

## Connaissance du génome de *Helicobacter pylori* : implications pour la physiopathologie et la thérapeutique

Agnès Labigne

La lecture linéaire de chacun des deux génomes publiés de *Helicobacter pylori* a permis de répondre à un certain nombre de questions relatives au pouvoir pathogène, au métabolisme, à l'organisation du génome et à l'expression des gènes de cette espèce bactérienne. Leur analyse comparative a révélé des propriétés uniques d'*H. pylori* et a mis en exergue l'extrême diversité de son génome. Mais c'est surtout en tant qu'outils que la disponibilité de ces séquences constitue un atout sans précédent pour le développement de recherches fondamentales ciblées (effets du fer, rôle des adhésines, variations antigéniques du LPS, des protéines de surface...), et pour élaborer des approches globales ou post-génomiques. C'est ainsi que l'étude de la diversité des souches isolées en clinique à l'échelle génomique a pu être envisagée, de même que l'expression différentielle des gènes en fonction des souches et de leur environnement. Enfin, cette information a aussi permis la mise en place de stratégies d'envergure visant à identifier de façon systématique des antigènes protecteurs et des cibles thérapeutiques.

Lorsqu'en juillet 1995 le journal *Science* publie la première séquence génomique complète d'un micro-organisme, celle de *Haemophilus influenzae* [1] réalisée par TIGR (*The Institute for Genomic Research*), la communauté scientifique n'imagine pas que cet événement marque le début d'une ère nouvelle dans le monde

de la microbiologie. Il est le résultat d'une avancée technologique sans précédent dans le domaine du séquençage, associant l'analyse automatisée des séquences d'ADN et la bioinformatique. Cette avancée ouvrait la voie à une science nouvelle, « la génomique », avec laquelle microbiologistes, industriels, et l'ensemble de la communauté scien-

### ADRESSE

A. Labigne : Unité de pathogénie bactérienne des muqueuses, Institut Pasteur, 28, rue du Docteur-Roux, 75724 Paris Cedex 15, France.

tifique non seulement allaient devoir cohabiter mais qu'ils allaient devoir intégrer, assimiler, exploiter.

La détermination de la séquence génomique de *H. influenzae*, n'a pas été un événement unique; depuis 1995 pas moins de 48 séquences complètes de génomes de micro-organismes, dont la taille variait de 580 kilobases (0,58 Mégabases [Mb] ou millions de bases) pour *Mycoplasma genitalium* à 13 Mb dans le cas de *Saccharomyces cerevisiae*, ont été décryptées (<http://www.cbs.dtu.dk/services/Genome Atlas/>). Pourquoi un tel engouement, mais aussi un tel investissement? Qu'apporte la connaissance des génomes et, en particulier, celle d'un pathogène tel que *Helicobacter pylori*? Ce sont les réponses à ces questions que l'on tentera d'apporter dans cette revue.

*Helicobacter pylori* est un agent pathogène, responsable des infections bactériennes associées à la majorité des maladies gastro-duodénales inflammatoires existant chez l'homme, les gastrites ou pangastrites chroniques, les ulcères duodénaux et gastriques, l'atrophie gastrique, les lymphomes de type MALT et les cancers gastriques. Depuis son identification en 1983, un certain nombre de propriétés ont été identifiées et analysées par des approches génétiques et moléculaires. Celles-ci permettent de mieux comprendre comment cette bactérie, retrouvée exclusivement au niveau de la muqueuse gastrique ou au niveau des métaplasies gastriques du duodénum, peut survivre au niveau d'une muqueuse que l'on croyait stérile du fait de l'acidité gastrique, s'y multiplier, persister en dépit de la forte réponse immunitaire qu'elle déclenche, et enfin engendrer des lésions. La publication en 1997 de la séquence génomique d'une première souche de *H. pylori* [2] (souche 26695 séquencée par TIGR) isolée d'un patient anglo-saxon présentant une gastrite, puis celle d'une deuxième souche [3] en 1999 (souche J99, séquencée par *Genome Therapeutics Corporation*) isolée d'un patient américain présentant un ulcère duodéal, mettaient à la disposition de la communauté scientifique une mine de données, de points de repères qui permettent maintenant à chacun de poser des questions précises, et d'avoir une

chance de pouvoir y répondre du fait de la puissance des outils qui peuvent être mis en œuvre.

### **Qu'a apporté la lecture linéaire et l'analyse comparative des deux séquences génomiques de *H. pylori* ?**

Cette revue n'a pas pour objet d'analyser de façon exhaustive l'ensemble des données génomiques. De nombreuses revues l'ont fait en détail [4-6]. Toutefois, un certain nombre d'exemples seront discutés ici, qui illustrent à la fois le type d'informations que l'on peut tirer de l'analyse des séquences génomiques et surtout de la génomique dite comparative, en présentant les questions qui émergent de ces nouvelles connaissances génomiques.

#### **Les caractéristiques globales du génome**

Le génome de la souche 26 695 et celui de la souche J99 de *H. pylori* consistent en un chromosome circulaire dont 91% est dit codant, c'est-à-dire qu'il possède des phases ouvertes de lecture permettant de prédire la synthèse d'une chaîne polypeptidique; le reste du génome est constitué de séquences intergéniques (6%), de séquences répétées non codantes (2,3%) ou d'ARN stables (0,7%, dont deux copies de 23S et de 16S ainsi que trois copies de 5S ribosomiques). Pour les deux génomes, 77% des phases ouvertes de lecture (ORF) prédites codent pour des protéines pour lesquelles il existe des homologues (orthologues) dans d'autres organismes. Certains de ces orthologues ont des fonctions connues (58% des ORF), ce qui permet alors d'assigner cette même fonction à la protéine de *H. pylori*; d'autres sont inconnus, mais on sait qu'une telle protéine existe dans plusieurs organismes occupant par exemple des niches très différentes, et donc que la protéine n'est pas unique et spécifique à *H. pylori*. En revanche, 23% des protéines prédites de *H. pylori* n'ont pas d'homologues connus actuellement dans le monde du vivant, et sont codées par des gènes dits « orphelins », une donnée

qui est toute relative et qui peut être amenée à changer au fur et à mesure de la publication de nouveaux génomes. De tels gènes apparaissent comme des candidats d'intérêt majeur car *a priori* spécifiques de *H. pylori*.

A ces données brutes, tirées de la lecture individuelle des deux génomes, s'opposent celles provenant de leur comparaison. En effet, les gènes codant pour une centaine de protéines dans chacun des deux génomes ne sont pas retrouvés dans l'autre génome et réciproquement, alors que les deux souches possèdent toutes deux l'îlot de pathogénicité « *cag* » comprenant 31 ORF qui sont connues pour être présentes ou absentes des génomes selon les souches [7] et qui sont plus fréquemment associés à des affections sévères chez les patients occidentaux [8, 9]. De même, alors que chacune des deux souches renferme 23% de ses gènes qui codent pour des protéines sans homologue dans le monde du vivant, les gènes spécifiques de *H. pylori* de la souche 26 695 ne sont pas tous retrouvés dans le génome de la souche J99 et réciproquement. L'analyse comparative des deux génomes permet donc de s'interroger sur le degré de diversité des isolats cliniques, sur la nature des gènes communs à toutes les souches et sur la nature des gènes communs et spécifiques de l'espèce *H. pylori*.

#### **Métabolisme**

Sans développer extensivement ce domaine, les données du métabolisme déduit de la séquence des deux génomes sont en accord, dans la grande majorité des cas, avec les analyses biochimiques qui avaient été réalisées avant l'analyse génomique. Le seul sucre utilisé par *H. pylori* comme source de carbone est le glucose, et le génome de *H. pylori* est dépourvu de tout système de transport des autres sucres (système PTS) ainsi que du système de contrôle catabolique (absence de CRP et d'adényl-cyclase). *H. pylori* possède également les voies métaboliques nécessaires à l'utilisation des acides aminés et des acides organiques comme source de carbone (voie d'Entner-Doudoroff, cycle des pentoses, et cycle des acides tricarboxy-

liques), et les voies métaboliques permettant l'assimilation de l'azote à partir de l'urée, de l'ammoniac et de certains acides aminés (glutamine, alanine, sérine). L'ensemble des voies de biosynthèse des acides aminés qui étaient connus comme des acides aminés dont l'ajout n'était pas requis pour la croissance *in vitro* de *H. pylori* a été identifié alors que les voies de synthèse des 6 acides aminés requis [10] n'ont pas été retrouvés, témoignant d'une bonne concordance entre études fonctionnelles et analyse génomique.

### Contrôle de l'expression des gènes

Une des surprises associées au décryptage du génome de *H. pylori* concerne la pauvreté des systèmes de régulation mis en œuvre par *H. pylori*. Tout d'abord, pour la transcription de ses gènes, *H. pylori* est équipé d'une ARN polymérase atypique. Cette polymérase fonctionne avec un nombre limité de facteurs sigma: seuls les sigmas 70, 54 et 28 ont été détectés, et très peu de protéines de type modulateur transcriptionnel avec un motif hélice-tour-hélice ont été retrouvés: seuls 4 ont été identifiés chez *H. pylori* contre 34 chez *Haemophilus influenzae* et 148 chez *E. coli*, et environ le tiers des protéines de type senseur trouvées chez *E. coli* a été identifié chez *H. pylori*. Cette information est importante car elle reflète très probablement la spécialisation et l'adaptation de *H. pylori* à une niche écologique unique. La cellule n'a sans doute pas besoin de faire face à des conditions de vie très différentes qui nécessiteraient l'expression de nouveaux gènes et une telle observation tend à favoriser l'hypothèse d'une transmission interhumaine exclusive des bactéries en l'absence de tout autre réservoir environnemental, ce qui auparavant était présenté comme une possibilité. Toutefois, l'analyse de chacun des génomes et leur comparaison a révélé un certain nombre de propriétés spécifiques de *H. pylori*: de nombreux gènes présentent à leur extrémité 5' des séries d'homopolymères d'un ou deux nucléotides répétés et la comparaison des deux souches a montré que, pour certains gènes, le nombre de répétitions permettait soit la mise en phase du cadre de lecture et la synthèse de protéines

complètes, soit celle de protéines incomplètes et donc non fonctionnelles [2, 11]. De même, de nombreuses phases ouvertes identifiées dans l'un des deux génomes comme une seule phase ouverte sont retrouvées comme deux phases ouvertes possibles dans l'autre génome, et réciproquement, et ce du fait d'une microdélétion ou d'une micro-insertion [12]. Il semble donc que, pour la synthèse de certaines des protéines et la mise *on/off* de leur synthèse, *H. pylori* utilise un moyen extrêmement précaire (comparée à une modulation transcriptionnelle) résultant de microdélétions ou de micro-insertions permettant un changement du cadre de lecture défini. Depuis la parution des deux génomes, ces observations *in silico*, ont été confirmées par l'étude de la variabilité de l'expression et de la synthèse des protéines associées à de nombreuses enzymes, parmi lesquels figurent certains des 15 systèmes de restriction et de modification répertoriés dans les deux génomes ainsi que certains des facteurs de virulence connus: mobilité [13], adhésines [2], biosynthèse du lipopolysaccharide (LPS) [14].

### Facteurs de virulence et pouvoir pathogène

Les données génomiques ont incontestablement confirmé le rôle de la mobilité dans le pouvoir de colonisation de la muqueuse gastrique; au-delà des gènes préalablement connus tels que les gènes *flaA* [15], *flaB* [16], *flbA* [17], *flgE* [18] codant respectivement pour les flagellines majeures et mineures, un modulateur, et le crochet d'ancrage du flagelle, une vingtaine d'autres gènes ont été identifiés qui codent pour des homologues de la machinerie d'assemblage et de sécrétion des flagellines présents chez d'autres bactéries. Pour l'un de ces gènes, le gène *fliP* [13], il a été récemment montré que son expression était soumise à une variation de phase: une phase d'expression dite *on* associée à la mobilité de la bactérie, correspond à la présence d'une série consécutive (homopolymère) de huit cytosines (C<sub>8</sub>) en 5' du gène *fliP*, alors que la présence de neuf cytosines (C<sub>9</sub>) correspond à l'absence de synthèse de la protéine FliP et est

associée à un phénotype non mobile pour la bactérie. Outre ces gènes impliqués dans la machinerie, un certain nombre de gènes codant pour des homologues de protéines impliquées dans le chimiotactisme ont été identifiés et leur rôle pourra être étudié par rapport aux chimio-attractants potentiels que constituent le bicarbonate et l'urée déjà décrits [19].

Aucun gène typiquement associé à la synthèse de structures piliées telles qu'observées chez les bactéries à Gram négatif n'a été identifié. De même, les machineries (système de sécrétion de type III) impliquées dans la sécrétion des adhésines et des invasines, trouvées chez *Shigella*, *Salmonella*, *Yersinia* et autres bactéries pathogènes, n'ont pas été détectées chez *H. pylori*. En revanche, l'analyse du génome a mis en évidence une particularité unique de *H. pylori*: ce dernier possède deux fois plus de gènes impliqués dans la synthèse de composants de l'enveloppe cellulaire que d'autres bactéries [2]. Parmi les protéines de surface identifiées qui pourraient jouer un rôle dans l'adhérence, on peut noter la présence de 19 lipoprotéines qui s'ajoutent aux deux lipoprotéines préalablement identifiées comme adhésines (HpaA) [20] et AlpA [21]. De même, le génome de *H. pylori* contient 32 phases ouvertes codant pour une famille de protéines de membrane externe (d'un poids moléculaire moyen de 70 kDa) dont certaines avaient été reconnues comme porines (HopA, HopB, HopC, HopD, HopE) [22, 23] ou, comme adhésine reconnaissant le récepteur Lewis<sup>b</sup> des cellules gastriques (BabA) [24]. Ces 32 protéines présentent des domaines extrêmement conservés dans leurs parties amino-terminale et carboxy-terminale, répondant très certainement à des contraintes conformationnelles et fonctionnelles. La présence d'un tel nombre de protéines homologues laisse présager d'une possibilité de réarrangements conduisant à des protéines mosaïques et à une diversité des protéines exprimées en surface comme possible moyen de variation antigénique. Pour bon nombre des gènes codant pour ces protéines, il existe des motifs dans la région 5' du gène consistant en une succession de

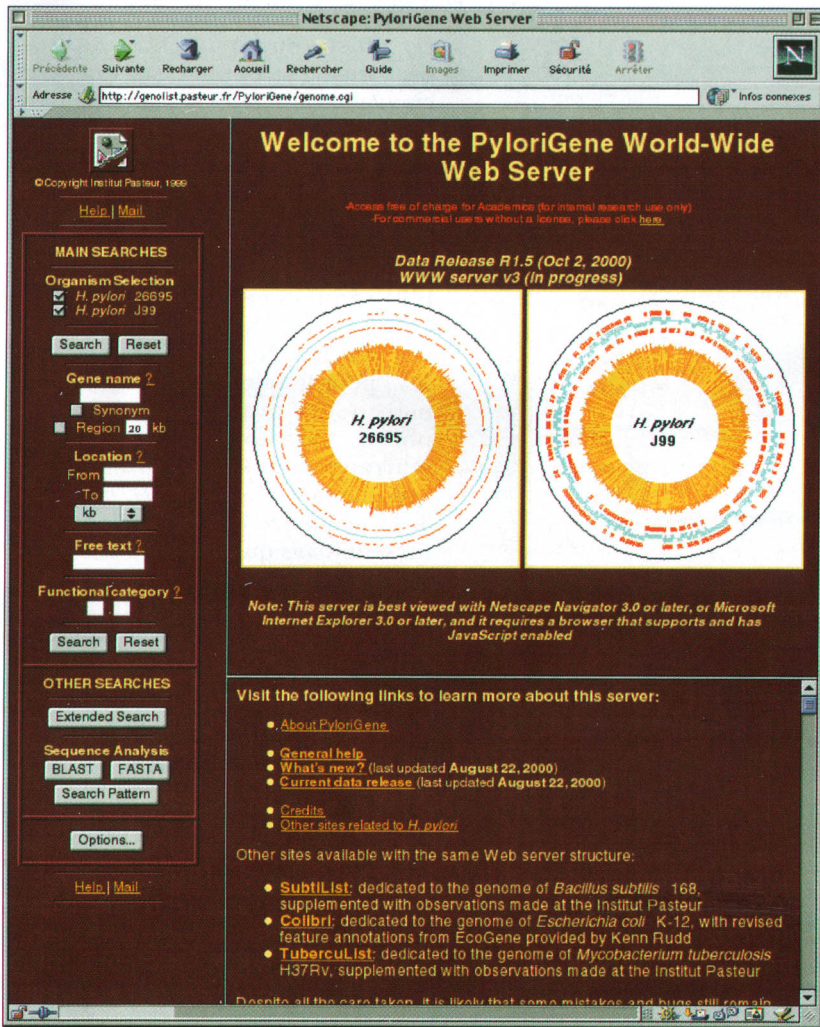


Figure 1. Site Internet de l'Institut Pasteur (<http://genolist.pasteur.fr/PyloriGene/>).

mononucléotides (polyA, polyG, polyC ou polyT) ou de dinucléotides (CTn, AGn) qui permettent ou non la synthèse d'une protéine fonctionnelle; cela confère une possibilité de choix multiples de synthèse de protéines de membrane. L'iver *et al.* [24] ont très récemment montré qu'un autre mécanisme pouvait intervenir *via* la duplication de 10 paires de bases en début de la phase ouverte qui permettait à l'une des deux copies d'un gène homologue de produire une protéine fonctionnelle. Le répertoire des gènes codant pour des protéines de surface de *H. pylori* est excessivement surprenant. La nouvelle connaissance que l'on a de cette situation, inhabituelle, devrait permettre de comprendre si une telle

possibilité de variation est une nécessité pour la bactérie comme moyen d'échappement à la réponse immunitaire. Dans ce cas, la structure primaire de ces protéines devrait changer au cours d'une infection persistante. Alternativement, ces protéines pourraient procurer à la bactérie des moyens de s'adapter à l'hôte en fonction de l'expression des récepteurs cellulaires fucosylés qu'il exprime. Dans ce cas, une fois implanté, l'expression d'un jeu de ces protéines devrait rester constante.

Ces dernières années une situation unique a été mise en évidence chez *H. pylori*: la présence de motifs oligosaccharidiques de type Lewis<sup>x</sup>, Lewis<sup>y</sup>, Lewis<sup>i</sup> et Lewis<sup>a</sup> associés au lipopoly-

saccharide (LPS) de la bactérie. D'abord soupçonnés de jouer un rôle dans un processus d'auto-immunité, il est plus probable que ces motifs jouent un rôle dans l'adhérence des bactéries à des lectines du mucus ou des cellules gastriques. Leur présence inhabituelle chez des bactéries confère à ces déterminants un intérêt particulier. Là encore, les gènes requis pour la biosynthèse du lipide A, de la région *core* du LPS, ont été identifiés par analogie avec d'autres systèmes de biosynthèse du LPS. Contrairement aux autres bactéries, ces gènes sont dispersés sur le génome chez *H. pylori*. Trois fucosyltransférases et une galactosyltransférase ont été identifiées dans les génomes, et la synthèse de chacune de ces quatre enzymes est elle aussi soumise à une variation de phase *on/off* associée à des microdélétions dans des homopolymères permettant ou non le branchement des résidus fucose pour l'expression des motifs Lewis du LPS [14].

Parmi les facteurs de virulence classiquement décrits chez les bactéries figurent les systèmes de captage du fer. Cela est d'autant plus une nécessité pour *H. pylori* que la bactérie vit dans un environnement naturellement riche en macromolécules capables de séquestrer le fer disponible au sein du mucus. La séquence du génome a mis en évidence l'existence de très nombreux systèmes de captage du fer présent soit sous forme libre (Fe<sup>2+</sup> et Fe<sup>3+</sup>) soit sous forme complexée:

- homologues de perméase des ions Fe<sup>2+</sup> (FeoB) [25];
- trois homologues de la protéine FecA, récepteur du citrate ferrique (Fe<sup>3+</sup>), alors que généralement un seul est présent chez d'autres bactéries;
- partenaires du récepteur FecA (FecD et FecE) alors que des protéines homologues aux protéines FecB et FecC de *E. coli* sont absentes. *H. pylori* contient aussi 5 gènes codant pour des homologues de la protéine FrpB, connue pour capter le fer, associées à l'hème ou à la lactoferrine. Enfin, *H. pylori* possède des gènes codant pour des homologues des protéines TonB, ExbB et ExbD (3 homologues de chaque) nécessaires à l'acquisition globale du fer par la cellule. Pourquoi une telle redondance des systèmes de captage des ions fer-

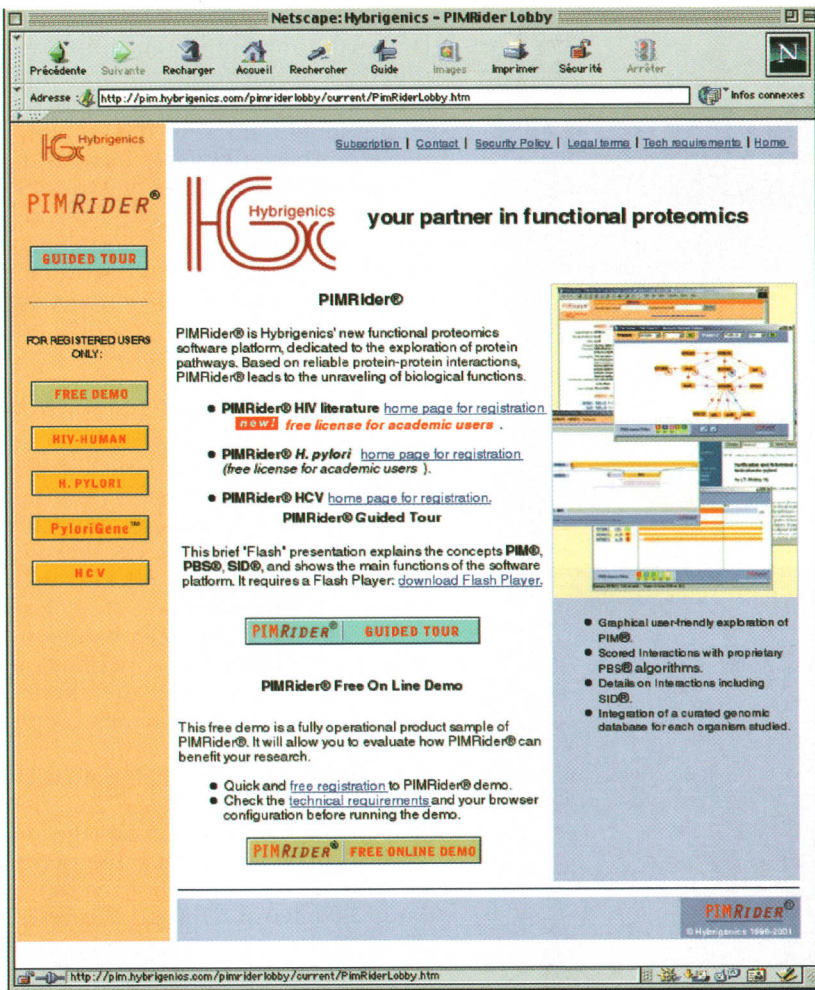


Figure 2. Carte d'interaction protéine/protéine (groupe Hybrigenics; <http://pim.hybrigenics.com>).

reux-ferriques, une telle complexité ? Tous ces systèmes sont-ils présents dans toutes les souches ? Présentent-ils une expression différentielle en fonction du stade de l'infection (lumière gastrique, contact avec l'épithélium, mucus) ? Ce sont des questions auxquelles les approches post-génomiques permettront de répondre.

### Qu'apporte la connaissance des génomes comme outils de travail pour la connaissance des micro-organismes ?

L'information disponible sur les différents sites Internet, parmi lesquels ceux de TIGR (<http://www.tigr.org/>

[tdb/mdb/hpdb/hpdb.html](http://tdb/mdb/hpdb/hpdb.html)), d'Astra (<http://scriabin.astrazeneca-boston.com/hpylori/>) ou de l'Institut Pasteur (<http://genolist.pasteur.fr/PyloriGene/>) (figure 1), est de trois ordres :

1. Une liste des phases ouvertes de lecture prédites et leur annotation. Cette annotation résulte de la comparaison des protéines déduites des phases ouvertes avec l'ensemble des protéines étudiées dans le monde procaryote et eucaryote dont les fonctions sont connues, ou inconnues, permettant de classer les ORF comme codant pour des fonctions spécifiques de *H. pylori*, pour des orthologues de fonctions connues ou pour des orthologues de fonctions inconnues.

2. La position relative de chacune de ces phases ouvertes (ORF) les unes par rapport aux autres, sur une carte circulaire de 1 600 000 paires de base, pour laquelle la position 1 a été fixée arbitrairement. Cette carte donne une vision globale de l'organisation du génome. D'ores et déjà, la comparaison de la carte génomique de la souche 26 695 avec celle de la souche J99, séquencée, ou de la souche CCUG17875 (NTCC11638) [26] et des souches NCTC11639 et UA861 [27] pour lesquelles une vingtaine de marqueurs avaient été cartographiées, font apparaître de sérieux réarrangements de l'ensemble du génome des souches de *H. pylori*.

3. La séquence brute du génome en un seul *contig* qui donne la possibilité pour tout informaticien d'y appliquer des algorithmes pour rechercher des duplications, des motifs nucléotidiques répétés, ou de potentielles séquences promotrices ou régulatrices. Une telle approche a été utilisée globalement par Vanet *et al.* [28] pour définir la séquence consensus associée aux promoteurs des gènes de *H. pylori*. Cette étude a permis de prédire une structure inhabituelle des promoteurs chez *H. pylori*: TTAAGCx(19-23)TATAAT comparée à la séquence consensus TTGACAx(15-19)TATAAT retrouvée classiquement chez les bactéries à Gram négatif [28].

L'utilisation de ces trois types d'outils permet de poser un nombre illimité de questions qui peuvent être abordées soit par des approches ciblées concernant l'étude de quelques gènes d'intérêt, soit par des approches systématiques visant à étudier le génome dans sa globalité, que nous développerons plus particulièrement dans cette revue.

### Étude globale de la diversité des souches

Une des questions soulevées par la comparaison des deux génomes concerne la diversité des souches naturelles. Si l'on considère comme point de référence la souche 26 695, on peut en effet se demander combien des 1 590 gènes décrits pour le génome de cette souche sont présents dans toutes les souches de *H. pylori* (gènes ubiquitaires), et plus encore combien parmi ceux qui sont

ubiquitaires sont spécifiques à *H. pylori*. En effet, l'identification de ces gènes permet *a priori* d'identifier un nombre limité de gènes, dont les fonctions sont inconnues, parmi lesquels figurent ceux qui confèrent à *H. pylori* l'aptitude de se multiplier dans une niche aussi particulière que l'estomac. Il est possible d'apporter une réponse à ces questions en utilisant des membranes de nylon ou des lames de verre sur lesquelles l'ensemble des 1590 ORF du génome de la souche 26 695, produits par amplification génique, ont été déposés sous forme de *spots*. L'ADN chromosomique des isolats cliniques est alors extrait, marqué à l'aide de nucléotides radioactifs ou fluorescents, et utilisé dans des réactions d'hybridation moléculaire pour déterminer celles des ORF qui sont présentes ou absentes des différents génomes. Une telle étude a récemment été publiée et montrait que 1 281 des 1 590 ORF étaient présentes dans les 15 souches d'origine occidentale analysées [29]. En réalisant de telles études à partir de 24 souches provenant pour 8 d'entre elles d'Europe, 8 d'Afrique subsaharienne, et 8 d'Asie, nous avons pu montrer que le *core* génomique des souches de *H. pylori* consistait en 1 097 ORF, que dans les différents isolats cliniques 4 % à 10 % des ORF de la souche 26 695 pouvaient être absents, que 242 des ORF du génome de la souche 26 695 n'étaient pas ubiquitaires (alors que la comparaison des deux génomes en avait révélé 98), enfin que 210 ORF étaient ubiquitaires et spécifiques de *H. pylori*. Enfin, une telle étude mettait en évidence que certaines des ORF n'étaient jamais retrouvées dans les souches asiatiques ou les souches africaines. Ces analyses comparatives permettent d'envisager d'étudier à plus grande échelle la distribution des ORF dites non ubiquitaires au sein d'un grand nombre de souches associées à des maladies bien précises (gastrite, ulcère, MALT, cancer) ; elles permettent également d'envisager une analyse fonctionnelle des ORF ubiquitaires et spécifiques de *H. pylori* sur un sous-jeu restreint puisque constitué de 210 ORF, plutôt que des 1 590 ORF issues du séquençage du génome.

### Étude de l'expression différentielle des gènes de *H. pylori* (transcriptome-protéome)

Un des intérêts majeurs de la fabrication des « puces à ADN » présentant l'ensemble des gènes fixés soit sur un support nylon, soit sur une lame de verre, est de les utiliser dans des expériences d'hybridation ADN/ARN. On peut ainsi comparer les expériences d'hybridation réalisées entre : (1) de l'ARN extrait de bactéries cultivées dans différentes conditions (*in vivo/in vitro*, milieu de culture à pH neutre ou pH acide, absence ou présence de source azotée), ou (2) de l'ARN extrait d'une souche parentale comparée à celui extrait d'une souche mutée dans un gène modulateur de façon à identifier globalement les gènes impliqués dans la réponse à un micro-environnement donné, ou sous le contrôle d'un même modulateur identifié. On appelle donc transcriptome l'analyse des transcrits d'une bactérie cultivée dans des conditions données. L'analyse différentielle des transcriptomes donne des informations sur la régulation de ces gènes, sur leur association au sein d'une voie métabolique, sur leur association en réponse à un stimulus, ou à l'expression d'un gène modulateur.

Une autre approche consiste non plus à visualiser le niveau d'expression des gènes par le niveau de synthèse des ARNm, mais à mesurer le niveau de synthèse des protéines dans la bactérie. C'est ce qui est désigné sous le terme de protéome. L'analyse consiste généralement à extraire les protéines et à les séparer en les faisant migrer dans des gels de polyacrylamide successifs permettant leur séparation par rapport à leur charge électrique puis en fonction de leur taille. Ces deux migrations successives doivent être réalisées dans des conditions très standardisées et reproductibles, et l'analyse, tout comme pour le transcriptome, passe par l'utilisation de logiciels informatiques permettant l'identification des taches protéiques variant dans deux conditions différentes. Une telle analyse a récemment été réalisée et publiée pour *H. pylori* [30], en comparant les protéines synthétisées après croissance des bactéries en milieu acide ou en milieu neutre. Ces

approches post-génomiques (transcriptome et protéome) en sont à leurs balbutiements et nécessitent standardisations et contrôles pour permettre de conclure avec certitude sur l'influence d'un facteur ou l'identification de gènes cibles pour un modulateur (régulon).

### Analyse fonctionnelle des protéines codées

Par analyse fonctionnelle, on entend une recherche conduisant à identifier la fonction des protéines codées. Cela implique tout d'abord de confirmer que la fonction assignée sur la base de la similitude avec une protéine de fonction connue dans un autre micro-organisme est correcte. Cette démarche n'est pas triviale, puisque l'on constate qu'en faisant l'analyse biochimique d'un certain nombre de protéines auxquelles étaient assignées des fonctions, certains auteurs ont montré que l'activité biochimique était différente de celle supputée sur la base de la similitude [31]. Indépendamment, l'analyse fonctionnelle a pour but de rechercher la fonction des ORF de fonction inconnue. Différentes approches sont utilisées pour cela. L'une d'entre elles a été récemment développée par le groupe Hybrigenics et a consisté à établir une carte d'interaction protéine/protéine (figure 2) entre des protéines de fonction inconnue et l'ensemble des protéines du génome par une approche de double hybride réalisée chez la levure [32]. Une telle approche permet dans certains cas de visualiser des interactions entre une protéine de fonction connue et une de fonction inconnue, et donc d'émettre des hypothèses sur son rôle. Une autre approche consiste à classer de façon très systématique tous les gènes selon leur caractère essentiel, non essentiel, ou essentiel conditionnel. Les gènes essentiels sont ceux dont l'inactivation spécifique ne permet pas à la bactérie de se multiplier, les gènes conditionnels sont ceux dont l'expression n'est pas essentielle pour la croissance de la bactérie *in vitro*, mais qui apparaissent comme essentiels à la survie de la bactérie dans un environnement particulier, tel que la muqueuse gastrique.

## Qu'apporte la séquence du génome dans la recherche de cibles thérapeutiques et prophylactiques ?

Il n'est de secret pour personne que la séquence génomique de *H. pylori* a été fortement convoitée par les principaux groupes industriels, qui cherchaient à identifier des antigènes protecteurs en vue de développer une immunisation active à visée thérapeutique ou prophylactique, ou d'identifier de nouvelles cibles thérapeutiques. De fait, les deux séquences génomiques indépendantes de *H. pylori* ont été exploitées à ces fins : la séquence de la souche 26 695 réalisée par TIGR, a été exploitée par la joint venture Aventis Pasteur/OraVax, avant sa publication en 1997, pour cribler à l'échelle génomique des antigènes protecteurs, et la séquence de la souche ATCC55679 (HP-J99), séquencée par Genome Therapeutics Corporation, a été exploitée par le groupe Astra avant sa publication en 1999, pour identifier des cibles thérapeutiques.

### Recherche d'antigènes protecteurs

La stratégie utilisée par Aventis Pasteur/OraVax a consisté à rechercher *in silico*, au sein des 1 590 phases ouvertes identifiées dans le génome de la souche 26 695, celles qui apparaissent comme spécifiques de *H. pylori* et/ou qui pouvaient être identifiées comme des protéines de surface ou périplasmiques ; cette assignation a été faite sur la base d'un certain nombre de critères qui permettaient de réduire le nombre de gènes d'intérêt quasiment de moitié. Chacune des 700 ORF ainsi sélectionnées a ensuite été amplifiée *in vitro*, et clonée en fusion avec une région *tag* dans un vecteur d'expression (une région *tag* est une séquence d'acides aminés courte qui a des propriétés d'affinité particulière à l'égard d'un substrat). Les protéines recombinantes ont été produites chez *Escherichia coli* et purifiées rapidement grâce à leur domaine *tag*. Enfin, le pouvoir protecteur de chacune des protéines recombinantes a été apprécié en l'administrant par voie orale, en présence d'un adjuvant mucosal (toxine cholérique), dans

un modèle d'infection murin. Le pouvoir protecteur de l'antigène recombinant est alors classé par rapport à celui des antigènes protecteurs déjà connus, à savoir l'uréease. Une approche générale et systématique a permis de cribler des centaines de molécules en réalisant successivement ces trois étapes, ADN, protéine et modèle animal, en grande série et a conduit à l'identification d'une dizaine d'antigènes protecteurs parmi les 300 testés.

### Criblage de molécules actives contre *H. pylori*

Toutes les compagnies pharmaceutiques disposent de banques de molécules de synthèse chimiques qui sont testées pour leur activité antimicrobienne. Un criblage de ces molécules en grandes séries à l'égard de *H. pylori* a toujours été limité du fait de la croissance fastidieuse de cette bactérie, qui ne se prête pas à une approche automatisée à grande échelle. Là encore, la connaissance du génome permet aux industriels de sélectionner quelques dizaines de cibles d'intérêt potentiel, tels que des enzymes et des protéines dont on a montré précédemment qu'elles jouaient un rôle essentiel pour la survie de la bactérie *in vitro* ou *in vivo*. Ayant sélectionné les cibles d'intérêt, il est alors possible de cloner le gène codant pour l'une de ces cibles, de produire une protéine recombinante et d'utiliser non pas la bactérie entière, mais la protéine recombinante pour rechercher des molécules inhibitrices soit d'une activité enzymatique, soit d'un transport actif, soit encore d'une interaction entre les constituants d'un homopolymère ou d'un hétéropolymère.

On voit que la connaissance du génome est très utile pour ce type d'approche. Sans résoudre l'ensemble des questions, elle permet d'identifier des antigènes protecteurs sur une période relativement courte et de cribler des molécules dirigées contre des cibles vitales pour la bactérie.

### Conclusions

Dans les années à venir, il s'agira d'apprécier le degré d'universalité de l'information associée à la publica-

tion d'une souche donnée, c'est-à-dire à sa valeur heuristique pour l'ensemble des souches cliniques de l'espèce à laquelle elle appartient. La connaissance des deux génomes de *H. pylori* offre une base de données et des outils accessibles à tous pour poser des questions de tout ordre. C'est avant tout un outil pour les « pyloristes » microbiologistes, cliniciens et industriels, et une donnée supplémentaire dans la compréhension globale de l'expression des gènes, de l'adaptation des microorganismes à leur environnement, de l'évolution des êtres vivants à l'ère post-génomique dans laquelle nous ne faisons qu'entrer ■

### RÉFÉRENCES

1. Fleischmann RD, Adams MD, White O, *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995 ; 269 : 496-512.
2. Tomb JF, White O, Kerlavage AR, *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 1997 ; 388 : 539-47.
3. Alm RA, Ling LS, Moir DT, *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 1999 ; 397 : 176-80.
4. Alm RA, Trust TJ. Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. *J Mol Med* 1999 ; 77 : 834-46.
5. Berg DE, Hoffman PS, Appelmelk BJ, Kusters JG. The *Helicobacter pylori* genome sequence: genetic factors for long life in the gastric mucosa. *Trends Microbiol* 1997 ; 5 : 468-74.
6. Marais A, Mendz GL, Hazell SL, Mégraud F. Metabolism and genetics of *Helicobacter pylori*: the genome era. *Microbiol Mol Biol Rev* 1999 ; 63 : 642-74.
7. Jenks P, Mégraud F, Labigne A. Clinical outcome after infection with *Helicobacter pylori* does not appear to be reliably predicted by the presence of any of genes of the *cag* pathogenicity island. *Gut* 1998 ; 43 : 752-8.
8. Akopyants NS, Clifton SW, Kersulyte D, *et al.* Analyses of the *cag* pathogenicity island of *Helicobacter pylori*. *Mol Microbiol* 1998 ; 28 : 37-53.
9. Censini S, Lange C, Xiang Z, *et al.* Cag, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc Natl Acad Sci USA* 1996 ; 93 : 14648-53.
10. Reynolds DJ, Penn CW. Characteristics of *Helicobacter pylori* growth in a defined medium and determination of its amino acid requirements. *Microbiology* 1994 ; 140 : 2649-56.

## RÉFÉRENCES

11. Marshall DG, Dundon WG, Beesley SM, Smyth CJ. *Helicobacter pylori*: a conundrum of genetic diversity. *Microbiology* 1998; 144: 2925-39.
12. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 1987; 4: 203-21.
13. Josenhans C, Eaton KA, Thevenot T, Suerbaum S. Switching of flagellar motility in *Helicobacter pylori* by reversible length variation of a short homopolymeric sequence repeat in *fljP*, a gene encoding a basal body protein. *Infect Immun* 2000; 68: 4598-603.
14. Appelmek BJ, Martino MC, Veenhof E, et al. Phase variation in H type I and Lewis a epitopes of *Helicobacter pylori* lipopolysaccharide. *Infect Immun* 2000; 68: 5928-32.
15. Leying H, Suerbaum S, Geis G, Haas R. Cloning and genetic characterization of a *Helicobacter pylori* flagellin gene. *Mol Microbiol* 1992; 6: 2863-74.
16. Suerbaum S, Josenhans C, Labigne A. Cloning and genetic characterization of the *Helicobacter pylori* and *Helicobacter mustelae* flaB flagellin genes and construction of *H. pylori* flaA- and flabB-negative mutants by electroporation-mediated allelic exchange. *J Bacteriol* 1993; 175: 3278-88.
17. Schmitz A, Josenhans C, Suerbaum S. Cloning and characterization of the *Helicobacter pylori* flbA gene, which codes for a membrane protein involved in coordinated expression of flagellar genes. *J Bacteriol* 1997; 179: 987-97.
18. O'Toole PW, Kostrzynska M, Trust TJ. Non-motile mutants of *Helicobacter pylori* and *Helicobacter mustelae* defective in flagellar hook production. *Mol Microbiol* 1994; 14: 691-703.
19. Mizote T, Yoshiyama H, Nakazawa T. Urease-independent chemotactic responses of *Helicobacter pylori* to urea, urease inhibitors, and sodium bicarbonate. *Infect Immun* 1997; 65: 1519-21.
20. Evans DG, Karjalainen TK, Evans DJ Jr, Graham DY, Lee CH. Cloning, nucleotide sequence, and expression of a gene encoding an adhesin subunit protein of *Helicobacter pylori*. *J Bacteriol* 1993; 175: 674-83.
21. Odenbreit S, Till M, Haas R. Identification of a *Helicobacter pylori*-specific protein involved in adherence to gastric epithelial cells. *Gut* 1995; 37: 3.
22. Exner MM, Doig P, Trust TJ, Hancock RE. Isolation and characterization of a family of porin proteins from *Helicobacter pylori*. *Infect Immun* 1995; 63: 1567-72.
23. Doig P, Exner MM, Hancock RE, Trust TJ. Isolation and characterization of a conserved poring protein from *Helicobacter pylori*. *J Bacteriol* 1995; 177: 5447-52.
24. Llover D, Arnqvist A, Ögren J, et al. *Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging. *Science* 1998; 279: 373-7.
25. Velayudhan J, Hughes NJ, McColm AA, et al. Iron acquisition and virulence in *Helicobacter pylori*: a major role for FeoB, a high-affinity ferrous iron transporter. *Mol Microbiol* 2000; 37: 274-86.
26. Bukanov NO, Berg DE. Ordered cosmid library and high-resolution physical-genetic map of *Helicobacter pylori* strain NCTC11638. *Mol Microbiol* 1994; 11: 509-23.
27. Jiang Q, Hiratsuka K, Taylor DE. Variability of gene order in different *Helicobacter pylori* strains contributes to genome diversity. *Mol Microbiol* 1996; 20: 833-42.
28. Vanet A, Marsan L, Labigne A, Sagot MF. Inferring regulatory elements from a whole genome. An analysis of *Helicobacter pylori* 880 family of promoter signals. *J Mol Biol* 2000; 297: 335-53.
29. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc Natl Acad Sci USA* 2000; 97: 14668-73.
30. Jungblut PR, Bumann D, Haas G, et al. Comparative proteome analysis of *Helicobacter pylori*. *Mol Microbiol* 2000; 36: 710-25.
31. O'Rourke EJ, Chevalier C, Boiteux S, Labigne A, Ielpi L, Radicella JP. A novel 3-methyladenine DNA glycosylase from *Helicobacter pylori* defines a new class within the endonuclease III family of base excision repair glycosylases. *J Biol Chem* 2000; 275: 20077-83.
32. Rain JC, Selig L, de Reuse H, et al. The protein-protein interaction map of *Helicobacter pylori*. *Nature* 2001; 409: 211-15.

## Summary

### The *Helicobacter pylori* genomes: new insights into physiopathology and therapeutic

Examination of the two genome sequences of *H. pylori* has already identified new putative virulence factors, unravelled basic metabolic pathways, and provided further information on genetic organisation and the regulation of gene expression. In the future, the genomes of these isolates will greatly facilitate research into specific areas of this bacteria's biology, such as iron regulation, the role of recently identified adhesins, and the relative importance of antigenic variation of LPS *in vivo*. It will also enable researchers to adopt a more global approach to the study of *H. pylori* and will allow: (1) examination of the genetic diversity of clinical isolates (gene polymorphism and diversity in genomic complement and organisation); (2) characterisation of differential gene expression of clinical isolates in relation to *in vitro* and *in vivo* conditions; (3) identification of genes essential for colonisation of the gastric mucosa. Finally, it already served for the development of systematic strategies to identify protective antigens for prophylactic or therapeutic vaccine development as well as for the search of essential and conserved targets for development of new drugs.

## TIRÉS À PART

A. Labigne.