(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété Intellectuelle
Bureau international

(43) Date de la publication internationale
21 janvier 2021 (21.01.2021)

WIPO | PCT

(10) Numéro de publication internationale
WO 2021/008878 A1

(51) Classification internationale des brevets :
G16H 50/80 (2018.01) *G16B 20/20* (2019.01)

(21) Numéro de la demande internationale :
PCT/EP2020/068611

(22) Date de dépôt international :
02 juillet 2020 (02.07.2020)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
19186032.9 12 juillet 2019 (12.07.2019) EP

(71) Déposant : **BIOMÉRIEUX** [FR/FR] ; 69280 Marcy l'Etoile (FR).

(72) Inventeurs : **KANEKO, Gaël** ; 1147 avenue Marcel Mérieux, 69280 Marcy l'Etoile (FR). **GUIGON, Ghislaine** ; 4 allée des Vignes, 69570 Dardilly (FR).

(74) Mandataire : **LE MAUFF, Frédéric** ; bioMérieux, 69280 MARCY L'ETOILE (FR).

(81) États désignés *(sauf indication contraire, pour tout titre de protection nationale disponible)* : AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) États désignés *(sauf indication contraire, pour tout titre de protection régionale disponible)* : ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), eurasien (AM, AZ, BY, KG, KZ, RU, TJ, TM), européen (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Publiée:
— *avec rapport de recherche internationale (Art. 21(3))*

WO 2021/008878 A1

(54) Title: METHOD FOR EPIDEMIOLOGICAL IDENTIFICATION AND MONITORING OF A BACTERIAL OUTBREAK

(54) Titre : PROCEDE D'IDENTIFICATION ET DE SURVEILLANCE EPIDEMIOLOGIQUE D'UN FOYER BACTERIEN

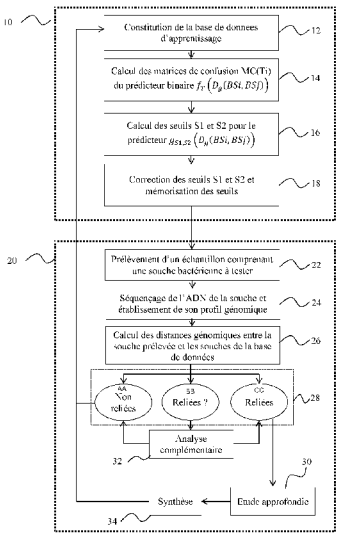


Figure 1

12 Creation of the learning database
14 Calculation of confusion matrices MC(Ti) of the binary predictor
16 Calculation of thresholds S1 and S2 for the predictor
18 Correction of thresholds S1 and S2 and saving of the thresholds
22 Collection of a sample comprising a bacterial strain to be tested
24 Sequencing the DNA of the strain and establishing its genetic profile
26 Calculation of genomic distances between the collected strain and the strains from the database
AA Not related
BB Related?
CC Related
30 Deeper study
32 Complementary analysis
34 Synthesis

(57) Abstract: The invention relates to a method for detecting and monitoring a bacterial outbreak comprising the prediction that a collected bacterial strain and a bacterial strain from a database belong to the bacterial outbreak if their genomic distance is less than a first predetermined threshold, do not belong to the bacterial outbreak if their genomic distance is greater than a second predetermined threshold which is strictly greater than the first threshold or may belong to the bacterial outbreak if their genetic distance is in between. The first threshold is greater than or equal to a third threshold, such that a prediction that two bacterial strains with a genomic distance less than the third threshold belong to the bacterial outbreak has a maximum specificity. The second threshold is less than or equal to a fourth threshold, such that a prediction that two bacterial strains with a genomic distance greater than the fourth threshold do not belong to the bacterial outbreak has a maximum sensitivity.

(57) Abrégé : L'invention a pour objet un procédé de détection et de surveillance d'un foyer bactérien comprenant la prédiction qu'une souche bactérienne prélevée et une souche bactérienne d'une base de données appartiennent au foyer bactérien si leur distance génomique est inférieure à un premier seuil prédéterminé, n'appartiennent pas au foyer bactérien si leur distance génomique est supérieure à un second seuil prédéterminé strictement supérieur au premier seuil, ou appartiennent peut être au foyer bactérien si leur distance génomique est comprise. Le premier seuil est supérieur ou égal à un troisième seuil tel qu'une prédiction d'appartenance au foyer bactérien de deux souches bactériennes ayant une distance génomique inférieure au troisième seuil a une spécificité maximale. Le second seuil est inférieur ou égal à un quatrième seuil tel qu'une prédiction de non appartenance au foyer bactérien de deux souches bactériennes ayant une distance génomique supérieure au quatrième seuil a une sensibilité maximale.

PROCEDE D'IDENTIFICATION ET DE SURVEILLANCE EPIDEMIOLOGIQUE D'UN FOYER BACTERIEN

DOMAINE DE L'INVENTION

La présente invention a trait au domaine de l'épidémiologie bactérienne, en particulier la détection et la surveillance de foyers bactériens en fonction des génomes de souches bactériennes, notamment le séquençage partiel ou total de l'ADN et/ou de l'ARN des souches bactériennes.

ETAT DE LA TECHNIQUE

La détection d'un foyer infectieux (ou « outbreak ») bactérien consiste classiquement à déterminer si plusieurs souches bactériennes prélevées chez des sujets (e.g. des patients et par extension chez des animaux) résultent d'une transmission récente d'une même souche entre les sujets, par exemple la transmission de la souche à plusieurs sujets depuis un sujet « source » ou la transmission de la souche de sujet à sujet. En se fondant sur les outils microbiologiques classiques, la détection est usuellement réalisée en deux temps :

a. dans un premier temps en suspectant un foyer bactérien, cette suspicion survenant lorsque des souches prélevées appartiennent à la même espèce bactérienne et partagent des caractéristiques phénotypiques communes, par exemple un antibiogramme identique ou voisin pour les bactéries pathogènes ;
b. et en cas de suspicion en menant une enquête épidémiologique visant à démontrer, ou infirmer, que ces souches résultent bien d'une transmission entre sujets. Ce type d'enquête consiste notamment à rechercher si les sujets objets des prélèvements ont été récemment en contact, ont partagé une même localisation (e.g. une même salle d'opération ou une même chambre dans un hôpital), ont été soignés par un même personnel soignant, etc. Ce type d'enquête est le plus souvent long et fastidieux, mobilise beaucoup de personnes. En outre, une enquête peut perturber grandement le fonctionnement d'une institution ou d'une société suspectée d'être l'objet d'une épidémie dans la mesure où des mesures prophylactiques sont le plus souvent mises en œuvre avant la fin de l'enquête, comme par exemple la mise en quarantaine d'une chambre, d'un service ou la fermeture d'une salle d'opération.

Dans ce contexte, l'avènement du séquençage, en particulier les séquençages du type WGS (« whole genome sequencing », séquençage de génome entier en français) représente une avancée notable dans l'épidémiologie bactérienne puisqu'un génome bactérien entier contient un niveau d'information bien plus important que celui délivré par les techniques

microbiologiques classiques. Non seulement, les critères pour décider de lancer une étude épidémiologiques sont plus précis mais de plus l'emploi de la génomique peut également grandement simplifier et normaliser celle-ci. Par exemple, si deux souches de staphylocoques dorés, prélevées au sein du même service hospitalier à quelques jours d'intervalle, sont strictement identiques du point de vue génomique, il peut être déterminé sans autre information que les deux souches font bien partie d'un même foyer bactérien.

Si le séquençage se révèle une avancée notable, il ne permet cependant pas à lui seul de déterminer la lignée de deux souches bactériennes quelle que soit l'espèce. En effet, certaines espèces bactériennes ont un génome plastique évoluant très rapidement en l'espace de quelques jours, et ce d'autant plus qu'un traitement antibiotique est mis en œuvre, de sorte qu'une stricte identité entre génomes ne peut être utilisée comme seul critère. Pour tenir compte de cette plasticité, des procédés de détection de foyers bactériens consistent à évaluer que des souches bactériennes appartiennent à un même foyer si leur différence génomique, par exemple calculée en fonction du nombre de polymorphismes d'un seul nucléotide (ou « single-nucleotide polymorphism» en anglais), est inférieure à un seuil prédéterminé. Comme décrit dans l'article « *Beyond the SNP threshold : identifying outbreak clusters using inferred transmission* » de J. Simson et al, dec. 2018, cette approche est cependant peu précise en raison de nombreuses sources d'incertitude, comme par exemple le contexte dans lequel évoluent les bactéries ou la variabilité du taux de mutation en fonction des espèces. Les auteurs de cet article proposent ainsi de tenir compte également de la chronologie des prélèvements d'échantillon contenant les souches bactériennes et de connaissances *a priori* sur les mécanismes de mutation et de transmission des souches bactériennes.

Outre la complexification des modèles de prédiction épidémiologiques, l'emploi d'un unique seuil mène nécessairement à un compromis difficile entre sensibilité et spécificité de la prédiction. D'un côté si la prédiction d'appartenance à un foyer bactérien est trop sensible mais trop peu spécifique, le déclenchement des enquêtes épidémiologiques menant à infirmer le caractère épidémique d'un évènement est trop fréquent, ce qui implique un cout important en termes de ressources, de fonctionnement et de budget. D'un autre côté si la prédiction d'appartenance à un foyer est peu sensible, alors des foyers bactériens ne sont pas détectés, avec des conséquences graves en termes de santé, par exemple celle de patients ou de consommateurs.

EXPOSE DE L'INVENTION

Le but de la présente invention est de proposer un procédé d'identification et de surveillance d'un foyer bactérien à base de comparaison de génomes bactériens qui offre une liberté en

termes de sensibilité et de spécificité tout en tenant compte explicitement des sources d'incertitude dans la prédiction d'appartenance de souches bactériennes au foyer bactérien.

A cet effet, l'invention a pour objet un procédé de détection et de surveillance d'un foyer bactérien lié à une espèce bactérienne au sein d'une zone géographique, comprenant :

- l'obtention d'un génome numérique d'une souche bactérienne prélevée au sein de la zone géographique et appartenant à l'espèce bactérienne ;
- le calcul d'une distance génomique du génome numérique obtenu avec un génome numérique d'une base de données, dite « épidémiologique », comprenant au moins un génome numérique d'une souche bactérienne appartenant à l'espèce bactérienne;
- la prédiction :
  - que la souche bactérienne prélevée et la souche bactérienne de la base de données appartiennent au foyer bactérien si leur distance génomique est inférieure à un premier seuil prédéterminé ; ou
  - que la souche bactérienne prélevée et la souche bactérienne de la base de données n'appartiennent pas au foyer bactérien si leur distance génomique est supérieure à un second seuil prédéterminé strictement supérieur au premier seuil ; ou
  - que la souche bactérienne prélevée et la souche bactérienne de la base de données appartiennent peut être au foyer bactérien si leur distance génomique est comprise entre le premier et le second seuil ;

procédé selon lequel :

- le premier seuil est supérieur ou égal à un troisième seuil tel qu'une prédiction d'appartenance au foyer bactérien de deux souches bactériennes ayant une distance génomique inférieure au troisième seuil a une spécificité maximale ; et
- le second seuil est inférieur ou égal à un quatrième seuil tel qu'une prédiction de non appartenance au foyer bactérien de deux souches bactériennes ayant une distance génomique supérieure au quatrième seuil a une sensibilité maximale.

En d'autres termes, deux seuils différents sont utilisés pour régler la sensibilité et la spécificité du procédé, le seuil le plus bas étant utilisé pour régler la spécificité de la prédiction d'appartenance d'une souche au foyer bactérien (ci-après « spécificité d'appartenance ») et le seuil le plus haut étant utilisé pour régler la sensibilité de cette prédiction (ci-après « sensibilité d'appartenance »). La zone comprise entre ces deux seuils est ainsi spécifiquement prévue pour tenir compte des incertitudes inhérentes à une prédiction basée sur des distances génomiques. En particulier les troisième et quatrième seuils, préalablement appris pour maximiser la spécificité et la sensibilité d'appartenance, définissent une zone où il est difficile de savoir si des souches appartiennent ou non à un même foyer en raison de données incomplètes ou trop peu diversifiées pour apprendre ces seuils, de méconnaissance des mécanismes de mutation qui

sont hétérogènes au sein de l'espèce bactérienne, d'une imprécision du procédé en raison du choix de la méthode de comparaison génomique ou encore des erreurs de caractérisation des foyers infectieux résultant des enquêtes épidémiologiques. Cette zone d'incertitude offre à l'utilisateur une souplesse dans la gestion des épidémies. En particulier, contrairement à la prédiction d'appartenance au foyer bactérien qui déclenche une enquête épidémiologique et des mesures prophylactiques pour endiguer le foyer bactérien, lorsqu'une souche est dans la zone intermédiaire, l'utilisateur peut mettre en place une enquête préliminaire, par exemple en recoupant avec le dossier du patient sur lequel le prélèvement a été réalisé ou en analysant son résistome, son virulome ou sa position phylogénique dans la biodiversité de l'espèce, pour décider si oui ou non une enquête épidémiologique poussée doit être mise en œuvre. Par ailleurs, la zone comprise entre les troisième et quatrième seuils peut dans certain cas être trop importante de sorte que la prédiction basée sur ces seuils n'est pas optimale. Les premier et second seuils, définissant une zone strictement comprise entre les troisième et quatrième seuils, autorisent une optimisation analytique de la prédiction d'appartenance ou de non appartenance au foyer bactérien.

Selon un mode de réalisation, le premier et le second seuils sont égaux à deux distances génomiques calculées :

- en constituant une base de données d'apprentissage de génomes numériques de souches bactériennes appartenant à l'espèce bactérienne, ladite base comprenant :
  - des couples de souches bactériennes préalablement déterminées comme appartenant à un même foyer bactérien, et étiquetés comme « couples de souches reliées » ;
  - des couples de souches bactériennes préalablement déterminée comme n'appartenant pas à un même foyer bactérien, et étiquetés comme « couples de souches non reliées » ;
- en choisissant un prédicteur binaire configuré pour prédire que deux souches bactériennes sont reliées ou non reliées par comparaison de leur distance génomique à un cinquième seuil ;
- pour chaque valeur de cinquième seuil appartenant à un ensemble prédéterminé de valeurs de cinquième seuil, en calculant
  - une matrice de confusion dudit prédicteur en fonction de la base de données d'apprentissage ;
  - un premier index de qualité du prédicteur en fonction de la matrice de confusion, ledit premier index étant différent de la sensibilité et de la spécificité du prédicteur ;
  - un second index de qualité, différent du premier index, en fonction de la matrice de confusion, ledit second index étant différent du premier index, de la sensibilité et de la spécificité du prédicteur ;

- en recherchant une première valeur de cinquième seuil qui optimise le premier index et une seconde valeur de cinquième seuil qui optimise le second index ;
- en posant le premier seuil comme égal au minimum de la première et de la seconde valeurs de cinquième seuil et en posant le second seuil comme égal au maximum de la première et de la seconde valeurs de cinquième seuil.

En d'autres termes, une prédiction basée sur une spécificité et une spécificité d'appartenance maximales ne constituent pas nécessairement une prédiction optimale au regard des données épidémiologiques disponibles, stockées dans la base de données d'apprentissage. En calculant des premier et second seuils qui optimisent la qualité de la prédiction binaire, il est obtenu de fait une optimisation de la gestion des évènements épidémiques, tout en conservant une zone intermédiaire suffisamment large pour continuer d'alerter l'utilisateur d'un possible foyer bactérien.

Selon un mode de réalisation, le premier index est choisi pour tenir compte du déséquilibre, dans la base de données d'apprentissage, entre le nombre de couples de souches reliées et le nombre de couples de souches reliées. En particulier, le premier index est le coefficient de corrélation de Matthews ou le F1 score. D'une manière générale, les données concernant les foyers bactériens, c'est-à-dire le nombre de souches considérées comme reliées, sont bien moins nombreuses que les souches considérées comme non reliées. En utilisant un index de qualité qui prend explicitement en compte ce déséquilibre, une meilleure optimisation de la prédiction est obtenue. En outre, le seuil correspondant au coefficient de Matthews ou du F1-score privilégie la spécificité sans pour autant ne prendre que la spécificité en compte.

Selon un mode de réalisation, le second index est l'index de Youden. Cet index, qui prend explicitement en compte la spécificité et la sensibilité permet naturellement d'optimiser la prédiction de non appartenance dont l'apprentissage est réalisée usuellement sur une donnée importante. Le déséquilibre de la base de données a pour effet que l'index de Youden est plus influencé par la sensibilité, la spécificité étant proche de 1 sur tout l'intervalle compris entre les troisième et quatrième seuils.

Selon un mode de réalisation, le prédicteur est choisi de sorte que :

- les vrais positifs correspondent aux couples de souches reliées ayant une distance génomique inférieure au cinquième seuil :
- les faux négatifs correspondent aux couples de souches reliées ayant une distance génomique supérieure au cinquième seuil ;
- la faux positifs correspondent aux couples de souches non reliées ayant une distance génomique inférieure au cinquième seuil ; et

- les vrais négatifs correspondent aux couples de souches non reliées ayant une distance génomique supérieure au cinquième seuil.

Selon un mode de réalisation, la base de données épidémiologique comprend la base de données d'apprentissage. En d'autres termes, la base de données d'apprentissage est complétée à mesure que le procédé est mise en œuvre, ce qui permet de raffiner les différents seuils à mesure que la base augmente.

Selon un mode de réalisation, la distance génomique est une distance normalisée. Plus particulièrement, la distance génomique entre deux souches bactériennes est calculée en :

- sélectionnant, dans un ensemble prédominé de loci, les loci communs aux génomes numériques desdites souches ;
- comptant le nombre de différences alléliques, aux loci communs, entre les deux génomes numériques desdites souches ;
- en divisant ledit nombre de différences par le nombre de loci communs.

En normalisant par le nombre de loci en communs, l'impact des erreurs de séquençage, en particulier le fait de ne pas identifier un loccus chez une souche bactérienne, est atténué.

Selon un mode de réalisation privilégié, si les premières et secondes valeurs de cinquième seuil sont supérieures à 0,1, alors :

- le second seuil est posé égal à 0,1 ;
- le premier seuil est posé égal à $\max(D_g \backslash D_g < 0{,}2)$, où $\max(D_g \backslash D_g < 0{,}2)$ est la plus grande distance génomique, parmi les couples de souches reliées, strictement inférieure à 0,2.

En particulier, les inventeurs ont constaté que des valeurs supérieures à 0,1, obtenues usuellement en raison d'une base de données d'apprentissage incomplète ou trop peu diverse, matérialisent un échec de l'apprentissage. Les inventeurs ont de plus noté que dans le cadre d'une base de données d'apprentissage appropriée, les premier et second seuils sont inférieurs ou égaux à 0,1. Un des deux seuils est ainsi fixé à cette borne supérieure. Par ailleurs, les inventeurs ont constaté que deux souches de même sous-type ont en très grande majorité une distance génomique inférieure à 0,2. Ainsi en posant l'autre seuil égal à $\max(d_r \backslash d_r < 0{,}2)$, deux souches de distance génomique supérieure à ce dernier, il est prédit que ces souches n'appartiennent pas au même sous-types bactérien, et donc n'appartiennent pas au même foyer, ce qui constitue un indice important pour la suspicion d'épidémie. Ainsi, alors même que les données sont encore insuffisantes pour calculer avec précision les premier et second seuils, l'utilisateur dispose d'un procédé par défaut.

Selon un mode de réalisation, les distances entre les génomes numériques sont calculées en fonction d'un base de marqueurs, en particulier une base wgMLST, cgMLST, MLST, de gènes ou de SNP.

Selon un mode de réalisation, lorsqu'une souche prélevée est prédite comme appartenant au foyer bactérien, elle est étiquetée dans la base de donnée épidémiologique comme étant « reliée » avec les souches bactériennes du foyer bactérien et comme étant « non reliée » avec les autres souches bactériennes.

Selon un mode de réalisation, lorsqu'une souche prélevée est prédite comme appartenant peut-être au foyer bactérien, une caractérisation supplémentaire de ladite souche est mise en œuvre pour déterminer si elle appartient effectivement audit foyer, et si tel est le cas la souche bactérienne prélevée est étiquetée, dans la base de donnée épidémiologique, comme étant « reliée » avec les souches bactériennes du foyer bactérien et comme étant « non reliée » avec les autres souches bactériennes.

Selon un mode de réalisation, le premier et le second seuil sont recalculés régulièrement et/ou dès que N nouvelles souches sont ajoutées à la base de données épidémiologique, où N est un entier supérieur ou égal à 1.

Selon un mode de réalisation, lorsqu'une souche est prédite comme appartenant au foyer bactérien, des mesures prophylactiques sont mises en œuvre pour enrailler ledit foyer.

BREVE DESCRIPTION DES FIGURES

L'invention sera mieux comprise à la lecture de la description qui va suivre, donnée uniquement à titre d'exemple, et faite en relation avec les dessins annexés, dans lesquels des références identiques désignent des éléments identiques, et dans lesquels :

- la figure 1 est un organigramme d'un mode de réalisation du procédé selon l'invention ;
- la figure 2 illustre une table de correspondance entre souches bactériennes mémorisées dans une base de données d'apprentissage ;
- la figure 3 est une matrice de confusion d'un prédicteur binaire prédisant l'état reliées ou non reliées de deux souches bactériennes ;
- la figure 4 illustre une distribution du nombre de couples de souches reliées et une distribution du nombre de couples de souches non reliées en fonction de leur distance génomique ainsi qu'un seuil Ti utilisé pour calculer la matrice de confusion de la figure 3 ;

- la figure 5 est un tracé illustrant différents seuils sur les distances génomiques utilisés par le procédé selon l'invention ;
- la figure 6 illustre un système informatique et de séquençage pour la mise en œuvre du procédé selon l'invention ;
- les figures 7A et 7B sont des distributions du nombre de couples de souches non reliées (distribution supérieure) et du nombre de couples de souches reliées (distribution inférieure) pour l'espèce bactérienne *Clostridium difficile*, la figure 7B étant un agrandissement entre 0 et 0,1 de la figure 7A ;
- les figures 8A et 8B illustrent, pour l'espèce *Clostridium difficile,* les distances génomiques pour différentes valeurs optimales d'indice de qualité, dont la sensibilité, la spécificité, la précision, la justesse (« accuracy » en anglais, i.e. $(TP + TN)/(N + P)$), le F1 score, l'indice de Youden, et le coefficient de corrélation de Matthews, la figure 8B étant un agrandissement entre 0 et 0,1 de la figure 7B ;
- les figures 9A et 9B sont des distributions du nombre de couples de souches non reliées (distribution supérieure) et du nombre de couples de souches reliées (distribution inférieure) pour l'espèce bactérienne *Staphylococcus aureus*, la figure 9B étant un agrandissement entre 0 et 0,1 de la figure 9A ;
- les figures 10A et 10B illustrent, pour l'espèce *Staphylococcus aureus,* les distances génomiques pour différentes valeurs optimales d'indice de qualité, dont la sensibilité, la spécificité, la précision, la justesse, le F1 score, l'indice de Youden, et le coefficient de corrélation de Matthews, la figure 10B étant un agrandissement entre 0 et 0,1 de la figure 10B ;

DESCRIPTION DETAILLEE DE L'INVENTION

Dans ce qui suit « inférieur » signifie « inférieur ou égal » et même « supérieur» signifie « supérieur ou égal » sauf s'il est stipulé strictement.

Il va à présent être décrit un mode de réalisation de l'invention en relation avec la détection et le suivi de foyers infectieux microbiologiques d'une espèce bactérienne particulière dans un hôpital.

En se référant à la figure 1, ce procédé comporte une première étape **10** d'apprentissage d'au moins deux seuils, notés *S1* et *S2*, sur la base desquels des comparaisons de génomes sont réalisées pour déterminer si une souche bactérienne appartient à non à un foyer bactérien, et une seconde étape **20** de mise en œuvre du procédé selon l'invention, paramétré avec les seuils appris lors de l'étape **10**. Plus particulièrement, le procédé se fonde sur la comparaison d'une distance génomique, notée $D_g(BSi, BSj)$ entre deux souches, notées *BSi* et *BSj*,

L'étape **10** débute par la constitution, en **12**, d'une base de données d'apprentissage pour l'espèce considérée comprenant :

- des génomes numériques de différentes souches BS1, BS2, BS3…., BSN appartenant à l'espèce ;
- une table de correspondance, illustrée à la figure 2, reliant chaque souche de la base de données à l'ensemble des autres souches, chaque lien entre deux souches de la base pouvant prendre un état « reliées » (cases noires) lorsque les deux souches ont été préalablement déterminées comme appartenant à un même foyer bactérien, et un état « non reliées » (cases blanches) lorsque les deux souches ont été préalablement déterminées comme n'appartenant pas à un même foyer bactérien, l'état du lien entre deux souches étant par exemple déterminé lors d'une étude épidémiologique antérieure. En outre, le lien d'une souche par rapport à elle-même est fixée à l'état « reliées ». Comme visible à la figure 2, plusieurs foyers infectieux pour l'espèce considérée peuvent être pris en compte pour déterminer les états « reliées » et « non reliées » des souches de la base de données d'apprentissage. Comme il sera décrit par la suite, la base de données d'apprentissage peut également contenir des souches déterminées comme étant « reliées » sans pour autant avoir été diagnostiquée comme appartenant à un quelconque foyer bactérien. De préférence, ladite table mémorise également les distances génomiques $D_g(BSi, BSj)$ entre chaque paire de souches *BSi* et *BSj* de la base de données d'apprentissage ;
- une table qui répertorie l'ensemble des foyers infectieux identifiés avec leurs souches associées ;
- les résistomes (ensemble de marqueurs génétiques contribuant à la sensibilité ou la résistance aux antibiotiques d'une bactérie) et les virulomes (ensemble de marqueurs génétiques contribuant à la virulence d'une bactérie) des souches BS1, BS2, BS3…., BSN ;

Le génome d'une souche bactérienne est de préférence obtenu par :

- le prélèvement d'un échantillon chez un patient comprenant la souche;
- la préparation d'un isolat de la souche, par exemple en étalant l'échantillon sur un milieu de culture gélosé et en réalisant une incubation de manière à faire pousser une colonie de la souche bactérienne ;
- le prélèvement d'une partie de la colonie et la préparation de la quantité prélevée pour un séquençage (e.g. une lyse pour libérer l'ADN des bactéries, le cas échéant amplification de l'ADN libéré et la préparation d'une librairie pour les techniques de séquençage le nécessitant) ;

- le séquençage, de préférence complet (ou séquençage WGS), de l'ADN de manière à produire des séquences numériques, communément appelées « read », par exemple à l'aide d'une technologie de type « next generation sequencing » telle qu'avec la plateforme de séquençage « MiSeq » de la société Illumina Inc., San Diego, Californie;
- optionnellement, l'assemblage des reads de manière à produire des séquences assemblées, connues sous le terme de « contig » ;
- la caractérisation selon la technique wgMLST (pour « whole genome multilocus sequensing typing ») du génome sous forme de contig ou de reads, communément appelé « profil wgMLST ». Comme cela est connu en soi, cette caractérisation consiste à localiser des loci dans le génome parmi un ensemble prédéterminé de loci, et pour chaque locus identifié, à déterminer l'allèle qui représente ce locus. La technique wgMLST est par exemple décrite dans le document « *MLST revisited: the gene-by-gene approach to bacterial genomics*» de Martin C.J. Maiden, Nature Reviews Microbiology, 2013.

L'apprentissage se poursuit par le calcul de seuils *S1* et *S2* en fonction de la base de données d'apprentissage. Plus particulièrement, ce calcul consiste à transformer :

- un premier prédicteur f_T d'appartenance ou non de deux souches à un foyer bactérien sur la base d'un unique seuil T sur les distances génomiques $D_g(BSi, BSj)$ divisant l'espace des distances génomiques en deux intervalles uniquement:

$$f_T\left(D_g(BSi, BSj)\right) = \begin{cases} 1 \text{ si les souches } BSi \text{ et } BSj \text{ sont reliées} \\ -1 \textit{ si les souches } BSi \text{ et } BSj \text{ sont non reliées} \end{cases}$$

- en un second prédicteur $g_{S1,S2}$ d'appartenance ou non de deux souches à un foyer bactérien sur la base de deux seuils *S1* et S2 sur les distances génomiques $D_g(BSi, BSj)$ divisant l'espace des distances génomiques en trois intervalles :

$$g_{S1,S2}\left(D_g(BSi, BSj)\right) = \begin{cases} 1 \text{ si les souches } BSi \text{ et } BSj \text{ sont reliées} \\ 0 \text{ si les touches } BSi \text{ et } BSj \text{ sont potentiellement reliées} \\ -1 \textit{ si les souches } BSi \text{ et } BSj \text{ sont non reliées} \end{cases}$$

Dans une variante privilégiée, le premier prédicteur f_T est défini tel que :

$$\begin{cases} \text{si } D_g(BSi, BSj) \leq T \textit{ alors } f_T = 1 \\ \text{si } D_g(BSi, BSj) > T \textit{ alors } f_T = -1 \end{cases}$$

et le second prédicteur est défini tel que :

$$\begin{cases} \text{si } D_g(BSi, BSj) \leq S1 \; alors \; g_{S1,S2} = 1 \\ \text{si } S1 < D_g(BSi, BSj) \leq S2 \; alors \; g_{S1,S2} = 0 \\ \text{si } D_g(BSi, BSj) > S2 \; alors \; g_{S1,S2} = -1 \end{cases}$$

De préférence, la distance génomique $D_g(BSi, BSj)$ est une distance normalisée, et donc comprise entre 0 et 1, calculée en :

a. identifiant dans les profils wgMLST des deux souches *BSi* et *BSj* les loci qu'elles ont en commun ;
b. pour chaque locus commun, en déterminant s'il existe une différence allélique entre les deux souches, et dans ce cas en incrémentant de 1 un compteur *Compt* de différences alléliques si au moins une différence allélique est constatée;
c. en calculant $D_g(BSi, BSj)$ selon la formule suivante, avec N_{lc} est le nombre de loci en commun:

$$D_g(BSi, BSj) = \frac{Compt}{N_{lc}}$$

Le calcul des seuils *S1* et *S2* débute, en **14**, par le calcul d'une matrice de confusion $MC(Ti)$ du prédicteur binaire f_T pour chacune des valeurs Ti d'un ensemble $\{T1, T2, \ldots, TM\}$ de valeurs de seuils T comprises entre 0 et 1, par exemple avec un incrément de 10^{-4}. Le calcul de la matrice de confusion $MC(Ti)$, illustrée à la figure 3, pour le seuil Ti est illustrée à la figure 4 et consiste à compter :

- les vrais positifs, notés « TPis », égal au nombre total de couples de souches reliées de la base tel que $D_g(BSi, BSj) \leq Ti$;
- les faux négatifs, notés « FNi », égal au nombre total de couples de souches reliées de la base tel que $D_g(BSi, BSj) > Ti$;
- la faux positifs, notés « FPi », égal au nombre total de couples de souches non reliées de la base tel que $D_g(BSi, BSj) \leq Ti$; et
- les vrais négatifs, notés « TNi », égal au nombre total de couples de souches non reliées de la base tel que $D_g(BSi, BSj) \leq Ti$.

Une fois l'ensemble des matrices de confusion $\{MC(T1), MC(T2, \ldots, MC(TM))\}$ calculé, le procédé se poursuit, en **16**, par le calcul de différents seuils illustrés à la figure 5:

- un seuil *S3* tel que la spécificité du prédicteur f_T est maximale, et donc tel que la spécificité de la prédiction que deux souches sont reliées est maximale, c'est-à-dire $S3 = arg \max_{Ti} \left(\frac{TNi}{N}\right)$, où N est le nombre de couples de souches non reliées ;
- un seuil *S4* tel que la sensibilité du prédicteur f_T est maximale, et donc tel que la sensibilité de la prédiction que deux souches sont non reliées est maximale, c'est-à-dire $S4 = arg \max_{Ti} \left(\frac{TPi}{P}\right)$, où P est le nombre de couples de souches reliées;

- le seuil *S1* optimisant un premier index de qualité du prédicteur f_T, différent de la sensibilité et de la spécificité et tenant compte explicitement du déséquilibre entre les nombres *P* et *N*, préférentiellement le coefficient de corrélation de Matthews (« MCC »), c'est-à-dire $S1 = arg \max_{Ti}\left(\frac{TPi.TNi-FPi.FNi}{\sqrt{(TPi+FPi).(TPi+FNi).(TNi+FPi).(TNi+FNi)}}\right)$;
- le seuil *S2* optimisant un second index de qualité du prédicteur f_T, différent de la sensibilité et de la spécificité, préférentiellement l'index de Youden, c'est-à-dire $S2 = arg \max_{Ti}\left(\frac{TPi}{P}+\frac{TNi}{N}-1\right)$.

Une étape **18** de contrôle de la qualité des seuils *S1* et *S2* est ensuite mise en œuvre. Plus particulièrement (le signe « \ » signifiant « tel que »):

- si les seuils *S1* et *S2* sont inférieurs ou égaux à 0,1, ils sont conservés, signifiant que la base de données d'apprentissage est appropriée pour leur calcul et leur usage ultérieur ;
- si les seuils *S1* et *S2* sont supérieurs à 0,1 ou diffèrent de moins de 1% , alors leurs valeurs sont fixées à 0,1 et $M = \max(D_g(BSi, BSj)\backslash D_g(BSi, BSj) < 0{,}2)$, où $\max(D_g(BSi, BSj)\backslash D_g(BSi, BSj) < 0{,}2)$ est ici la distance génomique maximale qui est la plus proche de 0,2 parmi les couples de souches reliées de la base de données d'apprentissage;
- si l'un des seuils *S1* ou *S2* est supérieur à 0,1, ce seuil est alors fixé à au minimum des valeur 0,1 et $\max(D_g(BSi, BSj)\backslash D_g(BSi, BSj) < 0{,}2)$ si cette valeur minimale est différente de l'autre seuil (e.g. diffère de plus de 1%), sinon ce seuil est fixé au maximum de ces deux valeurs.

Par soucis de simplification, il sera supposé par la suite que le seuil *S1* est inférieur au seuil *S2*, de sorte que, comme illustré à la figure 4, ces seuils divisent l'espace des distances génomiques en trois intervalles :

- un intervalle inférieur $]0, S1]$. Si la distance génomique entre deux souches est comprise dans cet intervalle, ces souches sont prédites comme étant « reliées » ($g_{S1,S2} = 1$);
- un intervalle supérieur $]S2, 1]$. Si la distance génomique entre deux souches est comprise dans cet intervalle, ces souches sont prédites comme étant « non reliées » ($g_{S1,S2} = -1$) ; et
- un intervalle intermédiaire $]S1, S2]$. Si la distance génomique entre deux souches est comprise dans cette intervalle, ces souches sont prédites comme « potentiellement reliées » ($g_{S1,S2} = 0$) .

Les seuils *S1* et *S2* sont alors mémorisés dans une mémoire informatique d'un système informatique utilisé pour mettre en œuvre l'étape **20** à présent décrite, ledit système comprenant en outre la base de données d'apprentissage. L'étape **20**, qui prend place au sein de l'hôpital pour détecter et surveiller les épidémies de nature bactérienne, est par exemple mise en œuvre

de manière systématique dès qu'un patient est atteint d'une infection bactérienne, qu'un échantillon environnemental comprend une bactérie pathogène ou qu'un patient présente des symptômes identiques ou similaires à un autre patient au sein de l'hôpital. D'autres critères peuvent bien entendu être employés pour lancer cette étape.

L'étape **20** débute, en **22**, par le prélèvement d'un échantillon contenant la souche pathogène, si ce prélèvement n'a pas encore eu lieu, puis se poursuit, en **24**, par le séquençage de la souche et l'établissement de son profil wgMLST tel que décrit en relation avec l'étape **12**. En **26**, la distance génomique $D_g(BSi, BSj)$ entre la souche prélevée et chacune des souches de la base de données d'apprentissage est ensuite calculée. Un premier diagnostic épidémiologique est alors émis en **28**. Plus particulièrement :

- si la souche prélevée n'est reliée à aucune souche de la base de données, c'est-à-dire que quel que soit la souche de la base de données la distance génomique $D_g(BSi, BSj)$ avec la souche prélevée est supérieure au seuil S2, alors il est déterminé que la souche prélevée n'appartient à aucun foyer bactérien ;
- si la souche prélevée est reliée à une souche de la base de données, c'est-à-dire que ces deux souches ont une distance génomique $D_g(BSi, BSj)$ inférieure ou égale au seuil S1, une alarme est déclenchée à l'attention de l'utilisateur et une étude épidémiologique approfondie **30** est lancée ainsi que, le cas échéant, des mesures prophylactiques pour lutter contre la transmission de la souche prélevée au sein de l'hôpital ;
- si la souche prélevée est potentiellement reliée à une souche de la base de données (noté « reliées ? » dans la figure 1), c'est-à-dire que ces deux souches ont une distance génomique $D_g(BSi, BSj)$ comprise entre les seuils S1 et S2, une analyse complémentaire est menée, en **34**, pour lever l'incertitude sur le lien entre ces deux souches. De manière préférentielle, le résistome et le virulome de la souche prélevée est déterminé puis comparé au résistome et le virulome de la souche à laquelle elle est potentiellement reliée. Si les résistomes et les virulomes sont concordants, les souches sont alors déterminées comme étant reliées, l'alarme est déclenchée et l'étude approfondie **30** menée. Dans le cas contraire, les souches sont déterminées comme étant non reliées. Enfin dans le cas où cette comparaison ne permet pas de trancher, l'étude approfondie **30** est menée. D'autres informations peuvent être utilisées lors de cette étude complémentaire, comme par exemple la durée écoulée entre le prélèvement et celui de la souche de la base de données, le nombre de SNP différents dans les gènes plastiques, etc.

Comme cela est connu en soi, un des objectifs de l'étude **30**, menée par l'équipe d'épidémiologie de l'hôpital, est de déterminer si différentes souches prélevées au sein de l'hôpital constituent une épidémie. A l'issue de cette étude, le lien entre différentes souches est

définitivement établi, à savoir « reliées » ou « non reliées ». Si par ailleurs une épidémie est détectée alors les souches de l'épidémie sont également étiquetées en fonction de cette épidémie. Le génome, le profils wgMLST, le resistome et le virulome de la souche prélevée, ses liens avec les autres souches de la base de données ainsi que les informations concernant le foyer bactérien sont ensuite mémorisés dans la base de données d'apprentissage pour pouvoir être utilisés ultérieurement. Les seuils S1 et S2 peuvent être ainsi actualisés régulièrement ou à chaque nouvelle entrée dans la base de données afin d'affiner leurs valeurs.

La figure 6 illustre un système informatique et de séquençage **40** pour la mise en œuvre du procédé selon l'invention. Le système **40** comporte une plateforme de séquençage **42** pour séquencer l'ADN bactérien d'un échantillon **44** et ainsi produire un ensemble de séquence numériques, ou « reads ». La plateforme **42** est connectée à une unité de traitement d'information **46**, par exemple un ordinateur personnel, qui reçoit les séquences, et met optionnellement un programme d'assemblage des reads pour produire des contigs. L'unité **46** est par ailleurs connectée à un serveur distant **48** mettant en œuvre un logiciel en tant que service (ou « Saas »), par exemple sous la forme d'une solution cloud. L'unité **46**, sur lequel tourne un logiciel « front end », envoie au serveur **48** les génomes séquencés par la plateforme **42** sous forme de reads ou de contigs. Le serveur **48**, sur lequel tourne le service informatique sous forme de « back end » et qui est connecté à la base de données d'apprentissage **50**, reçoit les génomes et met en œuvre les étapes de traitement du procédé selon l'invention (e.g. les étapes 14-18 et 24-32 de la figure 1), le serveur mémorisant dans une mémoire informatique l'ensemble des instructions nécessaires à cette mise en œuvre. Le serveur renvoie les résultats du traitement à l'unité **46** sous la forme d'un rapport **52**. Le système **40** comporte également un ou plusieurs serveurs **54** connecté(s) à l'unité **42**, ces serveurs étant notamment ceux du système informatique mémorisant les données patients et épidémiologiques, ces données étant utilisées lors des études approfondies pour caractériser les foyers bactériens épidémiologiques.

Les figures 7 et 9 illustrent des distributions du nombre de couples de souches reliées et de souches non reliées respectivement pour l'espèce *Clostridium difficile* (figures 7A et 7B) et *Staphylococcus aureus* (figures 9A et 9B). Comme on peut le noter sur ces figures, il existe des couples de souches reliées dont la distance génomique est importante (par exemple au-delà de 0,6 pour *Clostridium difficile*) et des couples de souches non reliées dont la distance génomique est faible (par exemple en deçà de 0,2 pour *Staphylococcus aureus*). Il existe ainsi une zone dans laquelle une distance génomique pourrait coder à la fois pour l'état « reliées » ou l'état « non reliées » si un seul seuil était employé. Cette zone intermédiaire est naturellement présente et correspond par exemple à des souches appartenant à un même sous-type mais n'ayant pas été jugées comme appartenant à un même foyer bactérien. De même, on observe au travers des figures 8A-B et 10A-B, qu'en choisissant les seuils S3 (spécificité maximale,

noté « spécificité ») et S4 (sensibilité maximale, notée « sensibilité ») pour diviser en trois l'espace des distances génomiques, la zone intermédiaire est si importante que bon nombre de souches seraient jugées comme potentiellement reliées. En utilisant les seuils S1 (e.g. maximisant le coefficient de Matthews MMC) et S2 (e.g. optimiant l'indice de Youden) qui optimisent la qualité de la prédiction, on note que la zone intermédiaire est sensiblement réduite tout en conservant une très bonne sensibilité globale.

Il a été décrit une application à l'épidémiologique de bactéries pathogènes au sein d'un hôpital. Evidemment, l'invention n'est pas limitée à cette application et peut être employée dans le domaine du contrôle microbiologique industriel (par exemple le contrôle agro-alimentaire), environnemental et vétérinaire.

Il a été décrit l'emploi de profil wgMLST pour calculer les distances génomiques. D'autres profils peuvent être employés, comme par exemple des profils cgMLST (« core genome multilocus sequencing typing »), MLST, des ensembles de SNP ou de gènes.

Il a été décrit l'emploi de l'indice de Youden et du coefficient de corrélation de Matthews. D'autres indices de qualité peuvent être employés, comme par exemple le F1 score (i.e. $2TP/(2TP + FP + FN)$), le coefficient χ_1 , la justesse (ou « accuracy », i.e. $(TP + TN)/(N + P)$), la précision (i.e. $TP/(TP + FP)$). De préférence au moins 1 de ces indices tient compte du déséquilibre de la base de données.

Il a été décrit une base de données d'apprentissage utilisée également pour la comparaison avec des souches prélevées. En variante, une base de données séparée, ou « base de données épidémiologique », peut être employée pour traiter les souches prélevées. Une telle base est par exemple propre à un hôpital, une institution, une société ou autre, et la base de données d'apprentissage est alors utilisée uniquement pour établir la valeur des seuils.

REVENDICATIONS

1. Procédé de détection et de surveillance d'un foyer bactérien lié à une espèce bactérienne au sein d'une zone géographique, comprenant :

- l'obtention d'un génome numérique d'une souche bactérienne prélevée au sein de la zone géographique et appartenant à l'espèce bactérienne ;
- le calcul d'une distance génomique du génome numérique obtenu avec un génome numérique d'une base de données, dite « épidémiologique », comprenant au moins un génome numérique d'une souche bactérienne appartenant à l'espèce bactérienne;
- la prédiction :
    - que la souche bactérienne prélevée et la souche bactérienne de la base de données appartiennent au foyer bactérien si leur distance génomique est inférieure à un premier seuil prédéterminé ; ou
    - que la souche bactérienne prélevée et la souche bactérienne de la base de données n'appartiennent pas au foyer bactérien si leur distance génomique est supérieure à un second seuil prédéterminé strictement supérieur au premier seuil ; ou
    - que la souche bactérienne prélevée et la souche bactérienne de la base de données appartiennent peut être au foyer bactérien si leur distance génomique est comprise entre le premier et le second seuil ;

procédé selon lequel :

- le premier seuil est supérieur ou égal à un troisième seuil tel qu'une prédiction d'appartenance au foyer bactérien de deux souches bactériennes ayant une distance génomique inférieure au troisième seuil a une spécificité maximale ; et
- le second seuil est inférieur ou égal à un quatrième seuil tel qu'une prédiction de non appartenance au foyer bactérien de deux souches bactériennes ayant une distance génomique supérieure au quatrième seuil a une sensibilité maximale.

2. Procédé selon la revendication 1, dans lequel le premier et le second seuils sont égaux à deux distances génomiques calculées :

- en constituant une base de données d'apprentissage de génomes numériques de souches bactériennes appartenant à l'espèce bactérienne, ladite base comprenant :
    - des couples de souches bactériennes préalablement déterminées comme appartenant à un même foyer bactérien, et étiquetés comme « couples de souches reliées » ;
    - des couples de souches bactériennes préalablement déterminée comme n'appartenant pas à un même foyer bactérien, et étiquetés comme « couples de souches non reliées » ;

- en choisissant un prédicteur binaire configuré pour prédire que deux souches bactériennes sont reliées ou non reliées par comparaison de leur distance génomique à un cinquième seuil ;
- pour chaque valeur de cinquième seuil appartenant à un ensemble prédéterminé de valeurs de cinquième seuil, en calculant
  - une matrice de confusion dudit prédicteur en fonction de la base de données d'apprentissage ;
  - un premier index de qualité du prédicteur en fonction de la matrice de confusion, ledit premier index étant différent de la sensibilité et de la spécificité du prédicteur ;
  - un second index de qualité, différent du premier index, en fonction de la matrice de confusion, ledit second index étant différent du premier index, de la sensibilité et de la spécificité du prédicteur ;
- en recherchant une première valeur de cinquième seuil qui optimise le premier index et une seconde valeur de cinquième seuil qui optimise le second index ;
- en posant le premier seuil comme égal au minimum de la première et de la seconde valeurs de cinquième seuil et en posant le second seuil comme égal au maximum de la première et de la seconde valeurs de cinquième seuil.

3. Procédé selon la revendication 2, dans lequel le premier index est choisi pour tenir compte du déséquilibre, dans la base de données d'apprentissage, entre le nombre de couples de souches reliées et le nombre de couples de souches reliées.

4. Procédé selon la revendication 3, dans lequel le premier index est le coefficient de corrélation de Matthews ou le F1 score.

5. Procédé selon l'une des revendications 2 à 4, dans lequel le second index est l'index de Youden.

6. Procédé selon l'une des revendications 2 à 5, dans lequel le prédicteur est choisi de sorte que :

- les vrais positifs correspondent aux couples de souches reliées ayant une distance génomique inférieure au cinquième seuil :
- les faux négatifs correspondent aux couples de souches reliées ayant une distance génomique supérieure au cinquième seuil ;
- la faux positifs correspondent aux couples de souches non reliées ayant une distance génomique inférieure au cinquième seuil ; et

- les vrais négatifs correspondent aux couples de souches non reliées ayant une distance génomique supérieure au cinquième seuil.

7. Procédé selon l'une des revendications 2 à 6, dans lequel la base de données épidémiologique comprend la base de données d'apprentissage.

8. Procédé selon l'une quelconque des revendications précédentes, dans lequel la distance génomique est une distance normalisée.

9. Procédé selon la revendication 8, dans lequel la distance génomique entre deux souches bactériennes est calculée en :
- sélectionnant, dans un ensemble prédominé de loci, les loci communs aux génomes numériques desdites souches ;
- comptant le nombre de différences alléliques, aux loci communs, entre les deux génomes numériques desdites souches ;
- en divisant ledit nombre de différences par le nombre de loci communs.

10. Procédé selon la revendication 9 et l'une des revendications 4 ou 5, dans lequel si les premières et secondes valeurs de cinquième seuil sont supérieures à 0,1, alors :
- le second seuil est posé égal à 0,1 ;
- le premier seuil est posé égal à $\max(D_g \backslash D_g < 0{,}2)$, où $\max(D_g \backslash D_g < 0{,}2)$ est la plus grande distance génomique, parmi les couples de souches reliées, strictement inférieure à 0,2.

11. Procédé selon l'une quelconque des revendications précédentes, dans lequel les distances entre les génomes numériques sont calculées en fonction d'une base de marqueurs, en particulier une base wgMLST, cgMLST, MLST, de gènes ou de SNP.

12. Procédé selon l'une quelconque des revendications précédentes, dans lequel lorsqu'une souche prélevée est prédite comme appartenant au foyer bactérien, elle est étiquetée dans la base de donnée épidémiologique comme étant « reliée » avec les souches bactériennes du foyer bactérien et comme étant « non reliée » avec les autres souches bactériennes.

13. Procédé selon l'une quelconque des revendications précédentes, dans lequel lorsqu'une souche prélevée est prédite comme appartenant peut-être au foyer bactérien, une caractérisation supplémentaire de ladite souche est mise en œuvre pour déterminer si elle appartient effectivement audit foyer, et si tel est le cas la souche bactérienne prélevée est étiquetée, dans la base de donnée épidémiologique, comme étant « reliée » avec les

souches bactériennes du foyer bactérien et comme étant « non reliée » avec les autres souches bactériennes.

14. Procédé selon l'une quelconque des revendications précédentes, dans lequel le premier et le second seuil sont recalculés régulièrement et/ou dès que N nouvelles souches sont ajoutées à la base de données épidémiologique, où N est un entier supérieur ou égal à 1.

15. Procédé selon l'une quelconque des revendications précédentes, dans lequel lorsqu'une souche est prédite comme appartenant au foyer bactérien, des mesures prophylactiques sont mises en œuvre pour enrailler ledit foyer.

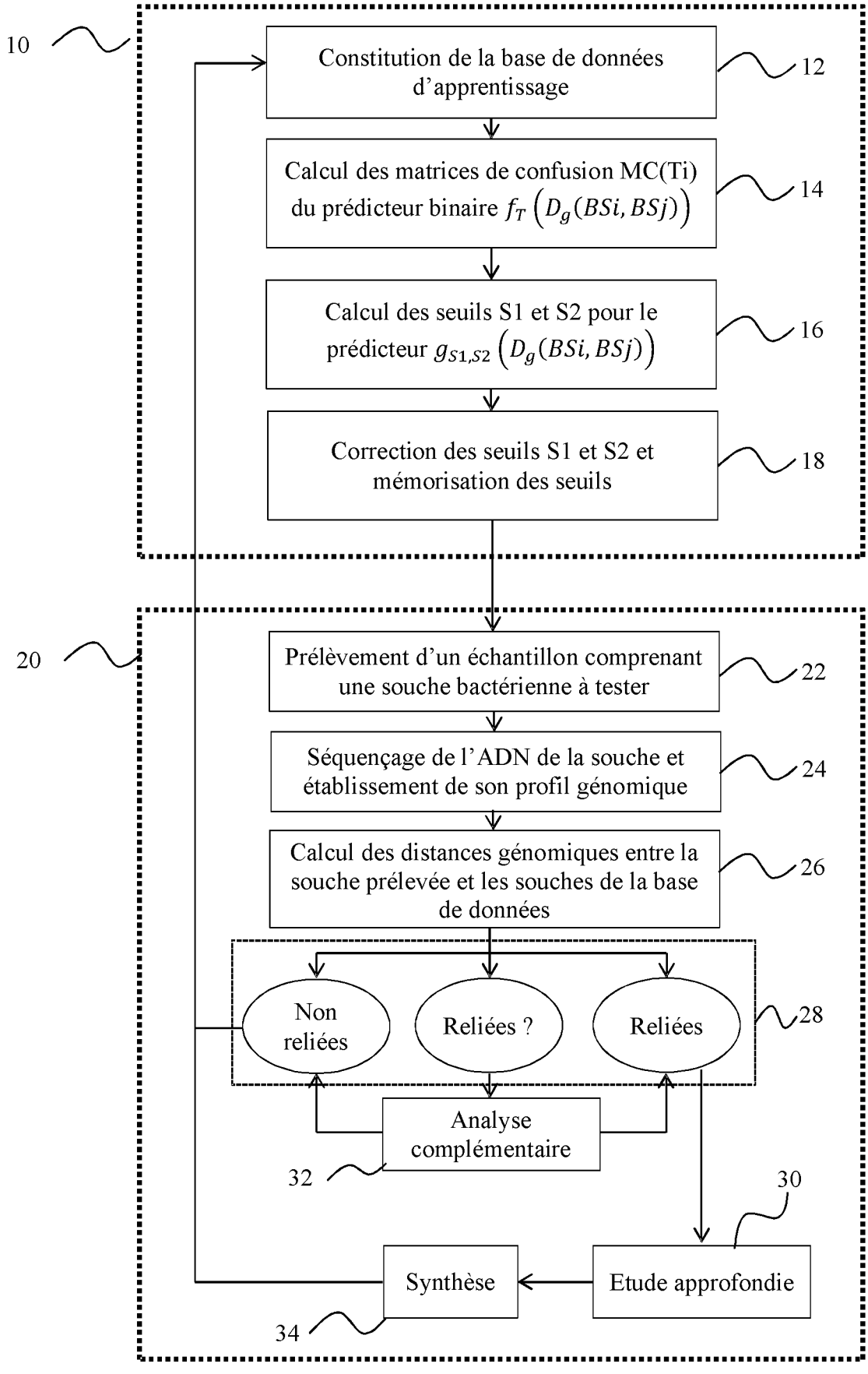

10
Constitution de la base de données d'apprentissage
12
Calcul des matrices de confusion MC(Ti) du prédicteur binaire $f_T\left(D_g(BSi, BSj)\right)$
14
Calcul des seuils S1 et S2 pour le prédicteur $g_{S1,S2}\left(D_g(BSi, BSj)\right)$
16
Correction des seuils S1 et S2 et mémorisation des seuils
18
20
Prélèvement d'un échantillon comprenant une souche bactérienne à tester
22
Séquençage de l'ADN de la souche et établissement de son profil génomique
24
Calcul des distances génomiques entre la souche prélevée et les souches de la base de données
26
Non reliées
Reliées ?
Reliées
28
Analyse complémentaire
32
30
Synthèse
Etude approfondie
34

Figure 1

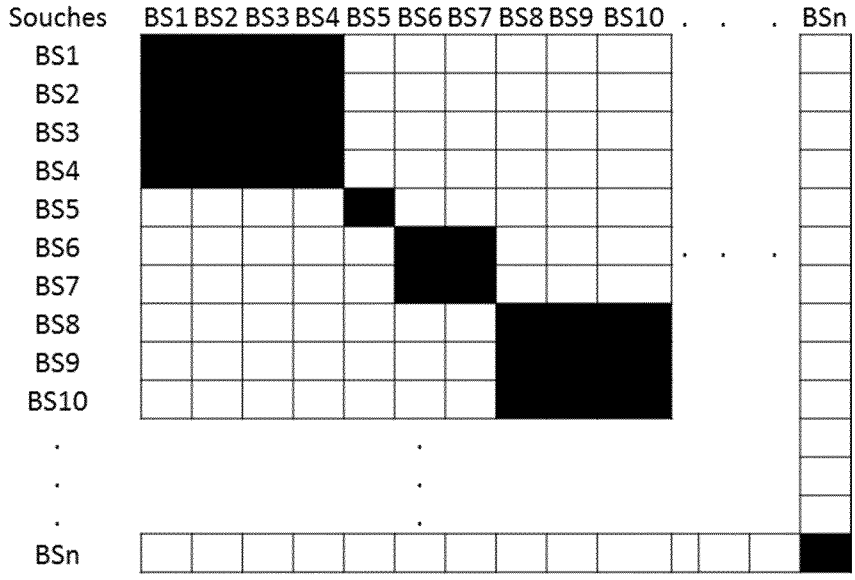


Figure 2

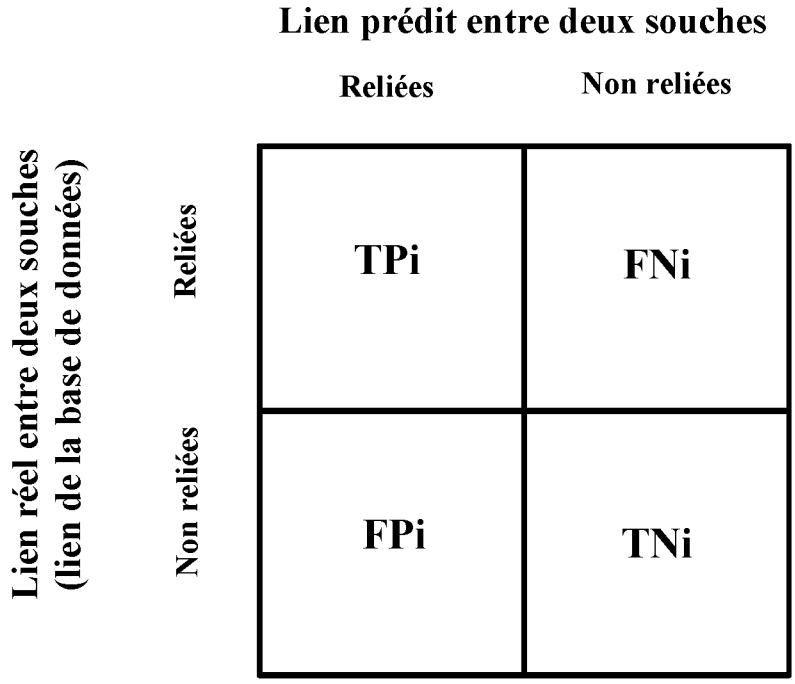


Figure 3

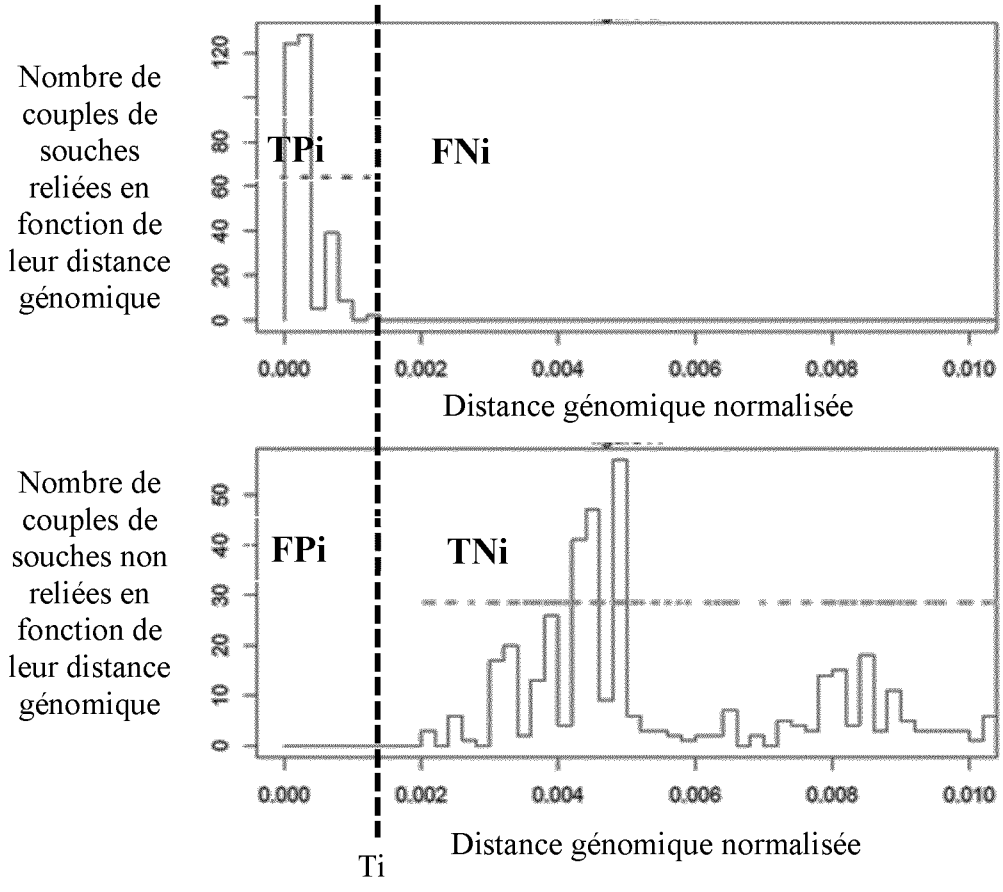


Figure 4

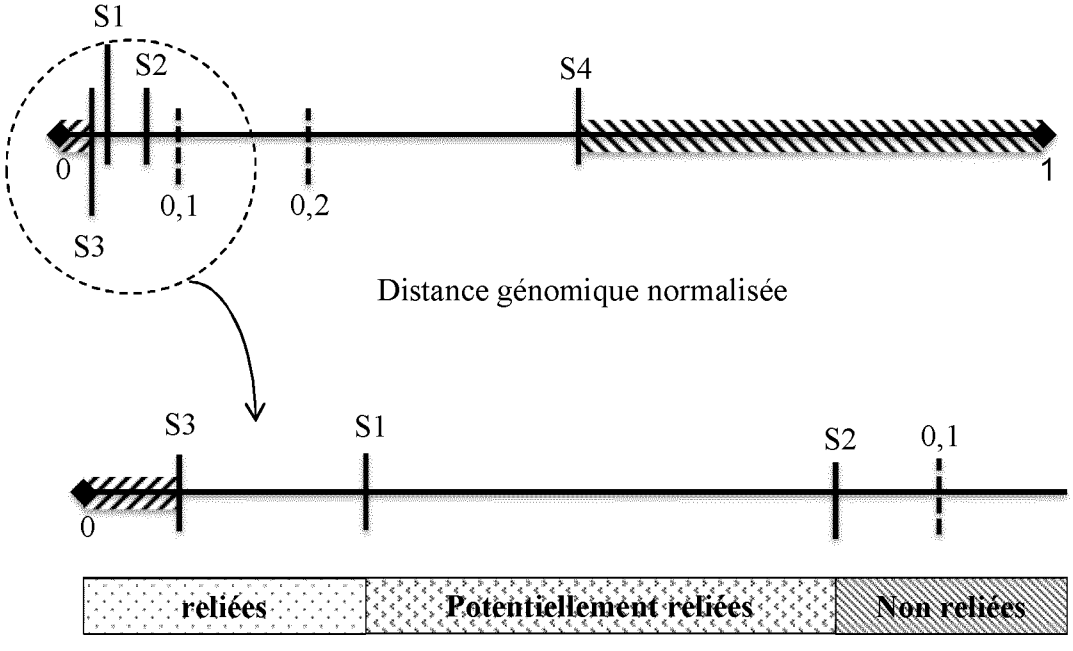


Figure 5

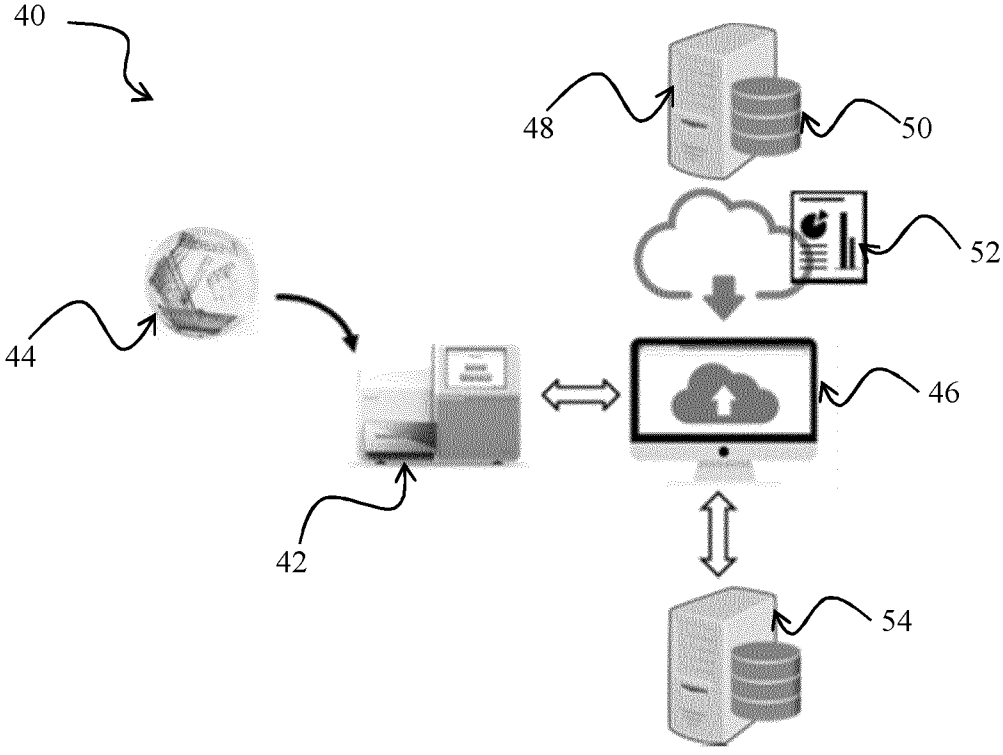


Figure 6

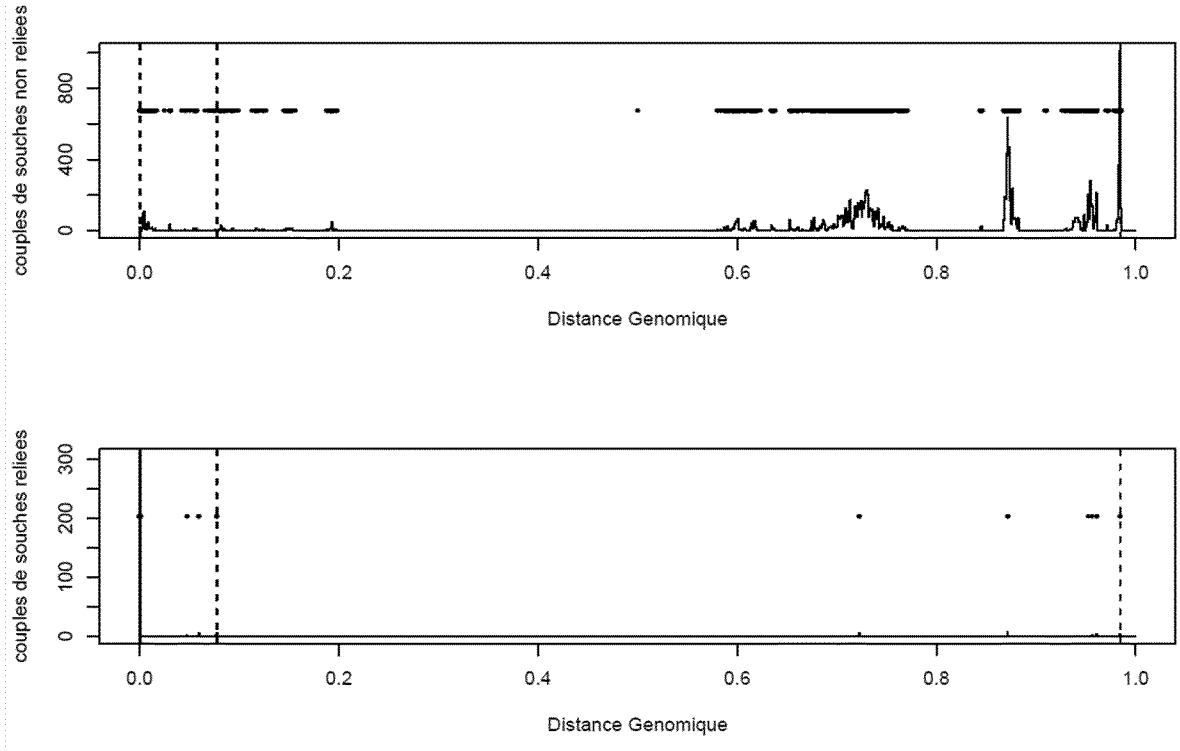


Figure 7A

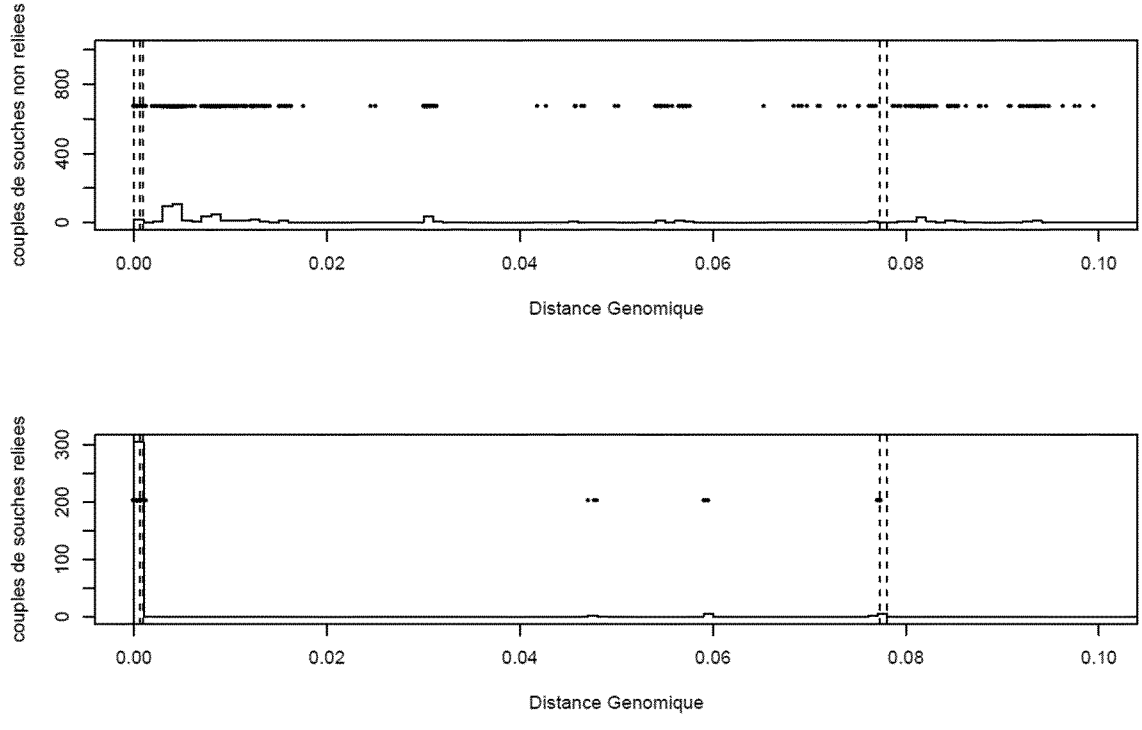


Figure 7B

Figure 8A

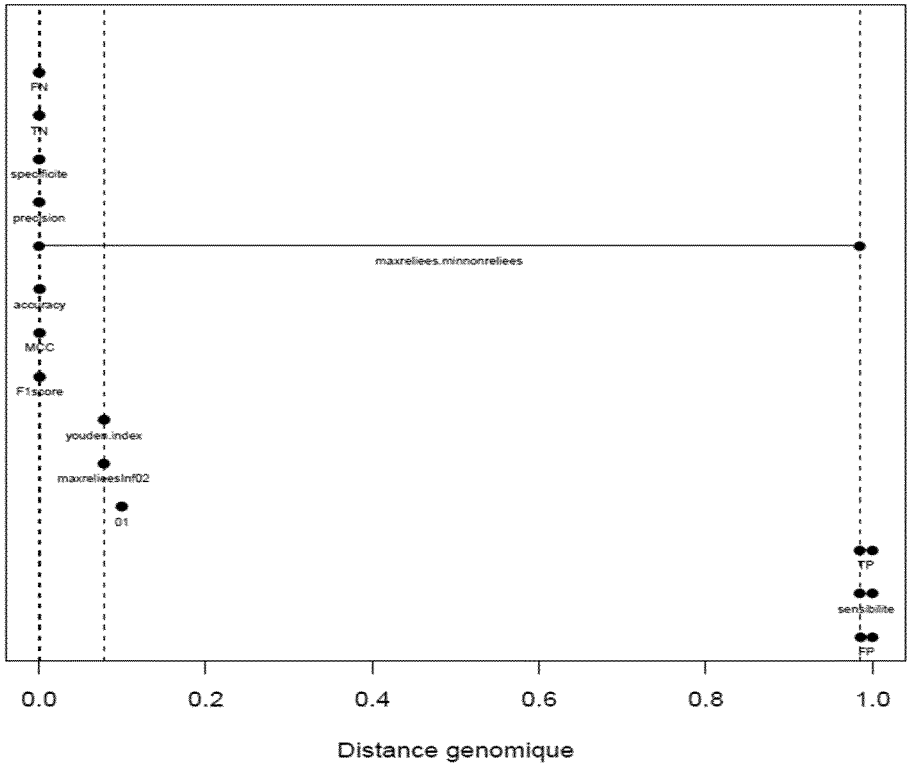
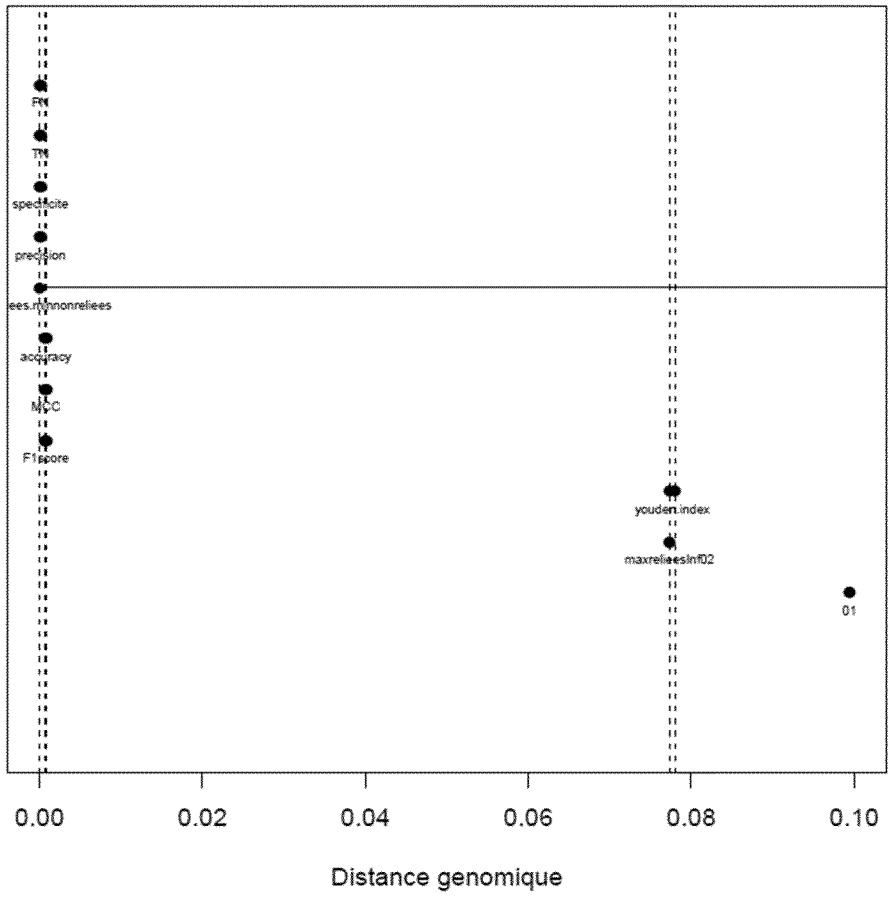


Figure 8B



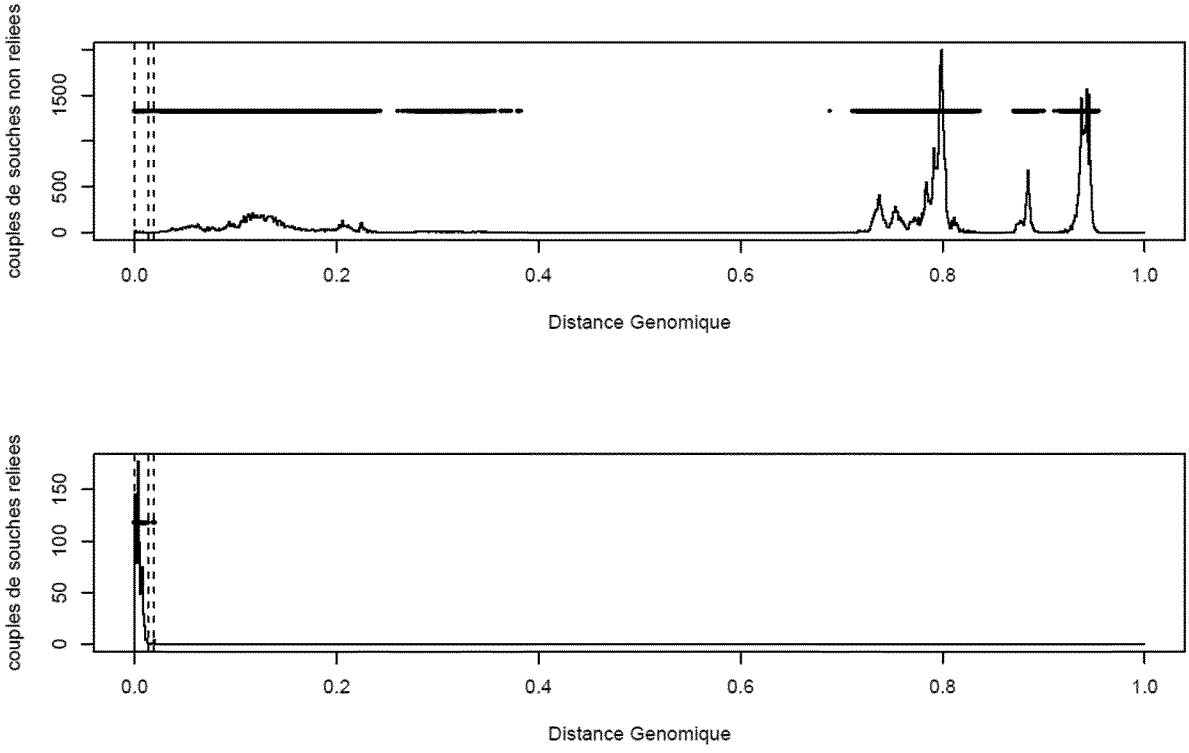


Figure 9A

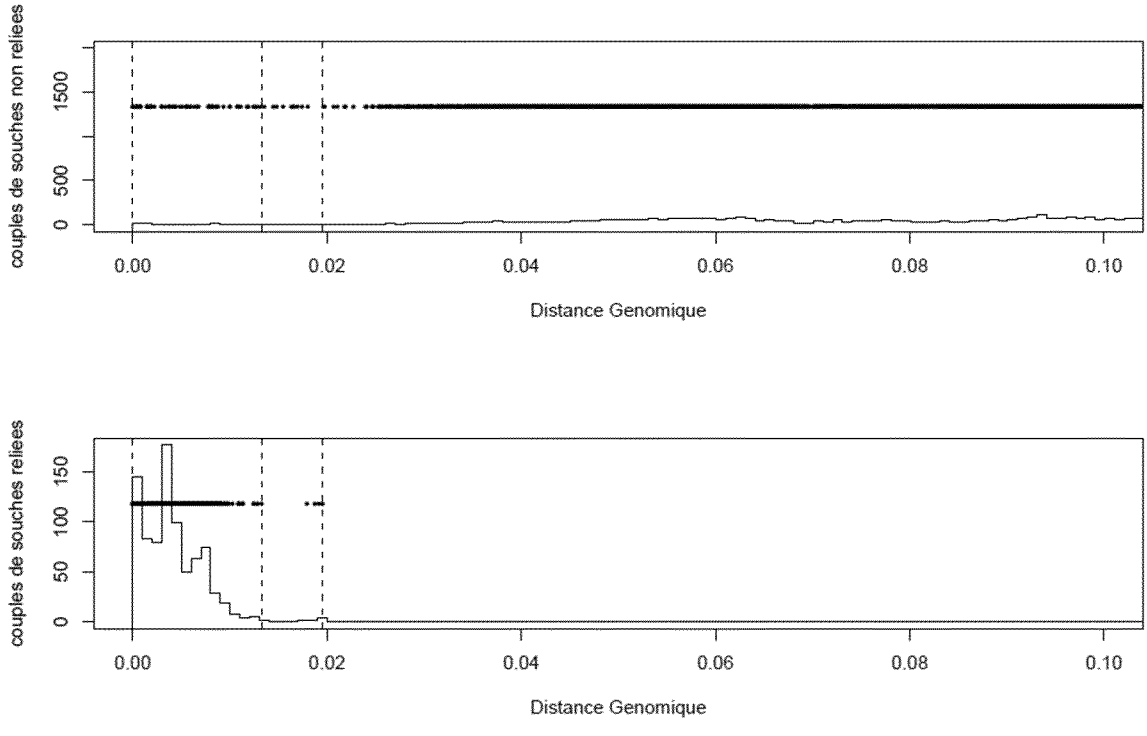


Figure 9B

Figure 10A

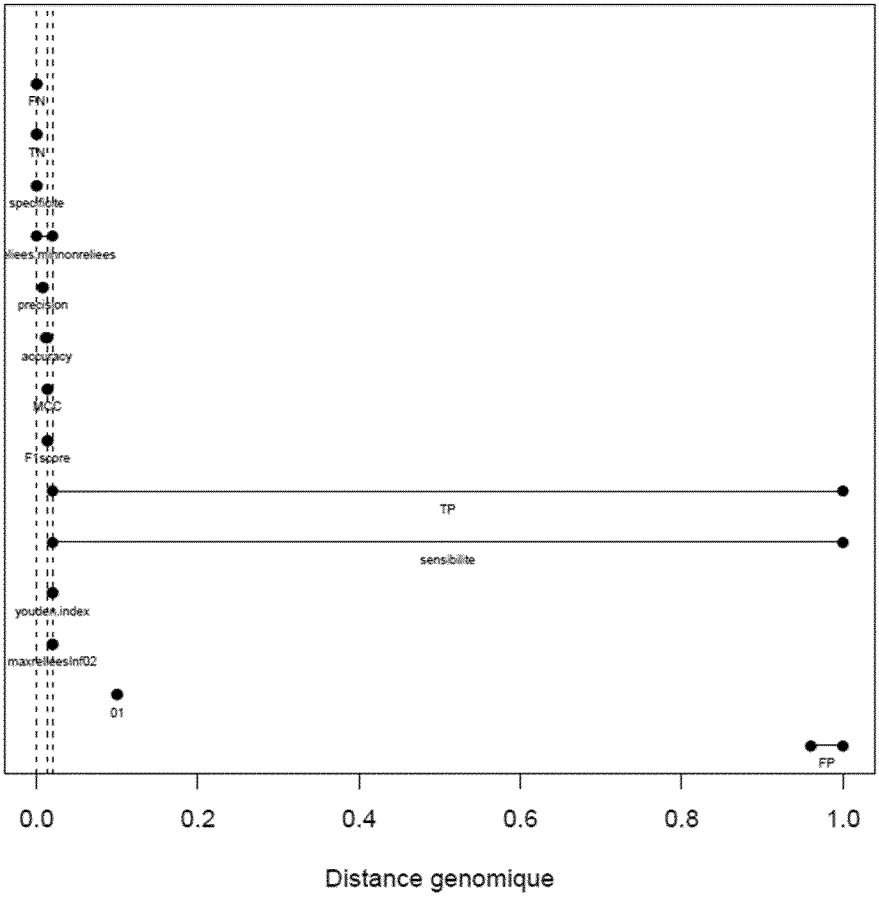
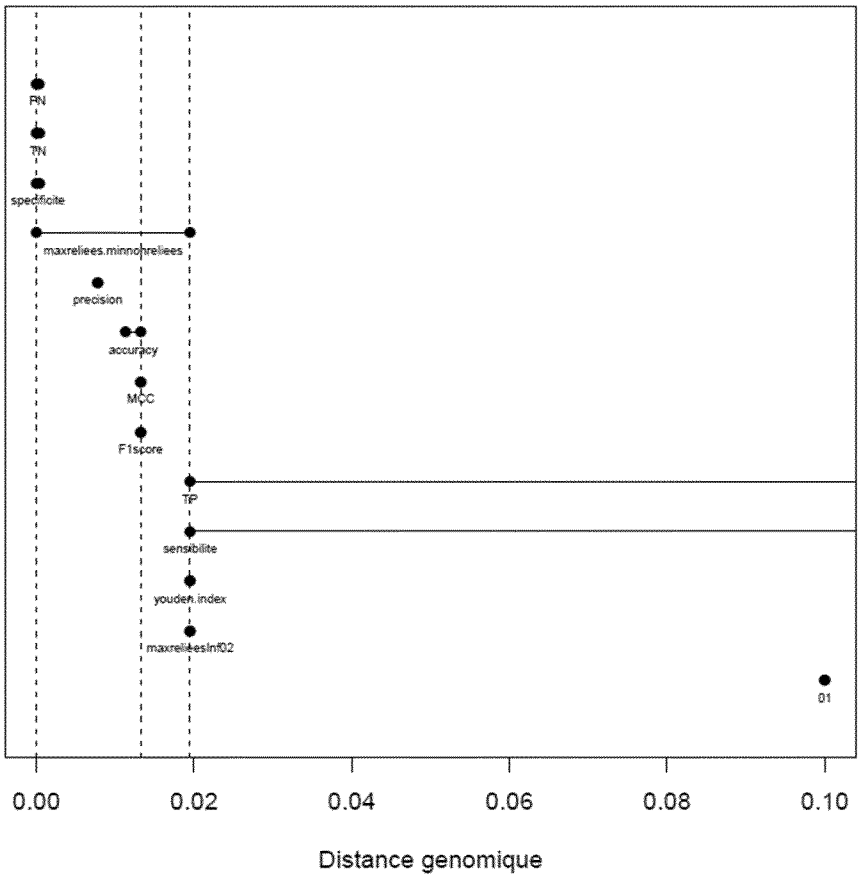


Figure 10B



INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2020/068611

A. CLASSIFICATION OF SUBJECT MATTER

G16H 50/80(2018.01)i; *G16B 20/20*(2019.01)n

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16H

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MARJOLEIN F. Q. KLUYTMANS-VAN DEN BERGH ET AL. "Whole-Genome Multilocus Sequence Typing of Extended-Spectrum-Beta-Lactamase-Producing Enterobacteriaceae" *JOURNAL OF CLINICAL MICROBIOLOGY*, US, Vol. 54, No. 12, 14 September 2016 (2016-09-14), pages 2919-2927 DOI: 10.1128/JCM.01648-16 ISSN: 0095-1137, XP055656833 abstract page 2929, column 2, paragraph 2 page 2923, column 1, paragraph 2 page 2921, column 1, paragraph 2 page 2919, column 1, paragraph 1	1-15

☑ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 July 2020	**06 August 2020**
Name and mailing address of the ISA/EP **European Patent Office** **p.b. 5818, Patentlaan 2, 2280 HV Rijswijk** **Netherlands** Telephone No. **(+31-70)340-2040** Facsimile No. **(+31-70)340-3016**	Authorized officer **Schmitt, Constanze** Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	James Stimson ET AL. "Abstract" *bioRxiv*, 03 December 2018 (2018-12-03), Retrieved from the Internet: https://www.biorxiv.org/content/10.1101/319707v3.f ull.pdf [retrieved on 2020-01-09] DOI: 10.1101/319707 XP055656344 abstract table 1	1-15
A	MICHAEL INOUYE ET AL. "Short read sequence typing (SRST): multi-locus sequence types from short reads" *BMC GENOMICS, BIOMED CENTRAL*, Vol. 13, No. 1, 24 July 2012 (2012-07-24), page 338 DOI: 10.1186/1471-2164-13-338 ISSN: 1471-2164, XP021115188 abstract; figure 1 page 2920, column 2, paragraph 2 - page 2921, column 1, paragraph 3	1-15
A	Suresh Babu ET AL. "Various performance measures in Binary classification -An Overview of ROC study" *International Journal of Innovative Science, Engineering & Technology*, 01 September 2015 (2015-09-01), Retrieved from the Internet: http://ijiset.com/vol2/v2s9/IJISET_V2_I9_72.pdf [retrieved on 2020-01-13] XP055657111 abstract page 597, paragraph 6 - page 601, paragraph 3	1-15

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°
PCT/EP2020/068611

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
INV. G16H50/80
ADD. G16B20/20

Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)
G16H

Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si cela est réalisable, termes de recherche utilisés)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	MARJOLEIN F. Q. KLUYTMANS-VAN DEN BERGH ET AL: "Whole-Genome Multilocus Sequence Typing of Extended-Spectrum-Beta-Lactamase-Producing Enterobacteriaceae", JOURNAL OF CLINICAL MICROBIOLOGY, vol. 54, no. 12, 14 septembre 2016 (2016-09-14), pages 2919-2927, XP055656833, US ISSN: 0095-1137, DOI: 10.1128/JCM.01648-16 abrégé page 2929, colonne 2, alinéa 2 page 2923, colonne 1, alinéa 2 page 2921, colonne 1, alinéa 2 page 2919, colonne 1, alinéa 1 ----- -/--	1-15

[X] Voir la suite du cadre C pour la fin de la liste des documents

[ ] Les documents de familles de brevets sont indiqués en annexe

* Catégories spéciales de documents cités:

"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent

"E" document antérieur, mais publié à la date de dépôt international ou après cette date

"L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée)

"O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens

"P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée

"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention

"X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément

"Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier

"&" document qui fait partie de la même famille de brevets

Date à laquelle la recherche internationale a été effectivement achevée	Date d'expédition du présent rapport de recherche internationale
28 juillet 2020	06/08/2020

Nom et adresse postale de l'administration chargée de la recherche internationale	Fonctionnaire autorisé
Office Européen des Brevets, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Schmitt, Constanze

1

C(suite). DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
A	James Stimson ET AL: "Abstract", bioRxiv, 3 décembre 2018 (2018-12-03), XP055656344, DOI: 10.1101/319707 Extrait de l'Internet: URL:https://www.biorxiv.org/content/10.1101/319707v3.full.pdf [extrait le 2020-01-09] abrégé tableau 1	1-15
A	MICHAEL INOUYE ET AL: "Short read sequence typing (SRST): multi-locus sequence types from short reads", BMC GENOMICS, BIOMED CENTRAL, vol. 13, no. 1, 24 juillet 2012 (2012-07-24), page 338, XP021115188, ISSN: 1471-2164, DOI: 10.1186/1471-2164-13-338 abrégé; figure 1 page 2920, colonne 2, alinéa 2 - page 2921, colonne 1, alinéa 3	1-15
A	Suresh Babu ET AL: "Various performance measures in Binary classification -An Overview of ROC study", International Journal of Innovative Science, Engineering & Technology, 1 septembre 2015 (2015-09-01), XP055657111, Extrait de l'Internet: URL:http://ijiset.com/vol2/v2s9/IJISET_V2_I9_72.pdf [extrait le 2020-01-13] abrégé page 597, alinéa 6 - page 601, alinéa 3	1-15

1