande de papier est ensuite mise dans l'eau avec la bactérie E.Coli pendant 30 minutes. Si le papier change de couleur alors la bactérie E.Coli est présente. [4]



4 Réalisation

Avant de pouvoir trouver les épitopes communs de toutes les souches de E.Coli, il va falloir faire un alignement de séquence avec toutes les souches afin de faire ressortir les régions de séquences similaires.

Un alignement de séquence est une manière de représenter deux ou plusieurs séquences de macromolécules biologiques les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires. L'objectif d'un alignement de séquence est donc d'identifier les zones de concordance.

L'alignement devra être un alignement multiples car nous avons bien plus que deux souches à analyser. Il existe déjà plusieurs algorithmes d'alignement de séquence multiple, nous allons comparer les différents logiciels qui proposent un MSA(Multiple Sequence Alignements) ci-dessous. [5]

ClustalW2 Un des premiers logiciels de MSA datant de 1988. Les auteurs cite que le logiciel est clairement dépassé. Nous n'allons donc pas utiliser ce logiciel.

MAFFT version 7 possède un des algorithmes les plus rapides à ce jour et est donc l'un des plus utilisés. Les alignements générés par cet outil sont de bonne qualité mais possède certaines fois quelques petites erreurs.

MUSCLE est un peu plus lent que MAFFT mais rien de très grave. Son alignement est un poile meilleur que celui de MAFFT. En plus de cela nous l'avons utilisé pendant notre laboratoire 2. Nous sommes déjà habitués à ce logiciel.

Clustal Omega il utilise une méthode d'alignement progressive tandis que Muscle utilise une méthode d'alignement itérative. L'avantage de la méthode itérative est que l'on peut revenir au calcule précédent du meilleur alignement. Il y'a donc plus d'erreur possible avec une méthode d'alignement progressive que itérative.

Nous allons donc nous orienter vers l'algorithme MUSCLE pour pouvoir aligner nos séquences de façon efficace. MAFFT aurait aussi pu être un choix mais le fait que nous sommes déjà familiarisé avec MUSCLE fait pencher la balance. [6]

4.1 Obtention des séquences

Nous avons commencé par faire des recherches sur GenBank pour pouvoir obtenir toutes les séquences possibles de E.Coli, ou plus précisément, de sa protéine OmpF. De toute façon nous n'allions analyser que la partie OmpF, donc nous n'avions pas besoin du reste de la séquence d'E.Coli, particulièrement les séquences partielles ne contenant pas OmpF.

La requête était relativement simple: `"ompf"[gene] AND "Escherichia coli"[porgn]`, dans laquelle nous avons filtré pour ne télecharger que les contenus des séquences des gènes OmpF.

Le résultat de cette recherche nous a fourni une liste de 1279 séquences, sur lesquelles nous avons du en éliminer deux, **EHX66884.1** et **AAP13245.1**. Alors que la taille moyenne des séquences était aux alentours de 360, ces deux avaient une taille de moins de 100, vu qu'elles correspondaient à des séquences partielles de OmpF Porin.

Ceci nous laisse 1277 séquences sur lesquelles travailler.

4.2 Critères de sélection d'épitope

Pour le résumer de manière concise, nous voulons trouver la séquence la plus longue commune entre toutes les souches, et qui n'est pas présente dans d'autres organismes. Avoir ceci parfaitement ne sera pas possible, mais nous allons tenter de maximiser la sensibilité et la spécificité de nos résultats.

Additionellement, chaque séquence va être testée pour voir si elle est propice à des épitopes ou non.

5 Résultats

Nous avons d'abord aligné les séquences obtenues avec MUSCLE, puis tenté de trouver les séquences les plus longues possibles, partagées par le plus de souches possibles, via AliView, ce qui nous donna cette liste:

Séquence	Longueur	Présence E.Coli	Epitopes
MKRNILAVIVPALLVAGTANAAEIYNKDGNKV... ...DLYGKAVGLHYFSKGNGENSYGGNGDMTYARL... ...GFKGETQINSDLTGYGGQWEYNFQGNNSE	92	83.95%	2
GADAQTGNKTRLAFAGLKYADVGSFDYGRNYG... ...VVYDALGYTDMLPEFGGDT	51	86.45%	1
AYSDDFFVGRVGGVATYRNSNFFGLVDGLNFA... ...VQYLGKNER	41	99.14%	0
DTARRSNGDGVGGSISYEYEGFGIVGAYGAAD... ...RTNLQEAQLLGN	44	27.25%	1
GKKAEQWATGLKYDANNIYLAANYGETRNATP	32	98.20%	1
FANKTQDVLLVAQYQFDFGLRPSIAYTKSKAKDVEGIG	38	93.81%	0
DVDLVNYFEVGATYYFNKNMSTYVDYIINQID	32	96.32%	0
GVGSDDTVAVGIVYQF	16	97.49%	0

La colonne **Présence E.Coli** correspond à dans quel pourcentage des 1277 souches cette séquence est présente. Nous ne considérons dans **Epitopes** que les épitopes possibles de longueur 8 et plus.

Ceci fait, nous avons utilisé un site pour détecter les épitopes y étant présents. [7]

Pour chacun de ces épitopes, nous avons calculé dans combien de souches il était présent, ainsi que fait une recherche sur BLAST pour vérifier sa présence dans d'autres organismes.

Epitope	Longueur	Sensibilité
ANAAEIYNKDGNKVD	15	96.00%
KGNGENSYGGNGDM	14	86.69%
GADAQTGNK	9	93.58%
DTARRSNGDGVGGSISY	17	95.54%
GKKAEQWA	8	98.82%

A la base nous comptions aussi calculer la spécificité de chaque épitope, mais nous avons eu deux problèmes:

1. La spécificité est calculée comme (# de vrais negatifs)/(# de vrais negatifs + # de faux positifs). Notre problème vient du fait que calculer le nombre de faux positifs est simple en théorie - cela revient à compter le nombre d'organismes qui pourraient contenir l'épitope. Par contre, pour le nombre de vrais négatifs, on ne sait pas par où commencer. Le plus logique serait de se concentrer "que" sur les organismes pouvant être présent dans l'eau, mais il n'existe pas de réference simple pour ces informations.

2. Obtenir le nombre de faux positifs est en effet simple en théorie, mais pas du tout en pratique. Nous nous attendions à pouvoir faire une recherche sur BLAST avec l'épitope, et comparer les résultats, mais comme plusieurs organismes ont des souches listées plus d'une fois, que certains sont listés comme étant E.Coli et d'autres seulement OmpF Porin, nous n'avions pas de moyen de trier la liste sans tout analyser à la main (à plus de 3000 résultats par épitope, pas faisable). Nous sommes certains que certains épitopes sont communs à d'autres organismes, mais nous n'avons pas pu calculer à quel point.

Ceci dit, en ne se basant que sur les informations que nous avons pu obtenir, notre meilleure option est probablement l'épitope **ANAAEIYNKDGNKVD**, ayant la deuxième sensibilité la plus haute, mais presque le double de la longueur du premier.

6 Conclusion

Ce projet fût très intéressant à suivre pour nous, il y a eu beaucoup de théorie à apprendre, particulièrement sur le début. Additionellement, même la partie *informatique* de la bioinformatique nous a beaucoup changé de notre travail habituel.

Après avoir compris la théorie, nous avons eu plus de peine que prévu pour trouver les épitopes, car malgré le fait que plusieurs étapes sont relativement simples en théorie, en pratique nous avons eu beaucoup de problèmes imprévus.

Pour les épitopes, il fallait constamment faire des compromis car il n'y avait aucunes parties entièrement en commun entre toutes les souches. Nous devions donc décider lesquelles exclure, et ensuite espérer que les épitopes résultants n'appartenaient pas aussi à d'autres organismes, chose que nous n'avons au final pas pu confirmer de manière quantitative.

References

[1] Trudy M. Wassenaar Oksana Lukjancenko and David W. Ussery. Comparison of 61 sequenced escherichia coli genomes. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2974192/, juillet 2010.

[2] Paul Champaloux. Function and structure of ompf porin. http://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2005/Champaloux/first.html, 2005.

[3] Wikipedia. Elisa. https://en.wikipedia.org/wiki/ELISA#Sandwich_ELISA, decembre 2015.

[4] Springer Link. Multiplexed paper test strip for quantitative bacterial detection. http://link.springer.com/article/10.1007%2Fs00216-012-5975-x, juin 2012.

[5] Wikipedia. Multiple sequence alignment. https://en.wikipedia.org/wiki/Multiple_sequence_alignment, mai 2016.

[6] Yoann M. Alignements multiples : quels logiciels choisir ? http://bioinfo-fr.net/alignements-multiples-quels-logiciels-choisir, mai 2012.

[7] IEDB Solutions Center. Antibody epitope prediction. http://tools.immuneepitope.org/bcell/, 2005-2016.

A Annexes

A.1 Repository GitHub

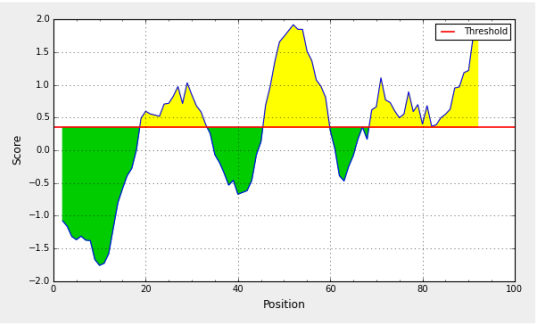
Les sources du projet sont disponibles sur un repo GitHub, à l'addresse
`https://github.com/younTheory/bbcislife`

A.2 Détection d'épitopes

Input Sequences

1 MKRNILAVIV PALLIAGTAN AAEIYNKDGN KVDLYGKAVG LHYFSKGNGE NSYGGNGDMT
61 YARLGFKGET QINSDLTGYG QWEYNFQGNN SE

Center position: 4 Window size: 7 Threshold: 0.35 Recalculate



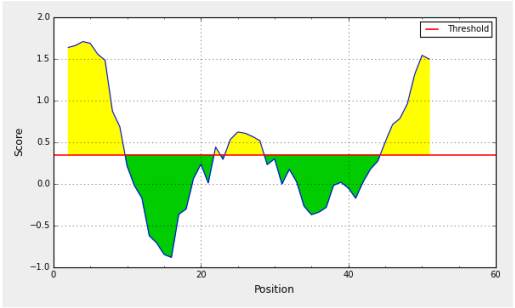
Average: 0.277 Minimum: -1.761 Maximum: 1.921

Predicted peptides:

No.	Start	End	Peptide	Length
1	19	33	ANAAEIYNKDGNKVD	15
2	46	59	KGNGENSYGGNGDM	14
3	67	67	K	1

1 GADAQTGNKT RLAFAGLKYA DVGSFDYGRN YGVVYDALGY TDMLPEFGGD T

Center position: 4 Window size: 7 Threshold: 0.35 Recalculate



Average: 0.396 Minimum: -0.885 Maximum: 1.712

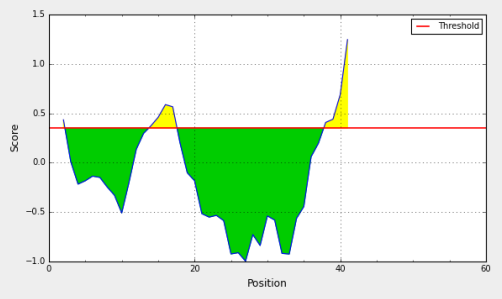
Predicted peptides:

No.	Start	End	Peptide	Length
1	1	9	GADAQTGNK	9
2	22	22	V	1
3	24	28	SFDYG	5

Input Sequences

1 AYSDDFFVGR VGGVATYRNS NFFGLVDGLN FAVQYLGKNE R

Center position: 4 Window size: 7 Threshold: 0.35 Recalculate



Average: -0.139 Minimum: -1.000 Maximum: 1.249

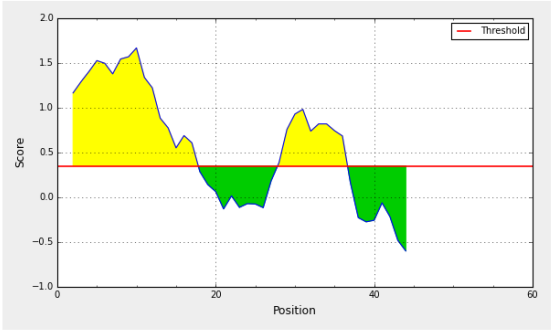
Predicted peptides:

No.	Start	End	Peptide	Length
1	1	2	AY	2
2	14	17	VATY	4

Input Sequences

1 DTARRSNGDG VGGSISYEYE GFGIVGAYGA ADRTNLQEAQ LLGN

Center position: 4 Window size: 7 Threshold: 0.35 Recalculate



Average: 0.576 Minimum: -0.602 Maximum: 1.672

Predicted peptides:

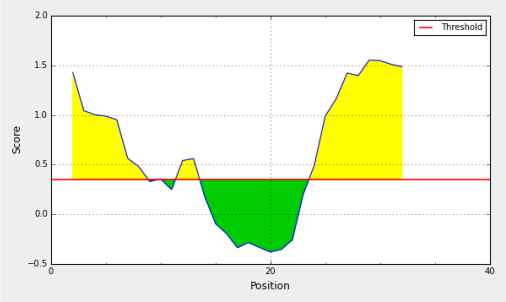
No.	Start	End	Peptide	Length
1	1	17	DTARRSNGDGVGGSISY	17
2	28	36	YGAADRTNL	9

Bepipred Linear Epitope Prediction Prediction

Input Sequences

```
1 GKKAEQNATG LKYDANNIYL AANYGETRNA TP
```

Center position: 4 **Window size:** 7 **Threshold:** 0.35 Recalculate



Average: 0.613 **Minimum:** -0.382 **Maximum:** 1.555

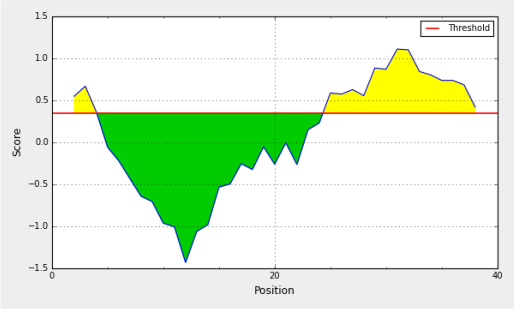
Predicted peptides:

No.	Start	End	Peptide	Length
1	1	8	GKKAEQNA	8
2	10	10	G	1
3	12	13	KY	2

Input Sequences

```
1 FANKTQDVLL VAQYQFDFGL RPSIAYTKSK AKDVEGIG
```

Center position: 4 **Window size:** 7 **Threshold:** 0.35 Recalculate



Average: 0.085 **Minimum:** -1.435 **Maximum:** 1.114

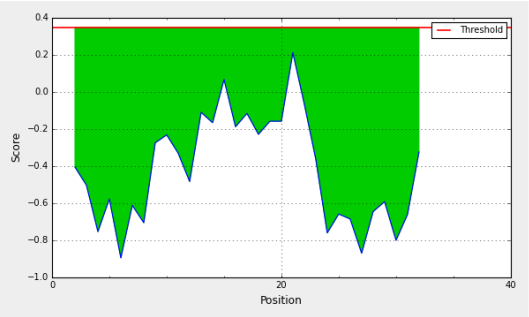
Predicted peptides:

No.	Start	End	Peptide	Length
1	1	4	FANK	4

Input Sequences

```
1 DVDLVNYFEV GATYYFNKNM STYVDYIINQ ID
```

Center position: 4 **Window size:** 7 **Threshold:** 0.35 Recalculate

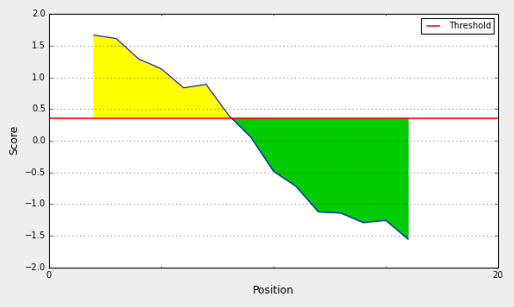


Average: -0.421 **Minimum:** -0.896 **Maximum:** 0.214

Input Sequences

```
1 GVGSDDTVAV GIVYQF
```

Center position: 4 **Window size:** 7 **Threshold:** 0.35 Recalculate



Average: 0.118 **Minimum:** -1.557 **Maximum:** 1.670

Predicted peptides:

No.	Start	End	Peptide	Length
1	1	8	GVGSDDTV	8