Développement d'une base de connaissances du virus de l'hépatite B, HBVdb, pour l'étude de la résistance aux traitements : intégration d'outils d'analyses de séquences et application à la modélisation moléculaire de la polymérase

Juliette Hayer

École Doctorale Interdisciplinaire Sciences-Santé

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Lyon I

Spécialité : Aspects Moléculaires et Cellulaires de la Biologie

par

Juliette HAYER

Développement d'une base de connaissances du virus de l'hépatite B, HBVdb, pour l'étude de la résistance aux traitements. Intégration d'outils d'analyses de séquences et application à la modélisation moléculaire de la polymérase.

Soutenue publiquement le 15 février 2013

Directeur de thèse : Pr. Gilbert DELEAGE

Membres du jury :

Rapporteurs :	Mme. Claudine MEDIGUE, Directrice de Recherche CNRS, Evry
	M. Camille SUREAU, Directeur de Recherche CNRS, Paris
Examinateurs :	M. Pedro COUTINHO, Professeur des Universités, Marseille
	M. Fabien ZOULIM, Professeur des Universités – Praticien Hospitalier, Lyon
Directeur de thèse :	M. Gilbert DELEAGE, Professeur des Universités, Lyon
Co-encadrant :	M. Christophe COMBET, Chargé de Recherche CNRS, Lyon

Equipe BioInformatique : Structure et Interactions (BISI)
BMSSI - UMR 5086 CNRS - Université Lyon 1 - http://www.ibcp.fr
IBCP - 7, passage du Vercors - 69367 Lyon - FRANCE

École Doctorale Interdisciplinaire Sciences-Santé

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Lyon I

Spécialité : Aspects Moléculaires et Cellulaires de la Biologie

par

Juliette HAYER

Développement d'une base de connaissances du virus de l'hépatite B, HBVdb, pour l'étude de la résistance aux traitements. Intégration d'outils d'analyses de séquences et application à la modélisation moléculaire de la polymérase.

Soutenue publiquement le 15 février 2013

Directeur de thèse : Pr. Gilbert DELEAGE

Membres du jury :

Rapporteurs :	Mme. Claudine MEDIGUE, Directrice de Recherche CNRS, Evry
	M. Camille SUREAU, Directeur de Recherche CNRS, Paris
Examinateurs :	M. Pedro COUTINHO, Professeur des Universités, Marseille
	M. Fabien ZOULIM, Professeur des Universités – Praticien Hospitalier, Lyon
Directeur de thèse :	M. Gilbert DELEAGE, Professeur des Universités, Lyon
Co-encadrant :	M. Christophe COMBET, Chargé de Recherche CNRS, Lyon

Equipe BioInformatique : Structure et Interactions (BISI)
BMSSI - UMR 5086 CNRS - Université Lyon 1 - http://www.ibcp.fr
IBCP - 7, passage du Vercors - 69367 Lyon - FRANCE

Remerciements

Table des matières

Remerciements ..... **I**

Table des matières ..... **III**

Table des figures ..... **IX**

Table des tableaux ..... **XI**

Table des annexes ..... **XIII**

Abréviations ..... **XV**

Introduction **1**

Chapitre 1 : Rappels bibliographiques **3**

1.1 L'hépatite B : Epidémiologie **5**

1.1.1 Prévalence et incidence de l'infection 5

1.1.2 Modes de transmission 6

1.1.3 Vaccination 7

1.2 Le virus de l'hépatite B **7**

1.2.1 Identification 7

1.2.2 Taxonomie 8

1.2.3 La particule virale 10

1.2.4 Organisation du génome 11

1.2.4.1 La polymérase/transcriptase inverse 12

1.2.4.2 Les protéines Core et Précore 13

1.2.4.3 Les protéines de surface 15

1.2.4.4 La protéine HBx 18

1.2.4.5 La protéine HBSP 18

1.2.4.6 Les séquences régulatrices 19

1.2.4.7 Les éléments structuraux 19

1.2.5 Le cycle de réplication 20

1.2.5.1 L'entrée virale 21

1.2.5.2 La conversion de l'ADN-RC à l'ADNccc 22

1.2.5.3 La transcription 22

1.2.5.4 La réplication 23

1.2.6 Variations génétiques 25

1.2.6.1 Variabilité génotypique 26

1.2.6.1.1 Sérotypes 26

1.2.6.1.2 Génotypes 27
1.2.6.2 Variabilité phénotypique 30
1.2.6.2.1 Les mutants de l'AgHBe 31
1.2.6.2.2 Les mutants PréS1 et PréS2 31
1.2.6.2.3 Les mutants de l'AgHBs 32
1.2.6.2.4 Les mutants de la Polymérase 32
1.3 L'hépatite B : la pathologie 32
1.3.1 Symptômes 32
1.3.2 Pathogénie 32
1.3.3 Diagnostic 34
1.3.4 Les traitements contre l'hépatite B 36
1.3.4.1 L'interféron alpha 36
1.3.4.2 Les inhibiteurs de polymérase : analogues de nucléos(t)ides (NA) 36
1.3.4.3 La résistance aux NA 38
1.3.4.3.1 Résistance associée aux L-nucléosides 39
1.3.4.3.2 Résistance associée aux cyclopentanes 39
1.3.4.3.3 Résistance associée aux phosphonates acycliques 40
1.3.4.3.4 Résistance croisée 40
1.3.4.3.5 Impact sur les gènes de surface 42
1.4 Bases de données biologiques généralistes 43
1.4.1 Bases de séquences nucléotidiques 45
1.4.2 Bases de séquences protéiques 48
1.4.3 Bases de structures de macromolécules biologiques 49
1.5 Outils bioinformatiques 53
1.5.1 Recherche d'homologie 54
1.5.1.1 Alignement de deux séquences 54
1.5.1.1.1 Matrices de substitutions 55
1.5.1.1.2 Pénalités pour les insertions et délétions (pénalités de gaps) 56
1.5.1.1.3 Signification statistique 57
1.5.1.2 Algorithmes de recherche de séquences similaires dans une banque de séquences 58
1.5.1.2.1 FASTA 58
1.5.1.2.2 BLAST 59
1.5.1.3 Méthodes des profils 60
1.5.1.3.1 L'algorithme PSI-BLAST 61
1.5.1.3.2 Le logiciel HMMER 61
1.5.2 Alignements multiples 63
1.5.2.1 Clustal W 63
1.5.2.2 MUSCLE 64

1.5.2.3 Entropie de Shannon et logos de séquences 64
1.5.3 Prédiction de la structure secondaire des protéines 65
1.5.3.1 La méthode DSC 66
1.5.3.2 Les méthodes SOPM et SOPMA 66
1.5.3.3 La méthode PHD 67
1.5.4 Prédiction de la structure tertiaire des protéines 67
1.6 Bases de données et outils spécialisés pour le virus de l'hépatite B 68
1.6.1 Bases de données VHB et outils associés 68
1.6.1.1 HepSEQ (Royaume-Uni) 68
1.6.1.2 Hepatitis Virus Database (Japon) 69
1.6.1.3 HBVRegDB (Nouvelle-Zélande) 70
1.6.1.4 HBVrtDB (Etats-Unis) 70
1.6.1.5 SeqHepB (Australie) 71
1.6.2 Outils de génotypage pour les séquences du VHB 72
1.6.2.1 NCBI Genotyping Tool 72
1.6.2.2 Oxford Subtyping Tool 73
1.6.2.3 jp-HMM HBV 73
1.6.2.4 HBV STAR 75
Chapitre 2 : La base de connaissances HBVdb 77
2.1 Introduction 79
2.2 Matériel et méthodes 80
2.2.1 Cluster de calcul et gestionnaire de ressources 80
2.2.2 PostgreSQL 80
2.2.3 Schémas relationnels des bases de données 81
2.2.4 Java 82
2.2.5 HTML 82
2.2.6 Servlets Java 83
2.2.7 La librairie ISA 83
2.3 Résultats 84
2.3.1 Processus de génération de HBVdb 84
2.3.1.1 La base hbvdbembl 85
2.3.1.2 Les bases maîtresses 85
2.3.1.3 La base de référence hbvdbref 86
2.3.1.4 La base hbvdb 86
2.3.1.5 Parallélisation 87
2.3.1.6 Contrôle des étapes 88
2.3.1.7 Mise à disposition de la base 88

2.3.2 Processus d'annotation automatique 89
2.3.2.1 Problèmes liés à la circularité du génome du VHB 89
2.3.2.2 Recherche du génome de référence le plus proche 90
2.3.2.3 Réarrangement de la séquence requête 91
2.3.2.4 Annotation du feature source 92
2.3.2.5 Cartographie des séquences codantes (CDS) et optimisation des alignements 93
2.3.2.6 Transfert d'annotations 94
2.3.2.6.1 Les séquences codantes 94
2.3.2.6.2 Les protéines 95
2.3.3 Processus de génotypage 97
2.3.4 Processus de détection de profils de résistance aux drogues 100
2.3.5 Interface web 102
2.3.5.1 Menu HBV 102
2.3.5.2 Menu Query 103
2.3.5.3 Menu Analysis 104
2.3.5.3.1 Les outils génériques 104
2.3.5.3.2 Les outils spécialisés pour l'analyse des séquences du VHB 105
2.3.5.4 Menu HBVdb 108
2.3.5.5 Menu Links 108
2.3.6 Statistiques 109
2.4 Discussion – Perspectives 111

Chapitre 3 : Analyses de séquences de virus des hépatites B et C 115
3.1 Introduction 117
3.2 Matériel et méthodes 118
3.2.1 Analyse qualitative et quantitative de la variabilité 118
3.2.2 Extraction des séquences de HBVdb et alignements multiples 119
3.2.3 Recherche d'empreinte pour la modélisation par homologie 119
3.3 Résultats et discussion 120
3.3.1 Analyses de séquences sur la glycoprotéine E2 du virus de l'hépatite C 120
3.3.2 Vers une modélisation du domaine RNase H du VHB 124
3.3.2.1 Recherche d'homologues pour la sélection d'une empreinte 126
3.3.2.2 Analyse de la variabilité de la RNase H 130
3.3.2.2.1 Positions présentant des variations 130
3.3.2.2.1.1 Etude des séquences de tous les génotypes à partir de HBVdb 130
3.3.2.2.1.2 Etude des séquences de génotype D issues de séquençage à haut débit 131
3.3.2.2.2 Analyse quantitative de la variabilité de la RNase H par l'entropie de Shannon 133
3.3.2.2.3 La tétrade catalytique 135
3.3.3 Vers une modélisation moléculaire des domaines RT et RNase H du VHB 138

3.3.3.1 Alignement multiple de polymérases du VHB 138
3.3.3.2 Recherche d'homologues pour la sélection d'une empreinte 140

Chapitre 4 : Modélisation moléculaire des domaines RT et RNase H de la polymérase du VHB 143

4.1 Introduction 145
4.2 Matériel et méthodes 147
4.2.1 Alignements et optimisations 147
4.2.2 Modélisation moléculaire par homologie 147
4.2.3 Superpositions structurales 148
4.2.4 Cartographie des positions variables sur le modèle 148
4.2.5 Analyses d'interactions avec un ligand 149
4.3 Résultats 149
4.3.1 La modélisation du domaine RNase H 149
4.3.1.1 L'alignement des RNases H du VHB et d'E. coli 149
4.3.1.2 La modélisation moléculaire du domaine RNase H du VHB 151
4.3.1.3 Superposition du modèle avec l'empreinte 152
4.3.1.4 Cartographie des positions variables de la RNase H sur le modèle 154
4.3.2 La modélisation des domaines RT et RNase H 155
4.3.2.1 L'alignement des domaines RT et RNase H du VHB et du VIH-1 156
4.3.2.2 La modélisation moléculaire des domaines RT et RNase H du VHB 160
4.3.2.3 Comparaisons structurales - Superpositions 161
4.3.2.4 Analyses de la poche catalytique 164
4.3.2.4.1 Cartographie des mutations de résistance 164
4.3.2.4.2 Interactions avec le ténofovir 165
4.3.2.5 Conception de mutants à partir des analyses sur le modèle 168
4.4 Discussion et perspectives 170
4.4.1 Le modèle de RNase H 170
4.4.2 Le modèle RT-RNase H 172

Conclusion 175

Références bibliographiques 177

Références personnelles 205

Annexes 207

Articles 243

Table des figures

Figure 1 : Répartition géographique de la prévalence de l'hépatite B. 6
Figure 2 : Particules virales du virus de l'hépatite B. 8
Figure 3 : Arbre phylogénétique des orthohepadnavirus. 9
Figure 4 : Représentation schématique d'un virion VHB et des particules subvirales. 10
Figure 5 : Génome circulaire du VHB. 11
Figure 6 : Les quatre domaines de la polymérase du VHB, avec les sous-domaines de Pol/RT. 13
Figure 7 : Structure tridimensionnelle de la protéine core du VHB (PDB : 1qgt). 14
Figure 8 : Cadre de lecture S sur le génome circulaire et les 3 protéines de surface. 16
Figure 9 : Topologie transmembranaire des protéines de surface. 17
Figure 10 : Le cycle de réplication du VHB dans la cellule hôte. 21
Figure 11 : Structure de la tige boucle ε. 23
Figure 12 : Modèle de la transcription inverse de l'ARNpg en ADN-RC. 25
Figure 13 : Répartition géographique des génotypes du VHB. 28
Figure 14 : Evolution de la concentration des marqueur sérologiques au cours de l'infection aigue par le VHB. 35
Figure 15 : Formules chimiques des analogues de nucléos(t)ides. 38
Figure 16 : Evolution du nombre de projets de séquençage de génomes complets. 44
Figure 17 : Croissance de l'EMBL-Bank. 45
Figure 18 : Qualifiers optionnels proposés par l'EMBL-Bank pour le feature CDS. 48
Figure 19 : Schéma représentant la recherche à l'aide d'un profil HMM. 62
Figure 20 : Page d'accueil de la base HepSEQ. 69
Figure 21 : Graphique de résultat du génotypage par l'outil du NBCI. 73
Figure 22 : Résultat du génotypage par jpHMM. 74
Figure 23 : Schéma relationnel d'une base de données au format EMBL. 81
Figure 24 : Processus de génération de la base de données HBVdb. 87
Figure 25 : Problèmes liés à la circularité du génome et organisation d'un génome dupliqué. 90
Figure 26 : Fonctionnement de l'algorithme cut&paste 92
Figure 27 : Schéma descriptif du processus d'annotation automatique. 96
Figure 28 : Une partie de l'entrée HBVdb annotée X02763. 97
Figure 29 : Schéma descriptif du processus de génotypage. 99
Figure 30 : Schéma descriptif du processus de détection des profils de résistance aux NA. 101
Figure 31 : Page web HBVdb des jeux de données protéiques. 104
Figure 32 : Table principale de résultats de l'outil Genotype et page détaillée des annotations. 107
Figure 33 : Nombre d'entrées de chaque génotype. 109
Figure 34 : Nombre d'entrées par release de HBVdb. 110
Figure 35 : Statistiques d'utilisation du site web HBVdb depuis juin 2012. 110
Figure 36 : Population virale des variants du VHC avant transplantation, et 7 jours après. 121

Figure 37 : Pourcentage des résidus observés aux positions 447, 458 et 478. ____ *122*
Figure 38 : Organisation de la RNase H de type 1 de E. coli présentant la tétrade catalytique. ____ *125*
Figure 39 : Superposition de 3 structures expérimentales de RNases H. ____ *126*
Figure 40 : Alignement structural de RNases H virales, procaryote, eucaryote, et d'Archae. ____ *127*
Figure 41 : RNase H du VHB alignée avec l'alignement structural de celles de MoMLV et E. coli. ____ *129*
Figure 42 : Diagramme de Venn des positions de la RNase H présentant des variations ____ *132*
Figure 43 : Entropie de Shannon normalisée à chaque position de la RNase H et pour chaque génotype du VHB. ____ *134*
Figure 44 : Entropie de Shannon normalisée à chaque position de la RNase H et pour les 4 jeux de données. ____ *135*
Figure 45 : Histogrammes des fréquences des résidus aux positions 807 et 817 ____ *137*
Figure 46 : Alignement montrant l'insertion dans les séquences des orthohepadnavirus. ____ *140*
Figure 47 : Alignement des séquences de RNases H de E. coli et du VHB. ____ *150*
Figure 48 : Superposition de la structure de la RNase H de E. coli et du modèle VHB sélectionné ____ *153*
Figure 49 : Cartographie des positions variables sur le modèle de RNase H du VHB. ____ *155*
Figure 50 : Alignement de la séquence du VHB à modéliser avec la séquence du VIH-1 ____ *158*
Figure 51 : Modèles des domaines RT-RNase H de la polymérase du VHB. ____ *161*
Figure 52 : Superposition du modèle avec l'empreinte 1T05. ____ *162*
Figure 53 : Cartographie des positions de mutations de résistance connues dans le domaine RT du modèle. *165*
Figure 54 : Interactions entre le ténofovir et les résidus de la structure du VIH-1 et des 3 modèles du VHB. *167*
Figure 55 : Mes travaux de thèse intégrés à ma vision simplifiée de la bioinformatique. ____ *176*

Table des tableaux

Tableau 1 : Sérotypes et acides aminés correspondants dans l'AgHBS. 27
Tableau 2 : Caractéristiques des génotypes du VHB. 29
Tableau 3 : Profils de substitutions et résistance aux analogues de nucléos(t)ides. 41
Tableau 4 : Mutations de résistance dans le domaine RT de la Pol et les substitutions correspondantes dans l'AgHBs. 42
Tableau 5 : Classes et divisions de la base EMBL et le nombres d'entrées correspondantes. 46
Tableau 6 : Composition de la Protein Data Bank. 50
Tableau 7 : Comparaison des résultats des outils de génotypage du VHB. 100
Tableau 8 : Nombre d'entrées présentant les mat_peptide et CDS spécifiés dans les colonnes. 109
Tableau 9 : Positions de la RNase H présentant des variations, pour chaque génotype. 130
Tableau 10 : Positions présentant des variations spécifiquement dans un génotype. 131
Tableau 11 : Valeurs des critères de sélection pour les 10 modèles de RNase H générés. 152
Tableau 12 : RMSD (en Å) entre les différents modèles de RT 163
Tableau 13 : Liste des résidus de 1T05 et des 3 modèles interagissant avec le ténofovir. 166
Tableau 14 : Résultats préliminaires pour les mutants des 5 positions interagissant avec le TFV 169

Table des annexes

Annexe 1 : Exemple d'entrée ENA (AC=X02763). 211
Annexe 2 : Schéma relationnel de la base de données metadb. 212
Annexe 3 : Tableau des 16 génomes complets de référence (image de HBVdb). 213
Annexe 4 : Tableau des PRABI_name utilisés pour annoter les CDS et les mat_peptides dans HBVdb 214
Annexe 5 : Entrée annotée HBVdb X02763. 218
Annexe 6 : Fichier de sortie du processus de genotypage GenotypeHBV. 220
Annexe 7 : Fichier de sortie du processus de détection des profils de résistance FindHBVRMut. 221
Annexe 8 : Nombre de séquences protéiques (mat_peptides) complètes pour chaque génotype. 222
Annexe 9 : Liste des positions de la RNase H avec une entropie de Shannon normalisée > 0,1 223
Annexe 10 : Alignement multiple des 66 séquences de polymérase d'hepadnavirus 224
Annexe 11 : Cartographie sur le modèle de RNase H de l'entropie de Shannon normalisée pour les 4 jeux de données. 239
Annexe 12 : Alignement des séquences de RT VIH et VHB de Das, Daga et Bartholomeusz. 241
Annexe 13 : Valeurs des critères de sélection pour les 10 modèles de RT-RNase H générés. 242

Abréviations

3TC : Lamivudine
aa : Acide Aminé
ADN-RC : ADN Relaxed Circular
ADNccc : ADN Covalently Closed Circular
ADV : Adéfovir
ARNpg : ARN prégénomique
AgHBc : Antigène HBc
AgHBe : Antigène HBe
AgHBs : Antigène HBs
BLAST : Basic Local Alignment Search Tool
CDS : Coding Sequence
CHB : Chronic Hépatitis B
DHBV : Duck Hepatitis B Virus
ENA : European Nucleotide Archive
ETV : Entecavir
FT : Feature
GSHV : Ground Squirrel Hepatitis Virus
HBSP : Hepatitis B Spliced Protein
HBVdb : Hepatitis B Virus Database
HBx : Hepatitis B X protein
HHBV : Heron Hepatitis B Virus
HMM : Hidden Markov Model
HSP : High-scoring Segment Pair
IFN-α : Interféron α
INSDC : International Nucleotide Sequence Database Collaboration
ISA : Integrated Sequence Analysis
LdT : Telbivudine
LHBs : Large Hepatitis B Surface protein
LMV : Lamivudine
MHBs : Middle Hepatitis B Surface protein
MoMLV : Moloney Murine Leukemia Virus
NA: Nucleos(ti)de Analogue
nt : Nucléotide
ORF : Open Reading Frame
PBC : Promoter Basal Core
PDB : Protein Data Bank
Pol : Polymérase
PSI-BLAST : Position Specific Iterative BLAST
PSSM : Position Specific Scoring Matrix
RMN : Résonnance Magnétique Nucléaire
RMSD : Root Mean Square Deviation
RNase H : Ribonuclease H
RT : Reverse Transcriptase
SHBs : Small Hepatitis B Surface protein
SQL : Structured Query Language
SVN : Subversion
TFV : Tenofovir
TP : Terminal Protein
UDPS : Ultra Deep Pyrosequencing
VHB : Virus de l'Hépatite B
VHC : Virus de l'Hépatite C
VIH : Virus de l'Immunodéficience Humaine
WHV : Woodchuck Hepatitis Virus
XMRV : Xenotropic Murine leukemia virus-Related Virus

Introduction

L'hépatite B est un enjeu majeur de santé publique. En effet, malgré l'existence d'un vaccin depuis les années 1980, on dénombre plus de 350 millions de personnes infectées et ayant développée une maladie chronique du foie, à travers le monde. Une hépatite chronique augmente grandement les risques pour le patient de développer une cirrhose puis un carcinome hépatique.

Le virus de l'hépatite B (VHB) est un virus à ADN circulaire partiellement double brin. Son génome possède quatre cadres ouverts de lectures chevauchants codant pour 7 protéines principales. Il utilise une phase de transcription inverse pour se répliquer, grâce à un domaine transcriptase inverse (RT) dans la polymérase. Ce domaine constitue d'ailleurs la cible principale d'un type de traitements utilisés à l'heure actuelle contre le VHB, les analogues de nucléos(t)ides, qui inhibent la RT. Malheureusement, l'utilisation prolongée de certaines de ces molécules a conduit à l'émergence de mutants du VHB résistants aux traitements.

La variabilité génétique du VHB a conduit à classifier les génomes selon 8 génotypes (de A à H), défini sur la base d'une variation nucléotidique de plus de 8% (en pourcentage d'identité sur génomes complets). Les huit génotypes du VHB présentent une distribution géographique distincte.

Afin de permettre aux chercheurs d'étudier la variabilité génétique des séquences HBV et la résistance virale au traitement, nous avons décidé de développer une base de connaissances des séquences du VHB, HBVdb. L'objectif est de récupérer toutes les séquences du VHB des bases publiques afin de les annoter de manière automatique, à partir d'un ensemble de génomes de référence annotés manuellement et génotypés. Ce processus permettra d'annoter toutes les caractéristiques des séquences, avec un vocabulaire standardisé.

Cette base de connaissances devra être disponible via une interface web permettant à l'utilisateur d'extraire des jeux de séquences protéiques et nucléotidiques, et également d'analyser ses propres séquences. L'interface devra intégrer des outils génériques d'analyses bioinformatiques, ainsi que des outils spécialisés pour le VHB,

permettant à l'utilisateur d'annoter et de génotyper ses séquences, ainsi que détecter les profils de résistance aux traitements dans ses séquences.

Dans un deuxième temps, nous nous sommes intéressés à la protéine portant les principales mutations qui confèrent des résistance aux traitements : la polymérase. Ces mutations sont principalement situées dans le domaine RT, puisque les traitements ciblent ce domaine. Cependant, cette protéine possède un autre domaine enzymatique très important pour la réplication du virus : la RNase H, qui permet la dégradation de l'ARN matrice utilisé par la RT pour la transcription inverse. Ce domaine, situé en C-terminal du domaine RT, pourrait être une cible intéressante pour le développement de nouvelles drogues anti-VHB. Or, pour développer des molécules inhibitrices spécifiques, il est préférable de connaître la structure tridimensionnelle de la protéine. Malheureusement, la structure de la polymérase n'a pas encore été résolue expérimentalement. Pour pallier ce manque de structure 3D, nous avons décidé de modéliser par homologie la structure de la RNase H, afin de déterminer les caractéristiques structurales et fonctionnelles principales de ce domaine.

Enfin pour la placer dans son contexte structural, et pour avoir une idée globale de la structure, nous avons décidé de construire un modèle moléculaire de la polymérase, incluant les 2 domaines RT et RNase H.

Ce manuscrit se divise en quatre chapitres principaux, dont le premier est une revue des connaissances dans le domaine du virus de l'hépatite B et dans celui de la bioinformatique. Les trois chapitres suivants présentent les travaux effectués, sous la forme d'articles, avec pour chaque chapitre les parties : introduction, matériel et méthodes, résultats et discussion.

Le chapitre 2 présente le travail effectué pour le développement de la base de données HBVdb. Dans le chapitre 3, je présenterai les analyses de séquences réalisées sur des séquences de virus des hépatites B et C, notamment à partir de HBVdb. Certaines de ces analyses ont été utilisées pour permettre la construction des deux modèles moléculaires cités précédemment, et qui seront développés dans le chapitre 4.

Chapitre 1 :

Rappels bibliographiques

1.1 L'hépatite B : Epidémiologie

L'hépatite B est une infection du foie par le Virus de l'Hépatite B (VHB), potentiellement mortelle. L'Organisation Mondiale de la Santé (OMS) estime à 2 milliards le nombre de personnes infectées dans le monde, et de 350 à 400 millions le nombre de personnes souffrant d'une infection chronique. Parmi ces derniers, 15 à 25% développeront une complication hépatique mortelle telle qu'une cirrhose ou un carcinome hépatocellulaire (Lavanchy, 2004).

1.1.1 Prévalence et incidence de l'infection

Malgré la mise au point de vaccins efficaces, l'infection par le VHB reste un problème planétaire. De manière générale, l'incidence et la prévalence de la maladie sont inversement proportionnelle au développement économique. On peut distinguer trois zones géographiques en fonction de la prévalence (Figure 1) (Lavanchy, 2004) :

- une zone de faible endémie en Australie, Amérique du Nord et Europe de l'Ouest, du Nord et Centrale, où la prévalence de l'antigène HBs est inférieure à 2%.
- une zone de moyenne endémie en Europe de l'Est, Russie, Amérique du Sud au Proche Orient et dans les pays méditerranéens, où la prévalence de l'antigène HBs est comprise entre 2 et 7%.
- une zone de forte endémie en Asie du Sud-Est, Chine et Afrique subsaharienne, où la prévalence de l'antigène HBs est supérieure à 7% et peut atteindre 20%.

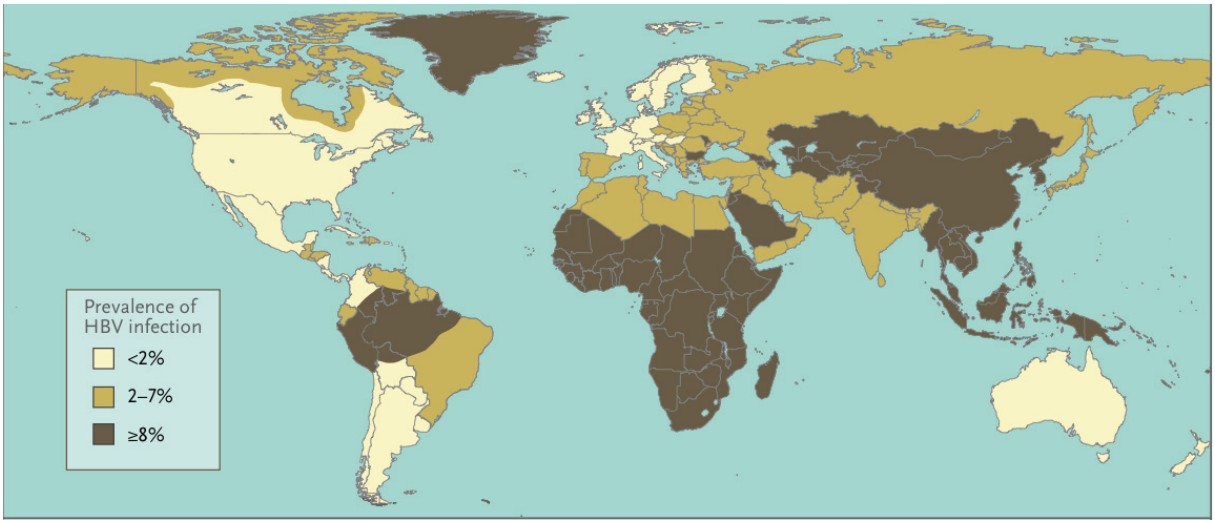


Figure 1 : Répartition géographique de la prévalence de l'hépatite B.
Les zones foncées correspondent à des zones où la prévalence de l'antigène HBs est supérieure ou égale à 8%, dans les zones en beige foncé la prévalence est comprise entre 2 et 7%, et dans les zones les plus claires, la prévalence est inférieure à 2% (Dienstag, 2008).

1.1.2 Modes de transmission

Le VHB est présent dans le sang, les secrétions sexuelles, la salive, et dans une moindre mesure le lait maternel, les larmes et l'urine d'une personne infectée. La transmission interhumaine du virus de l'hépatite B s'effectue par contact direct entre le sang de deux personnes ou par contact avec des sécrétions sexuelles d'une personne infectée. Les modes de transmission sont identiques à ceux du virus de l'immunodéficience humaine (VIH). De plus, le virus peut survivre à l'extérieur du corps pendant au moins 7 jours (Bond et al., 1981). Durant cette période, il reste capable d'infecter une personne non vaccinée. Tout comme pour la prévalence, les modes de transmission varient en fonction du niveau de développement économique. Dans les pays en voie de développement, le mode de transmission le plus fréquent reste le mode de transmission périnatal (de la mère à l'enfant pendant l'accouchement). La transmission du VHB d'un enfant à un autre est aussi très fréquente, et résulte souvent de contact de lésions cutanées ou de muqueuses avec du sang ou des secrétions de plaies. Le VHB peut également être transmis par contact avec la salive. Ces modes de transmission ont un impact clinique très important, car lorsque le VHB est transmis au cours des premières années de la vie, la probabilité d'un passage à chronicité est très forte. Dans les pays de moyenne endémie, la transmission est principalement périnatale ou horizontale (par contact étroit avec une personne infectée). La majorité des

infections recensées dans les pays développés, correspondant à des régions de faible endémie, sont contractées au stade de jeune adulte, à travers l'activité sexuelle ou la consommation de drogues injectables (Kane et al., 1999).

1.1.3 Vaccination

Un vaccin contre l'hépatite B existe depuis 1981. Au départ, il était recommandé pour les personnes appartenant à des groupes à haut risque telles que :

- les personnes ayant un comportement sexuel à haut risque,
- les partenaires ou personnes partageant le foyer de personnes infectées,
- les consommateurs de drogues injectables,
- les personnes fréquemment transfusées,
- les receveurs de transplantation d'organes,
- les personnes exposées à un risque professionnel d'infection (personnel de santé),
- les personnes voyageant dans des pays ayant un taux élevé d'infection.

Cette approche n'a pas permis de maitriser le taux d'infection dans la population générale. C'est pour cela qu'en 1992, l'OMS a recommandé d'introduire la vaccination universelle contre l'hépatite B dans les programmes nationaux de vaccination (Lavanchy, 2004).

Le vaccin présente une innocuité et une efficacité attestées remarquables. Dans bon nombre de pays où 8 à 15% des enfants devenaient des porteurs chroniques, la vaccination a permis de ramener le taux d'infection chronique à moins de 1% parmi les enfants vaccinés. En juillet 2011, 179 pays vaccinaient les nourrissons contre l'hépatite B dans le cadre de leur calendrier de vaccination – soit une augmentation substantielle de la couverture par rapport aux 31 pays concernés en 1992 (chiffres de l'Organisation Mondiale de la Santé).

1.2 Le virus de l'hépatite B

1.2.1 Identification

Le virus de l'hépatite B (VHB) a été identifié au début des années soixante par Baruch Blumberg. Il mis en évidence, dans des sera d'aborigènes australiens ainsi que

dans des sera de patients américains atteints de leucémie, une lipoprotéine inconnue qu'il nomma antigène « Australia » (Blumberg et al., 1965). Par la suite, il mis en évidence des anticorps précipitant dans des sera de patients ayant reçu plusieurs transfusions (Alter and Blumberg, 1966). Il découvrit que ces anticorps étaient dirigés contre l'antigène « Australia ». C'est en 1967 qu'il montra le lien entre l'antigène « Australia » et l'hépatite (Blumberg et al., 1967). L'antigène « Australia » correspond en réalité à l'antigène connu sous le nom de HBs actuellement (AgHBs). D.S. Dane *et al*. ont découvert la particule virale en 1970 par microscopie électronique (Dane et al., 1970) (Figure 2).

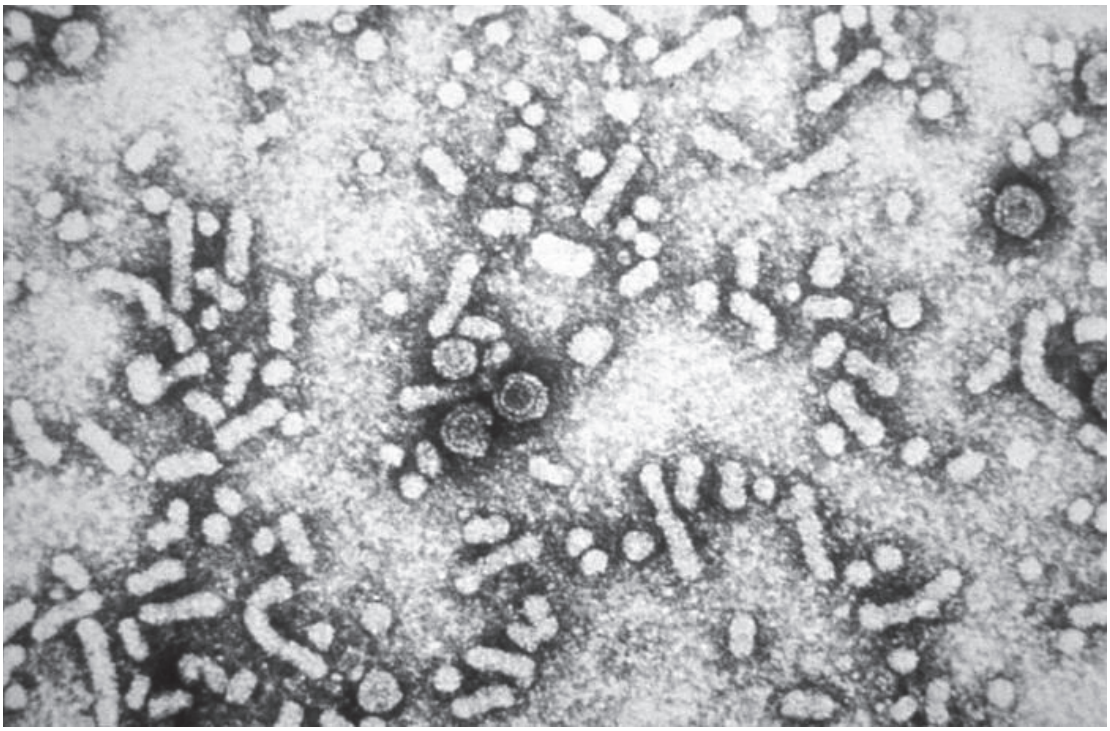


Figure 2 : Particules virales du virus de l'hépatite B.
Images de microscopie électronique (Center for Disease Control and Prevention)

1.2.2 Taxonomie

Le virus de l'hépatite B appartient à la famille des *hepadnaviridae* qui sont de petits virus enveloppés. C'est une famille de virus à ADN hépato-tropiques qui se divise en deux genres : les *orthohepadnavirus* qui infectent certains mammifères, et les *avihepadnavirus* qui infectent certains oiseaux (Schaefer, 2007). La famille des *orthohepadnavirus* comprend des virus de primates supérieurs (gorille, chimpanzé, gibbon, orang-outan) qui sont très similaires au VHB humain, mais aussi des virus de rongeurs, tels que celui de la marmotte (WHV) et de l'écureuil (GSHV), qui sont

similaires au VHB humain à plus de 80% (Figure 3) (Galibert et al., 1982). Chez les *avihepadnavirus*, on retrouve, par exemple, le virus du canard de Pékin (DHBV) et celui du héron (HHBV), similaires à environ 40% au VHB humain (Mandart et al., 1984). Tous les membres de cette famille peuvent causer des infections hépatiques chroniques ou aigües, ce qui, dans le cas du VHB humain, constitue un problème majeur de santé publique (Glebe and Urban, 2007).

Les *hepadnavirus* se répliquent par une phase de transcription inverse via un intermédiaire ARN, l'ARN pré-génomique (ARNpg). Mais ce qui les différencie des rétrovirus, c'est que le matériel génétique renfermé dans la capside est de l'ADN et non de l'ARN (Beck and Nassal, 2007).

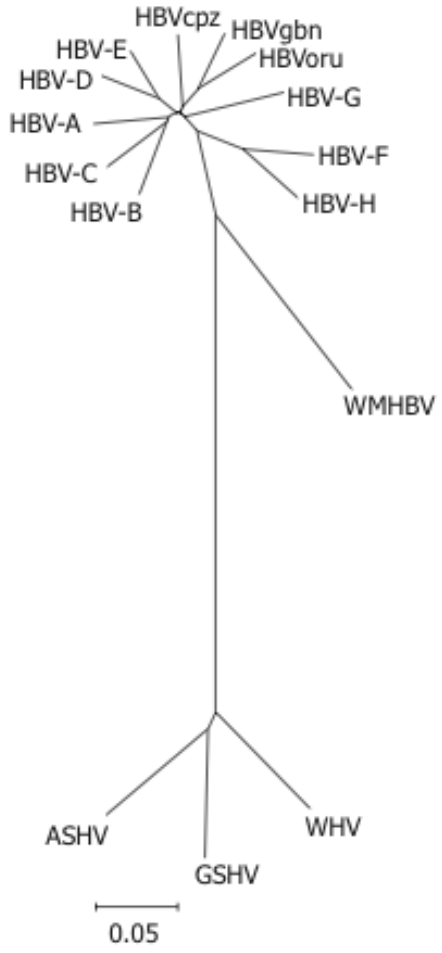


Figure 3 : Arbre phylogénétique des orthohepadnavirus.
(Schaefer, 2007)

1.2.3 La particule virale

La particule virale infectieuse du VHB, ou virion, consiste en une nucléocapside icosaédrique renfermant le génome viral et l'ADN polymérase. Cette nucléocapside, d'un diamètre approximatif de 30 nm, est enveloppée par une bicouche lipidique dans laquelle sont insérées les protéines de surface virale. Ces virions ont un diamètre de 40 nm et sont aussi appelés particules de Dane (Dane et al., 1970).

Une caractéristique de l'infection par un *hepadnavirus* est la sécrétion constitutive de particules d'enveloppes vides. Elles sont retrouvées dans le sérum sous forme de sphères de 22 nm de diamètre ou de bâtonnets qui sont aussi de 22 nm de diamètre mais de longueur variable (Figure 4) (Glebe and Urban, 2007). Ces particules vides (subvirales) sont sécrétées en large excès par rapport aux particules infectieuses (parfois plus de 1000 fois), et servent sans doute de leurres pour le système immunitaire.

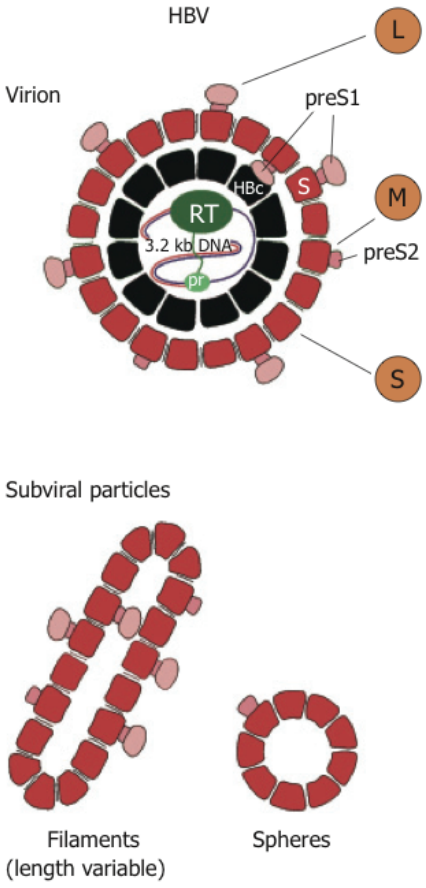


Figure 4 : Représentation schématique d'un virion VHB et des particules subvirales.
(Glebe and Urban, 2007)

1.2.4 Organisation du génome

Le génome du VHB est une molécule d'ADN circulaire partiellement double brin, non fermée de manière covalente, et dont la taille varie de 3182 à 3248 nucléotides. Cette molécule d'ADN, relâchée tout en étant circulaire, est appelée ADN-RC. C'est cette forme qui est renfermée par les virions (Seeger and Mason, 2000). Le brin moins ou brin L est complet et possède une courte redondance terminale en 5', à laquelle est attachée, de manière covalente, la polymérase virale. Le deuxième brin, brin plus ou S, est incomplet (environ 2400 nucléotides) avec l'extrémité 5' fixe et l'extrémité 3' libre. L'extrémité 5' du brin plus chevauche les 2 extrémités du brin moins, permettant ainsi d'assurer la circularité du génome (Figure 5).

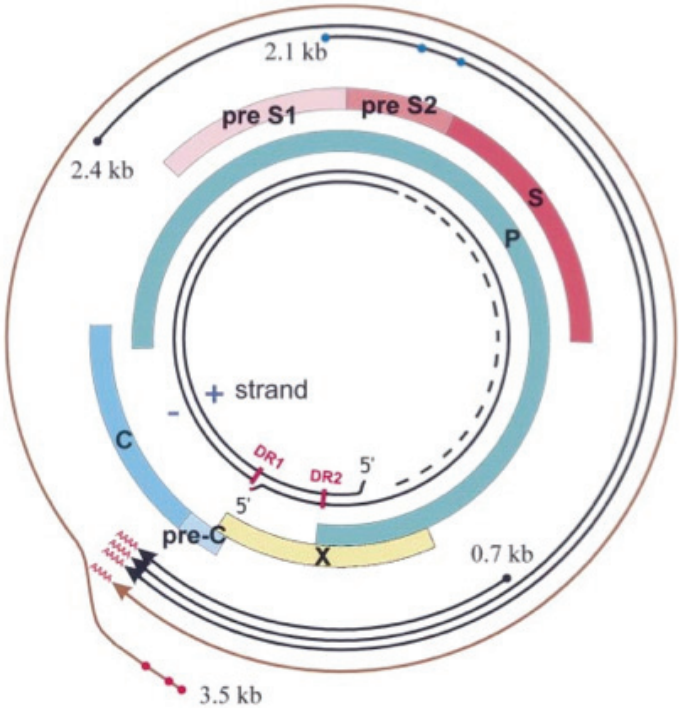


Figure 5 : Génome circulaire du VHB.
ADN partiellement double brin (brin – et brin +), avec les 4 cadres de lectures chevauchants, et les 4 transcrits. Le transcrit marron est l'ARNpg. (Kidd-Ljunggren et al., 2002)

Il existe également une forme circulaire close de manière covalente, l'ADNccc, qui est la forme réplicative, présente dans la cellule hôte au moment de la réplication. Le génome du VHB est optimisé au maximum puisque chaque nucléotide du génome est codant. En effet, le génome présente 4 cadres ouverts de lecture (ORF) qui se chevauchent (Nassal, 2008). Sept protéines, codées par ces 4 ORFs, ont été identifiées :

la polymérase, la protéine X (HBx), la protéine composant la capside ou core (HBc) ainsi que la protéine précore ou antigène HBe, et 3 protéines d'enveloppe : une large (LHBs), une moyenne (MHBs), une petite (SHBs) (Glebe and Urban, 2007). Ces protéines sont traduites à partir de 4 transcrits différents, dont le plus long est l'ARNpg (3,5 kb), qui sert également de matrice pour la transcription inverse.

1.2.4.1 La polymérase/transcriptase inverse

Le cadre de lecture P codant pour la polymérase (Pol) couvre quasiment 80% du génome du VHB. Pol est l'enzyme permettant la réplication du virus. Sa taille varie de 832 à 845 acides aminées (aa). Cette protéine a la particularité de posséder une double activité transcriptase inverse (ADN polymérase ARN-dépendante) et ADN polymérase ADN-dépendante. L'initiation de la transcription inverse et l'assemblage de la nucléocapside sont effectués par la polymérase (Summers and Mason, 1982; Bartenschlager and Schaller, 1992). Elle est composée de quatre domaines distincts (Figure 6).

Le premier domaine, situé à l'extrémité N-terminale, allant du résidu 1 à 183 environ, est appelé domaine protéine terminale (TP, terminal protein). Le domaine TP est utilisé comme amorce protéique pour initier la transcription inverse catalysée par le domaine RT (Wang and Seeger, 1992; Pollack and Ganem, 1994; Zoulim and Seeger, 1994). La synthèse du brin moins est amorcée par l'addition d'une molécule de dGTP au groupement hydroxyle d'un résidu tyrosine du domaine TP (résidu Y63). Cet amorçage protéique nécessite la reconnaissance et la liaison de la protéine à une structure tige-boucle, appelée epsilon (ε), située à l'extrémité 5' de l'ARNpg (précédemment transcrit à partir de l'ADNccc, voir paragraphe 1.2.5.3). Ainsi, la polymérase reste liée de manière covalente, par le domaine TP, à l'extrémité 5' du brin moins de l'ADN-RC. En plus de son rôle d'amorce et de polymérase pour la synthèse de l'ADN, Pol, grâce sa reconnaissance spécifique avec l'ARN ε, déclenche l'assemblage de la nucléocapside (Pollack and Ganem, 1994). Ceci conduit à l'incorporation spécifique de l'ARNpg et de Pol dans une nucléocapside prête à se répliquer.

Le deuxième domaine, allant des résidus 184 à 348, est un domaine charnière. Puisque le domaine TP est lié à l'extrémité 5' du brin moins, ce domaine charnière offre une flexibilité dans la protéine, permettant aux domaines suivants, qui portent des activités enzymatiques, de progresser dans la synthèse du brin.

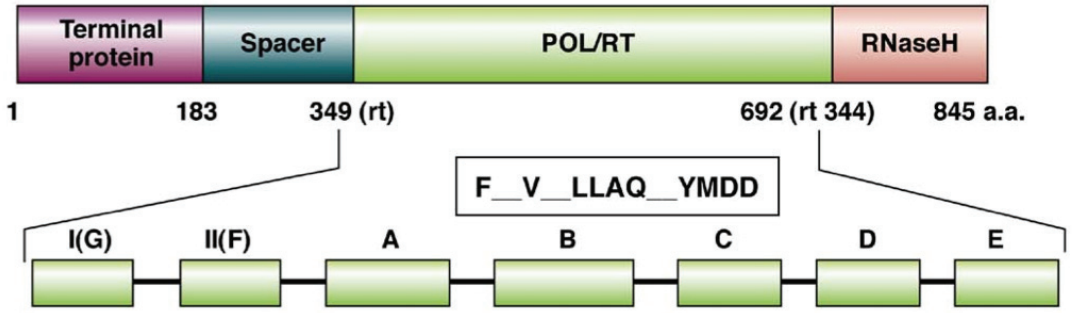


Figure 6 : Les quatre domaines de la polymérase du VHB, avec les sous-domaines de Pol/RT. (Zoulim and Locarnini, 2009).

Le troisième domaine (349-692) est celui portant les activités polymérase et transcriptase inverse. Par la suite, nous le désignerons par « domaine RT ». Ce domaine est très conservé et est subdivisé en 7 sous-domaines, désignés de A à G, correspondant à des blocs de séquences très conservées parmi les polymérases VHB (Stuyver et al., 2001; Bartholomeusz et al., 2004; Zoulim and Locarnini, 2009). Dans le sous-domaine C, on trouve un motif YMDD qui est très conservé chez les *hepadnavirus* et qui est aussi présent dans les séquences de transcriptases inverses des rétrovirus (Bartholomeusz et al., 2004). Il a été décidé d'une numérotation standard (Stuyver et al., 2001) ou nomenclature des résidus de ce domaine. Le domaine RT commence donc au niveau du motif très conservé EDWGPCDEHG, le premier résidu E étant désigné comme rt1. Le domaine RT s'étend du résidu rt1 au résidu rt344.

Le quatrième domaine, allant du résidu 693 à 845, correspond au domaine portant l'activité ribonucléase H (RNase H), qui dégrade l'ARN de l'hétéro-duplex ARN/ADN résultant de la transcription inverse.

1.2.4.2 Les protéines Core et Précore

La nucléocapside du VHB est composée de multiples copies d'une protéine : la protéine core ou protéine de capside ou antigène HBc (AgHBc), codée par le cadre de lecture C. La synthèse de la protéine core est initiée au deuxième codon ATG du cadre de lecture C pour former une protéine de 183 à 185 aa. Elle est divisée en deux domaines : le domaine N-terminal d'assemblage et le domaine fonctionnel en C-terminal. Le domaine N-terminal est nécessaire à la formation des particules de capside (Nassal, 1992). Le domaine C-terminal n'est pas indispensable pour l'assemblage mais est

important pour l'empaquetage de l'ARNpg et la synthèse d'ADN. Ce domaine possède une queue basique, riche en arginine, et contient de nombreux sites de phosphorylation, portés par des sérines et thréonines (Nguyen et al., 2008). Les propriétés de la capside et les interactions avec l'acide nucléique sont modulées par l'état de phosphorylation de la protéine core (Melegari et al., 2005). La structure cristalline de la protéine core, tronquée en C-terminal (queue basique), a été déterminée et est disponible à la Protein Data Bank (PDB) sous les codes 1QGT (Figure 7) (Wynne et al., 1999), 2G34 (Bourne et al., 2006) et 3KXS. (Packianathan et al., 2010)

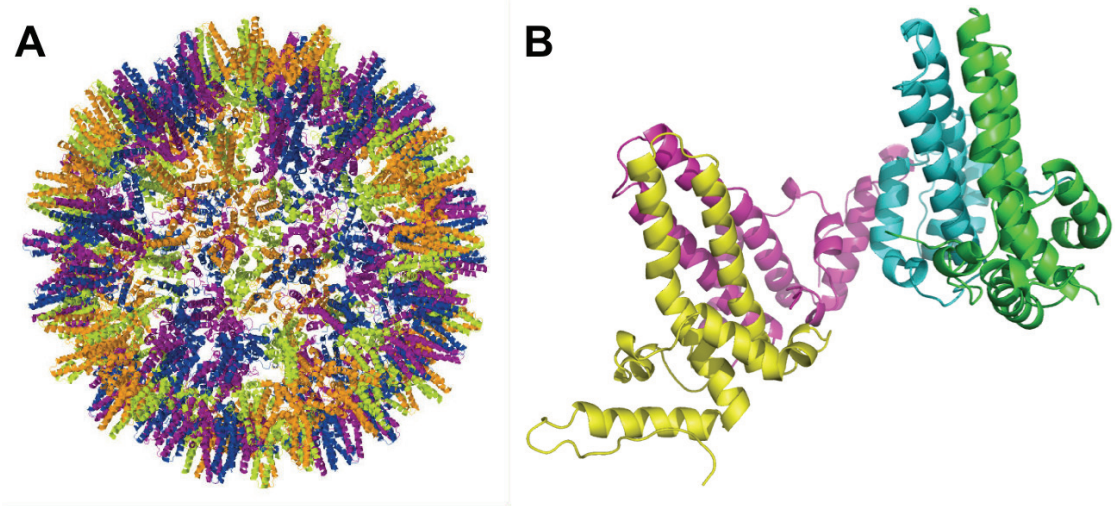


Figure 7 : Structure tridimensionnelle de la protéine core du VHB (PDB : 1qgt).
Assemblage de la protéine core en capside (modèle à 60 dimères). (B) Structure de 2 dimères core (A, B, et C, D). Coloration par chaîne.

L'assemblage commence par la formation d'homo-dimères (Zhou and Standring, 1992) liés par un pont disulfure entre les résidus cystéine 61 (Zheng et al., 1992). La capside finale est maintenue par des interactions faibles entre les dimères (Ceres and Zlotnick, 2002). Les particules se présentent sous deux formes : l'une a un diamètre de 30 nm et est constituée de 90 dimères disposés selon une symétrie icosaédrique T = 3, l'autre est un peu plus grande (34 nm de diamètre) et contient 120 dimères organisés avec une symétrie T = 4 (Crowther et al., 1994). Ces deux formes de particules peuvent être retrouvées dans un foie humain infecté (Kenney et al., 1995; Zlotnick et al., 1996).

La protéine précore est le précurseur de l'antigène HBe (AgHBe). Elle est synthétisée à partir du premier codon ATG du cadre de lecture C, et possède donc tous les résidus de la protéines core et 29 aa supplémentaires en N-terminal. Cette région est hydrophobe

et forme un peptide signal qui dirige la protéine vers le réticulum endoplasmique (RE). De là, la protéine transite par l'appareil de Golgi vers la surface cellulaire et, pendant le transport, le peptide signal et la queue basique sont éliminés. La protéine mature, AgHBe, est sécrétée sous forme de protéine monomérique et soluble, qui sert de leurre pour le système immunitaire (Bruss, 2004).

1.2.4.3 Les protéines de surface

Les virions et les particules vides ou subvirales contiennent des proportions variables des trois glycoprotéines de surface :

- la grande : LHBs ou PréS1,
- la moyenne : MHBs ou PréS2,
- la petite : SHBs, S ou antigène HBs (AgHBs) (Glebe and Urban, 2007).

Les protéines de surface du VHB sont les produits d'un cadre de lecture unique S codant pour trois domaines :

- le domaine préS1, d'une taille de 108 ou 119 aa, qui n'est présent que dans la grande protéine LHBs, aussi appelée PréS1,
- le domaine préS2, composé de 55 aa, qui est présent dans la grande protéine LHBs et dans la moyenne MHBs, aussi appelée PréS2,
- le domaine S, composé de 226 aa, est commun aux trois protéines de surface. La petite proteine SHBs, S ou AgHBs n'est composée que de ce domaine.

En réalité, ces 3 protéines codées par le carde de lecture S, sont le résultat de la transcription à partir de deux promoteurs : PS1 et PS2. La traduction de l'ARN messager (ARNm) transcrit à partir du promoteur PS1, mènera à la grande protéine LHBs. La transcription à partir du promoteur PS2 mène à deux transcrits de tailles différentes. Le plus long transcrit contient un codon initiateur, le codon 109 ou 120, et mène, après traduction à la protéine moyenne MHBs ; alors que le premier codon initiateur du transcrit plus court se trouve 55 codons plus loin, et la traduction de ce transcrit court mène à la petite protéine SHBs, S ou AgHBs (Bruss, 2004). Les protéines d'enveloppe sont donc co-carboxyterminales (Figure 8).

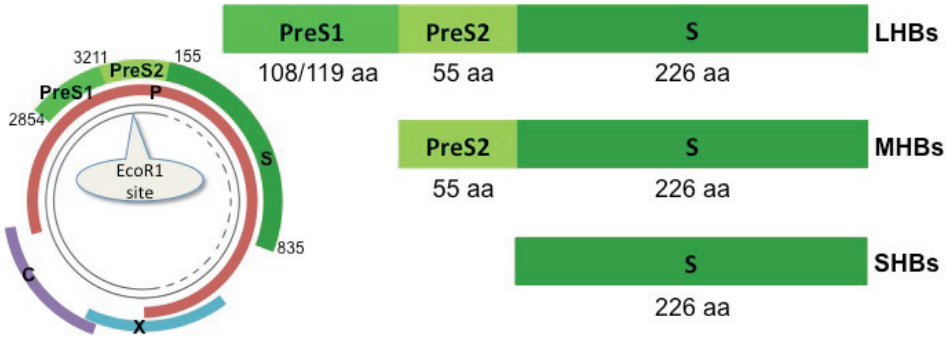


Figure 8 : Cadre de lecture S sur le génome circulaire et les 3 protéines de surface.
La numérotation est basée sur l'entrée X02763 de HBVdb.

L'AgHBs est le composant majeur de l'enveloppe virale et des vaccins anti-VHB. En effet, la plupart des anticorps neutralisants ciblent une partie de l'AgHBs qui est exposée à la surface des particules virales. La structure tridimensionnelle de l'AgHBs n'est pas connue. Cependant, on sait qu'elle présente une topologie transmembranaire complexe (Bruss, 2004). Elle possèderait 4 segments hydrophobes transmembranaires, et une large boucle hydrophile centrale, exposée à la surface de la particule virale (Figure 9).

Les protéines d'enveloppe sont synthétisées dans le RE. La translocation de la protéine S à travers la membrane du RE est initiée par un signal de type I en N-terminal, qui correspond à une séquence hydrophobe entre les résidus 8 et 22 de la protéine S, mais qui n'est pas clivé par une peptidase. Un second signal de type II, une séquence hydrophobe entre les résidus 80 et 98 de S, ancre la chaine dans la membrane du RE de telle manière que le domaine qui suit est orienté vers le lumen du RE (Eble et al., 1987).

Il est supposé que le signal N-terminal, qui n'a pas de fonction d'ancrage intrinsèque, est fixé passivement dans la membrane lipidique de telle sorte que la région comprise entre les résidus 23 et 79 forme une boucle du côté cytosolique de la membrane du RE. Les 57 résidus de l'extrémité C-terminale de S (aa 170 à 226) sont fortement hydrophobes et il est supposé que ce domaine est intégré dans la membrane du RE. Cette extrémité C-terminale de S est orientée vers la lumière du RE, car des domaines étrangers fusionnés à S sont exposés vers le lumen (Eble et al., 1987). La boucle dirigée vers le lumen (aa 99 à 169) située entre le second domaine transmembranaire et le domaine hydrophobe C-terminal porte l'épitope majeur de l'antigène HBs (AgHBs) qui est exposée sur la surface externe des particules virales et subvirales après bourgeonnement (Glebe and Urban, 2007; Bruss, 2007).

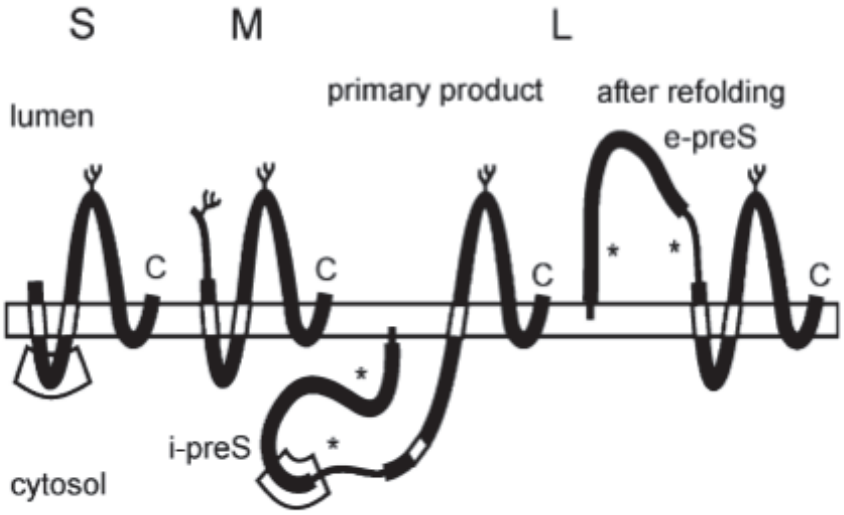


Figure 9 : Topologie transmembranaire des protéines de surface.
Les 3 protéines L, M et S sont représentées en fonction de leur ancrage dans la membrane du RE. Les deux topologies de LHBs sont représentées (i-preS et e-preS). Les signaux de type I et II sont représentés par les segments blancs. Les astérisques indiquent les sites potentiels de glycosylation. La région comportant le déterminant a est représentée par les régions encadrées. (Bruss, 2004)

Cette large région hydrophile est structurée en boucles par des ponts disulfures. Deux de ces boucles constituent le déterminant « a » (résidus 124 à 147), qui est la cible d'anticorps neutralisants importants. Il contient un site de glycosylation, mais qui n'est pas utilisé systématiquement. La région hydrophile semble jouer un rôle dans la reconnaissance des hépatocytes par les virions, mais l'acteur principal de cette reconnaissance semble être la protéine LHBs ou PréS1. Les particules vides, qui sont très appauvries en LHBs et MHBs, n'entrent donc pas en compétition avec les particules infectieuses pour les récepteurs hépatocytaires. La protéine MHBs ou PréS2 possède un site de glycosylation qui est systématiquement glycosylé. A l'extrémité N-terminale, la protéine LHBs possède un site de myristoylation, le myristoyle étant un acide gras qui sert à ancrer sa partie N-terminale à la membrane. Cette protéine peut donc adopter deux topologies membranaires. Initialement, les régions PréS1, PréS2 et le premier segment transmembranaire sont retenus du coté cytoplasmique (topologie i-preS), ou ils participent surement à l'enveloppement de la nucléocapside. Pendant le transport du virion vers la surface cellulaire, certaines protéines LHBs changent de conformation. Le premier segment transmembranaire s'insère dans la membrane, et les régions PréS1 et PréS2 s'externalisent (topologie e-preS). Ce changement de topologie est nécessaire car des séquences proches de l'extrémité N-terminale de la protéine LHBs jouent un rôle

dans la reconnaissance des hépatocytes par le virion (Bruss, 2007; Glebe and Urban, 2007).

1.2.4.4 La protéine HBx

La protéine HBx est la plus petite protéine du VHB. Elle est composée de 154 résidus, et son rôle exact est mal connu. Il a été montré qu'elle pourrait être un régulateur clé de la réplication virale, ainsi que des fonctions de la cellule hôte, en modulant une grande variété de processus cellulaires, y compris la transcription, la progression du cycle cellulaire, la réparation de l'ADN et l'apoptose (Benhenda et al., 2009; Tang et al., 2006). En effet, il a été suggéré qu'HBx pouvait transactiver une variété de promoteurs viraux et cellulaires, mais c'est une protéine transactivatrice relativement faible. De plus, elle ne se lie pas à l'ADN, donc son rôle dans la transactivation est certainement indirect (Bouchard and Schneider, 2004). HBx est essentielle pour initier et maintenir la réplication virale après infection (Lucifora et al., 2011). Il a été montré que HBx n'est pas empaquetée dans les virions mais elle est exprimée après infection dans la nouvelle cellule hôte, pour permettre le contrôle épigénétique de la transcription de l'ADN viral (Lucifora et al., 2011). Son rôle pourrait reposer sur la modification de l'environnement cellulaire par transactivation des gènes cellulaires dans les hépatocytes infectés, afin de faciliter la réplication virale. HBx peut également jouer un rôle dans la pathogenèse du carcinome hépatocellulaire induite par le virus (Koike, 2009).

1.2.4.5 La protéine HBSP

Les protéines du VHB sont codées par des ARN non épissés. Cependant, Soussan *et al*. (Soussan et al., 2000) ont montré qu'un ARN épissé du VHB codait pour une nouvelle protéine du VHB *in vivo*. Cette protéine du VHB générée par épissage (HBSP, Hepatitis B Spliced Protein) correspond à la fusion d'une partie de la polymérase virale et d'un nouveau cadre de lecture ouvert qui est créé par l'événement d'épissage. *In vivo*, la protéine HBSP a été trouvée dans des échantillons de foie infecté par le VHB et des anticorps anti-HBSP sont apparus dans un tiers des échantillons de sérum prélevés chez des porteurs chroniques du VHB. *In vitro*, l'expression ectopique de HBSP n'a aucun effet sur la réplication de l'ADN viral ou la transcription, mais induit l'apoptose cellulaire.

HBSP pourrait jouer un rôle dans l'histoire naturelle de l'infection par le VHB et pourrait être impliquée dans la pathogénie et / ou la persistance de l'infection par le VHB (Soussan et al., 2003). HBSP active la réponse immunitaire cellulaire par les cellules

T (Bayard et al., 2012). Il a également été montré que l'interaction de HBSP avec une protéine cellulaire, la Cathepsin B pouvait induire la motilité et l'invasion des cellules d'hépatome, dans les carcinomes hépatocellulaires (Chen et al., 2012).

1.2.4.6 Les séquences régulatrices

Le VHB possède 4 promoteurs : core, PréS1, S, X, et deux « enhanceurs » (Enh). L'Enh I se trouve en amont du cadre de lecture X et l'Enh II se trouve en amont du promoteur basal du core (PBC) et du cadre de lecture C. Cet Enh possède des sites de fixation pour des facteurs de transcription essentiellement hépatiques, tout comme le PBC. Ainsi, leur activité est maximale dans les cellules du foie. Le PBC contrôle la transcription de l'ARNm de la protéine PréC et de l'ARNpg qui est utilisé comme ARNm de la polymérase et de la protéine core. L'activité du PBC est modulée positivement par des éléments en *cis* appelés CURS (core upstream regulatory sequence) faisant partie de l'Enh II (Yuh et al., 1992) et négativement par des éléments régulateurs négatifs (NRE, negative regulatory element) situés en amont de l'Enh II (Moolla et al., 2002).

Le promoteur PréS1 initie la transcription de la protéine PréS1, et son activité est modulée par l'Enh I. Il est le seul promoteur du VHB à posséder une « TATA box », et son activité est régulée négativement par une boite « CAAT » située en aval, dans le promoteur S.

Le promoteur S contrôle la transcription de deux ARNm : celui de PréS2 et celui de l'AgHBs. Son activité est maximale dans le foie, car il est régulé positivement par les deux enhanceurs.

Le promoteur X est situé entre l'Enh I et le cadre de lecture X et sa régulation est mal connue. Il a été montré in vitro que c'est un promoteur fort, mais paradoxalement, dans le foie infecté, il est difficile de détecter l'ARNm de HBx (Moolla et al., 2002).

1.2.4.7 Les éléments structuraux

Le génome du VHB possède plusieurs éléments structuraux. Ces éléments comprennent : un signal d'encapsidation ε, deux répétitions directes DR1 et DR2, et un signal de polyadénylation. A l'exception de la DR2, tous ces éléments se trouvent sur une courte région du génome de moins de 100 nucléotides. Le signal d'encapsidation est constitué de deux tiges boucles superposées, et va permettre l'encapsidation simultanée de l'ARNpg et de Pol, et est également étroitement lié à l'initiation de la synthèse du brin moins par transcription inverse (Beck and Nassal, 2007).

La DR1, de 11 nucléotides, est située en 5' du signal d'encapsidation. Les répétitions DR1 et DR2 vont définir les extrémités 5' de l'ADN du brin moins et du brin plus respectivement. Le signal de polyadénylation est situé immédiatement en 3' du signal d'encapsidation ε (Nassal, 2008).

1.2.5 Le cycle de réplication

La réplication du génome du VHB peut être divisée en trois phases (Beck and Nassal, 2007) :

- les virions infectieux contiennent, dans leur capside icosaédrique, un génome d'ADN partiellement double brin, circulaire, mais non clos de façon covalente d'environ 3,2 kb de longueur (circulaire relâché ou ADN-RC),
- lors de l'infection, l'ADN-RC est converti, à l'intérieur du noyau de la cellule hôte, en un plasmide d'ADN circulaire fermé de manière covalente (ADNccc),
- à partir de l'ADNccc, plusieurs ARN génomiques et subgénomiques sont transcrits par l'ARN polymérase cellulaire II.

Ces transcrits sont traduits, notamment les protéines core qui vont ensuite s'assembler pour former les capsides.

Parallèlement, la protéine précore (AgHBe) et les protéines d'enveloppe, pendant ou après la traduction, sont dirigées vers le RE où la protéine précore est maturée et l'AgHBs s'assemble pour former des particules d'enveloppe vides. Après transport à travers l'appareil de Golgi vers la surface cellulaire, l'AgHBe et les particules vides sont sécrétées.

Parmi les transcrits, l'ARNpg est sélectivement emballé avec la Pol dans des capsides nouvellement générées, par traduction et assemblage des protéines core (HBc), et est inversement transcrit par la Pol en nouveau génome d'ADN-RC. Les nucléocapsides matures contenant de l'ADN-RC, mais pas les immatures contenant l'ARNpg, peuvent être recyclées vers le noyau, pour l'amplification intracellulaire du pool d'ADNccc, mais la majorité sera dirigée vers le RE où elles seront enveloppées (Beck and Nassal, 2007). Après transport vers la surface cellulaire et maturation du virion avec le changement de topologie de la protéine LHBs (Bruss, 2007), les particules infectieuses seront secrétées à leur tour par la cellule (Figure 10).

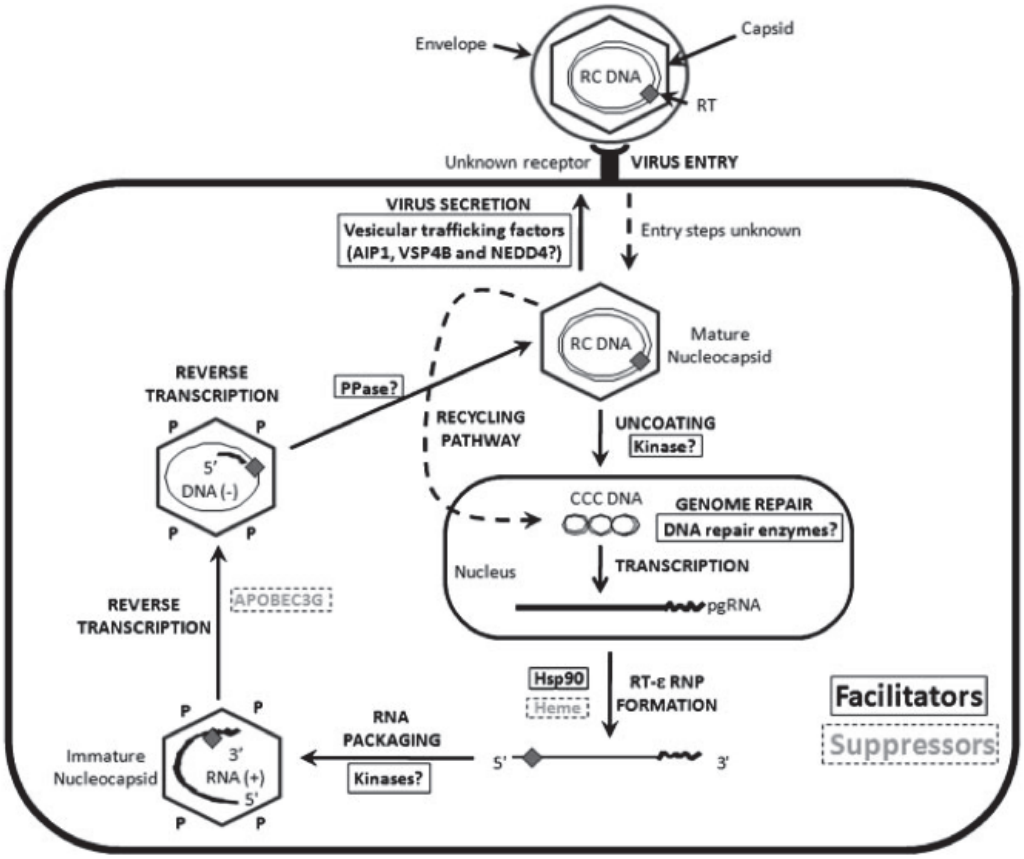


Figure 10 : Le cycle de réplication du VHB dans la cellule hôte.
On peut voir l'entrée du virus dans la cellule hôte, la décapsidation, et l'ADN-RC entre dans le noyau pour être réparé et former l'ADNccc, qui sera transcrit en ARNpg. Puis, l'ARNpg forme un complexe avec la Pol (RT-ε RNP), qui est encapsidé. L'ARNpg est retro-transcrit pour former le nouveau génome d'ADN-RC. Les facteurs facilitateurs et suppresseurs sont indiqués dans les encadrés (P : Phosphates, Ppase : Phosphatase, Hsp 90 : heat-shock protein 90) (Nguyen et al., 2008).

1.2.5.1 L'entrée virale

Les premières étapes de l'infection par le VHB sont mal connues. Le virion reconnaît et s'attache à l'hépatocyte via un ou plusieurs récepteurs cellulaires qui restent inconnus à l'heure actuelle. Les résidus 21 à 47 de la protéine PréS1 (LHBs) sont fortement impliqués dans cette reconnaissance mais il semble que quelques séquences de la grande boucle hydrophile de l'AgHBs y participent aussi (Glebe and Urban, 2007). Ensuite, il y a une série d'événements mal élucidés, incluant probablement une étape de fusion entre l'enveloppe virale et la membrane cellulaire, qui aboutissent à la libération de la nucléocapside dans le cytoplasme de l'hépatocyte. Puis, cette dernière est dirigée vers le noyau grâce à des signaux de localisation cellulaire situés sur la protéine core (AgHBc) (Kann et al., 1999).

1.2.5.2 La conversion de l'ADN-RC à l'ADNccc

Il faut éliminer l'oligoribonucléotide qui se trouve à l'extrémité 5' du brin plus d'ADN, compléter la synthèse de ce brin plus, éliminer la redondance et la Pol qui se trouvent à l'extrémité 5' du brin moins, et enfin ligaturer les 2 brins de manière covalente. Le déroulement exact et la localisation de ces événements sont mal connus (Beck and Nassal, 2007). C'est ainsi que l'ADN-RC (ADN relâché circulaire) est converti en ADNccc (covalently closed circular), qui représente la forme réplicative du génome du VHB. L'ADNccc nucléaire, produit à partir de l'ADN-RC génomique entrant, sert de matrice pour la transcription par l'ARN polymérase cellulaire II. Par conséquent, tous les transcrits viraux, y compris le l'ARNpg d'environ 3,5 kb, portent une coiffe en 5' et une queue polyA en 3'.

L'ADNccc avec les histones cellulaires et d'autres protéines nucléaires forment des mini-chromosomes servant de matrice pour la transcription des ARNm viraux, qui seront exportés vers le cytoplasme pour être traduits (Bock et al., 2001).

La persistance de l'ADNccc dans des hépatocytes joue un rôle clé dans la persistance virale, la réactivation de la réplication virale après l'arrêt du traitement antiviral, et dans la résistance au traitement. L'ADNccc peut persister dans le noyau des hépatocytes aussi longtemps que les cellules infectées survivent. L'infection chronique des hépatocytes est maintenue par le pool d'ADNccc viral (Nassal, 2008; Liu et al., 2004; Lee et al., 2004).

1.2.5.3 La transcription

Le promoteur basal du core (PBC) contrôle la synthèse de deux transcrits : un initié en amont de l'ATG initiateur de la protéine précore, et l'autre immédiatement en aval de ce codon. Le premier est l'ARNm de la protéine précore et le deuxième est l'ARNpg. Ces deux transcrits sont plus longs que le génome (environ 3500 nucléotides), et comprennent une queue polyA et une redondance terminale d'environ 130 nucléotides. Cette redondance dans l'ARNpg est essentielle pour que le virus ne perde pas d'information génétique pendant la réplication (Beck and Nassal, 2007). L'ARNpg possède deux signaux d'encapsidation. L'ARNpg sert de matrice pour la traduction de la protéine de capside et pour la Pol.

Le promoteur PS1 initie un transcrit d'environ 2400 nucléotides, à l'origine de la protéine LHBs. Comme le PBC, le promoteur de PS2 peut initier des transcrits à des sites multiples : un en amont de l'ATG initiateur de PréS2 (MHBs), et un en aval de ce codon

pour générer l'AgHBs (SHBs). Le promoteur X contrôle la transcription d'un ARNm d'environ 700 nucléotides, qui code pour la protéine HBx (Nassal, 2008).

1.2.5.4 La réplication

La transcription inverse a lieu dans les nucléocapsides se trouvant dans le cytoplasme de la cellule infectée. Les deux premières étapes sont l'encapsidation de l'ARNpg avec la Pol, et l'initiation de la synthèse du brin moins d'ADN. Pol se lie préférentiellement à la molécule d'ARNpg à partir de laquelle elle a été traduite, en particulier à la structure en tige-boucle appelée ε (Figure 11) à l'extrémité 5' (Bartenschlager and Schaller, 1992; Pollack and Ganem, 1994; Nassal, 2008).

Cette liaison déclenche l'encapsidation du complexe dans une nouvelle nucléocapside. Elle déclenche également la transcription inverse par amorçage protéique réalisé par la tyrosine 63 du domaine TP de Pol, ce qui initie la synthèse du brin moins d'ADN en utilisant des nucléotides d'epsilon comme modèle. Un seul ARNpg et une seule protéine Pol sont conditionnés par particule. L'encapsidation sélective de l'ARNpg dans la capside icosaédrique est médiée par des interactions spécifiques avec la transcriptase inverse virale (RT) (Liu et al., 2004; Lee et al., 2004).

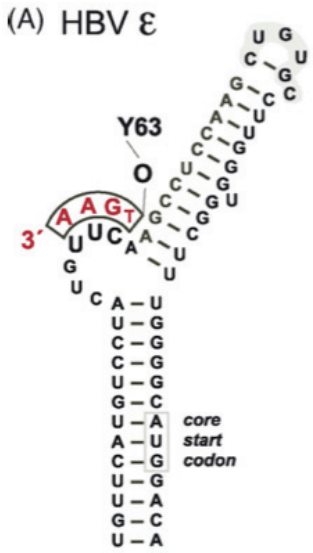


Figure 11 : Structure de la tige boucle ε.
Les bases en rouge représentent l'amorce du brin d'ADN, liée à la tyrosine 63 (Y63) du domaine TP de la Pol (Nassal, 2008).

Après la synthèse de trois ou quatre nucléotides au niveau d'epsilon (Figure 12 (A)), il y a translocation de la Pol et du brin moins d'ADN naissant au niveau d'une séquence complémentaire à DR1, qui se trouve près de l'extrémité 3' de l'ARNpg (Figure 12 (B)). La synthèse du brin moins ADN reprend à partir de cette position. Suite à l'allongement, le domaine RNase H de la protéine Pol dégrade le l'ARNpg modèle qui a été copié (Figure 12 (C)). La synthèse du brin moins se poursuit jusqu'à l'extrémité 5' de l'ARNpg. De cette synthèse résulte un brin moins de longueur complète, avec la protéine Pol toujours liée de façon covalente à son extrémité 5'. Le produit final de clivage par la RNase H est un fragment d'ARN court de 17 ou 18 nucléotides qui amorce l'initiation de la synthèse de brin d'ADN plus. L'extrémité 3' de l'amorce contient la séquence de DR1 (Figure 12 (D)).

Pour la synthèse de l'ADN-RC, deux changements de matrices sont nécessaires lors de la synthèse du brin plus d'ADN. Tout d'abord, il y a une translocation et l'amorce du brin plus s'hybride à la DR2 (Figure 12 (E)), qui se trouve près de l'extrémité 5 'du brin négatif. Ceci déclenche la synthèse du brin plus à partir de DR2 (Figure 12 (F)). La synthèse du brin plus se poursuit à l'extrémité 5' de l'ADN brin moins. Le brin plus naissant est soumis à un autre changement de matrice, la circularisation, pour permettre l'élongation du brin plus à partir de l'extrémité 3' du brin moins comme matrice (Figure 12 (G)). L'ADN-RC est ainsi formé (Figure 12 (H)).

Il existe une seconde voie de synthèse du brin plus d'ADN. Cette voie génère un duplex linéaire (DL) d'ADN parce que la synthèse du brin plus est amorcée à partir de DR1 sur le brin négatif (Figure 12 (I)). Il n'y a donc pas permutation de matrice lors de la synthèse du brin plus de l'ADN DL (Liu et al., 2004; Lee et al., 2004).

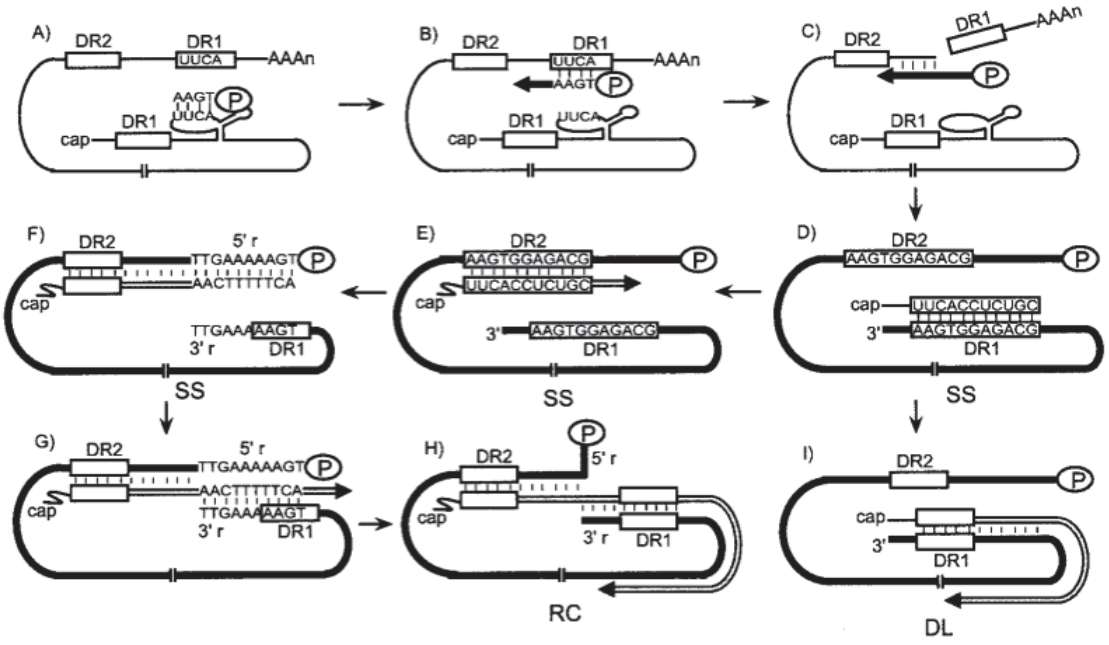


Figure 12 : Modèle de la transcription inverse de l'ARNpg en ADN-RC.
On distingue deux voies de synthèse : celle menant à l'ADN-RC et celle menant à un ADN linéaire double brin. (Liu et al., 2004)

1.2.6 Variations génétiques

Le taux de mutation est de l'ordre de 10^{-5} pour un cycle de réplication (Stuyver et al., 2001; Sanjuán et al., 2010; Margeridon-Thermet and Shafer, 2010). La Pol ne possède pas d'activité de « proof reading » permettant de vérifier la base qui vient d'être ajoutée, et la production journalière de virions est de 10^{11} (Hollinger, 2007).

Les mutations qui sont générées par transcription inverse sont initialement trouvées dans l'ADN-RC. Les mutations peuvent être transmises de façon stable : l'ADN-RC génomique muté est recruté dans le pool d'ADNccc du patient, soit par recyclage dans le noyau, soit par une infection d'un nouvel hépatocyte par le virus mutant. Dans les deux cas, le virus mutant sera en concurrence avec un grand nombre d'autres génomes viraux de type sauvage ou contenant d'autres mutations. L'émergence de mutants du VHB est donc beaucoup plus lente, même en présence d'une pression sélective, par rapport à d'autres virus tels que le VIH ou le VHC, où les mutations générées par des erreurs à l'étape de réplication, de transcription inverse ou de transcription, peuvent avoir des effets phénotypiques immédiats (Kay and Zoulim, 2007).

Une deuxième conséquence de la double nature du génome du VHB est la présence chez un patient chroniquement infecté de deux quasi-espèces. La première est la quasi-

espèce du pool d'ADNccc et la seconde est la quasi-espèce de l'ADN-RC qui reflète à la fois la quasi-espèce ADNccc et les nouvelles mutations générées au cours de la réplication.

Enfin, la diversité des génomes du VHB peut être divisée en deux catégories :

- une variabilité génotypique ou polymorphique, qui est le résultat d'une évolution progressive du génome en l'absence de pression sélective (accumulation de mutations au cours du temps qui ne changent pas les propriétés du virus)
- une variabilité phénotypique, qui résulte de la sélection de mutations qui altèrent les propriétés du virus face à une pression immunologique ou pharmacologique.

Avec la variation génotypique, le « fitness » viral (ou capacité de réplication) est le facteur le plus important. Un virus mutant qui est nettement moins bon que les autres VHB circulant finira par être éliminé, même si la l'ADN-RC génomique muté intègre le pool d'ADNccc. Avec la variation phénotypique, la force motrice est la sélection, puisque la capacité de résister à la pression antivirale l'emporte généralement sur le fitness réduit d'un mutant par rapport à un virus de type sauvage qui ne peut pas résister (Kay and Zoulim, 2007).

1.2.6.1 Variabilité génotypique

1.2.6.1.1 Sérotypes

Très rapidement après l'identification du VHB, il est devenu évident que des isolats du VHB ne réagissaient pas de la même façon avec des sérums provenant de différents patients anti-HBs positifs. Le premier classement des isolats de VHB a donc été fait par sérotypage, selon la réactivité de l'antigène HBs de l'isolat avec un panel d'antisérums standardisés (Couroucé et al., 1983). La région immunogène majeure, le déterminant «a» couvrant les résidus 124 à 147 de l'AgHBs, est commune à la quasi-totalité des isolats VHB et n'est donc pas informative pour la classification. La classification est donc faite en utilisant d'autres déterminants dont les plus importants sont deux couples de déterminants *d/y* et *r/w* *(Kay and Zoulim, 2007)*.

Ces deux déterminants sont composés de deux épitopes mutuellement exclusifs qui dépendent de la nature des acides aminés aux positions 122 et 160 de l'AgHBs, respectivement. Si l'acide aminé en position 122 est une arginine (122R), le sous-type est *y*, et si c'est une lysine (122K) le sous-type est *d*. De même, 160R définit le sous-type *r* et 160K définit le sous-type *w* (Okamoto et al., 1987). Les quatre combinaisons

possibles définissent les principaux sous-types et en rajoutant quelques déterminants mineurs, on arrive à 10 sous-types distincts (Tableau 1) (Norder et al., 1992). Le sérotypage a été utile pour des études épidémiologiques, y compris des études sur des infections nosocomiales et iatrogènes et de transmission intra-familiale.

Sérotype	**aa dans la séquence de l'AgHBs**
ayw1	122R + 160K + 127P + (134F et/ou 159A)
ayw2	122R + 160K + 127P
ayw3	122R + 160K + 127T
ayw4	122R + 160K + 127L
ayr	122R + 160R
adw2	122K + 160K +127P
adw3	122K + 160K +127T
adw4q-	122K + 160K +127L + 178Q
adrq+	122K + 160R +177V + 178P
Adrq-	122K + 160R+177A

Tableau 1 : Sérotypes et acides aminés correspondants dans l'AgHBS.
(Kay and Zoulim, 2007)

1.2.6.1.2 Génotypes

La première séquence d'un génome complet du VHB a été publiée en 1979 (Galibert et al., 1979). À la fin des années 1980, suffisamment de séquences de génomes complets s'était accumulées dans les bases de données pour permettre une classification des souches de VHB par séquence génomique plutôt que par l'antigénie de la protéine de surface. Okamoto *et al.* ont analysé 18 génomes complets et les ont divisés en quatre groupes, ou génotypes, nommés de A à D (Okamoto et al., 1988). Un génotype du VHB a été défini comme une séquence ou un groupe de séquences qui diffèrent des génotypes connus de 8% ou plus, en pourcentage d'identité sur le génome complet. Depuis lors, quatre génotypes humains de plus ont été identifiés, de E à H (Arauz-Ruiz et al., 2002; Norder et al., 1994; Stuyver et al., 2000). Il existe donc 8 génotypes du VHB, nommés de A à H. Le classement par génotypes met bien en relief l'importance de la géographie dans l'évolution du VHB, le virus évoluant différemment dans différents endroits du globe (Figure 13).

Figure 13 : Répartition géographique des génotypes du VHB.

Il y a une certaine concordance entre sérotype et génotype (Norder et al., 1992). D'ailleurs, le Tableau 2 présente les sérotypes les plus fréquemment retrouvés pour chaque génotype. De plus, il y a également un lien entre certaines mutations dans le PBC et/ou dans la région précore et le génotype. Il y a donc des différences dans les manifestations cliniques de l'infection pour les différents génotypes (Kramvis et al., 2008).

Un génome « type » du VHB a une longueur de 3215 nt, comme c'est le cas pour les génotypes B, C, F et H. Les autres génotypes présentent des génomes de longueurs différentes, dues à des insertions et délétions. Ainsi, le génotype A présente des génomes de 3221 nt (insertion de 6 nt), alors et les génomes de génotype D ont une longueur de 3182 nt (délétion de 33 nt), ceux de génotype E une longueur de 3212 nt (délétion de 3 nt) et ceux de génotype G de 3248 nt (insertion de 33 nt). Les caractéristiques de chacun des génotypes sont présentées dans le Tableau 2.

Génotype	Sérotype	Distribution géographique	Longueur génome (nt)	Pol (aa)	AgHBc (aa)	LHBs (aa)
A	adw2, ayw1	Afrique, Asie, Europe du Nord, Amérique du Nord	3221	845	185	400
B	adw2, adw3, ayw1	Asie	3215	843	183	400
C	adrq+, ayr, adw2, ayw1, adrq−	Asie du Sud-Est, Australie	3215	843	183	400
D	ayw2, adw1, ayw1, ayw3, ayw4	Europe, Etats-Unis, Australie, Asie du Sud-Est, Afrique du Sud	3182	832	183	389
E	ayw4, ayw2	Afrique subsaharienne, Rouyaume-Uni, France	3212	842	183	399
F	adw4, ayw4	Amérique (Centrale, Nord, Sud),	3215	843	183	400
G	adw2	Etats-Unis, Japon, Allemagne, France, Amérique-Centrale	3248	842	195	399
H	adw4	Etats-Unis, Japon, Amérique-Centrale	3215	843	183	400

Tableau 2 : Caractéristiques des génotypes du VHB.
Les longueurs des protéines Pol, HBc et LHBs sont présentées car elles varient en fonction du génotype (HBx, MHBS et SHBS ont des longueurs fixes ; respectivement 154, 281, 226 aa).

Certains génotypes peuvent être subdivisés en sous-génotypes, qui diffèrent par au moins 4% du génome complet (Kramvis and Kew, 2005), et montrent aussi une distribution géographique distincte. Excepté pour les génotypes E, G et H, tous les génotypes peuvent être subdivisés en sous-génotypes. Ainsi, le génotype A peut être divisé en sous-génotypes Aa ou A1 (Afrique, Asie) et Ae ou A2 (Europe, Amérique du Nord), et le génotype B en Bj ou B1 (Japon) et Ba ou B2 (reste de l'Asie). En réalité, le sous type Bj semble être le seul véritable génotype B, puisque les génomes de sous-types Ba sont essentiellement de génotype B mais avec un gène codant pour HBc dérivé du génotype C, probablement par un lointain événement de recombinaison (Sugauchi et al., 2003).

Certaines études ont montré des doubles infections avec des souches de deux génotypes différents (Tabor et al., 1977). Ces coinfections avec deux génotypes différents du VHB, chez un même patient, peuvent mener à des échanges de matériel génétique entre les deux souches. Malgré une bonne connaissance de la réplication du VHB, le mécanisme de ces événements de recombinaison reste énigmatique. Plusieurs

génomes recombinants ont été décrits, notamment des recombinants A/D, A/C, B/C, et C/D (Simmonds and Midgley, 2005). Des études ont également permis d'identifier des recombinants de génomes de VHB humains avec des variants des virus de chimpanzés et de gibbons (Simmonds and Midgley, 2005).

Les échelles de temps de recombinaison chez le VHB sont clairement compatibles avec l'existence de recombinants individuels géographiquement dispersés. Au cours de cette longue échelle de temps, chacune des formes recombinantes du VHB peut se propager et remplacer ses ancêtres et donc occulter la séquence des événements de recombinaison qui ont conduit à son origine (Simmonds and Midgley, 2005).

Concernant les recombinaisons intra-génotype, associées à de multiples événements de recombinaison avec des variants du VHB souvent divergents, l'étude de Simmonds *et al.* (Simmonds and Midgley, 2005) indique que la plupart des génotypes du VHB actuellement classés sont eux-mêmes mosaïques.

À l'avenir, les différences observées dans la réponse au traitement (Akuta and Kumada, 2005), combinées à l'évidence croissante de la coinfection fréquente avec différents génotypes, peuvent fournir une force motrice supplémentaire de recombinaison pour générer des virus résistants.

Enfin, la preuve de recombinaison entre variants du VHB de l'homme et du singe ont des implications importantes pour l'éradication et les mesures de contrôle pour le VHB. Leur présence démontre que les variants VHB humains et non humains peuvent partager les hôtes dans la nature. L'existence d'une infection endémique chez les primates non humains et son potentiel de transmission entre espèces, et chez l'humain, va entraver les efforts pour l'éradication de l'infection par le VHB dans les régions où les singes et les humains partagent l'habitat (Simmonds and Midgley, 2005).

1.2.6.2 Variabilité phénotypique

Les variants phénotypiques émergent en réponse à une pression sélective. Cette pression peut être due à la réponse immunitaire de l'hôte ou à des mesures préventives ou thérapeutiques (vaccination, traitements antiviraux). Les variants phénotypiques ont généralement une capacité de réplication moins bonne que les variants génotypiques normaux, en l'absence de pression selective. Ceci est illustré par le fait qu'ils n'apparaissent pas comme populations virales majeures chez les patients en l'absence de pression sélective et, dans le cas de mutants résistants, les mutations de résistance

aux médicaments sont habituellement rapidement perdues lorsque le médicament est enlevé. Pour le virus, il y a donc un compromis entre un fitness réduit et une résistance accrue, et cet équilibre permettra de déterminer la facilité avec laquelle un mutant donné peut émerger. Comme une grande partie du génome du VHB est composé de chevauchement des cadres de lecture, l'équilibre est délicat (Kay and Zoulim, 2007). Par exemple, les mutations de la polymérase virale qui confèrent la résistance aux médicaments peuvent provoquer des changements dans les antigènes de surface, et potentiellement affecter l'assemblage des virions, la stabilité ou l'infectiosité (Locarnini and Yuen, 2010).

1.2.6.2.1 Les mutants de l'AgHBe

Parmi ces variants phénotypiques, on distingue les mutants précore, qui possèdent des mutations abolissant l'expression de l'AgHBe. Dans la plupart des cas, il s'agit d'une mutation du codon 28 de la région préC menant à un codon de terminaison. Il faut noter que tous les génomes de génotype G ont cette mutation précore. En revanche, cette mutation est très rarement retrouvée dans les génomes de génotype A, pour lesquels ce codon 28 se situe au niveau du signal d'encapsidation. Cette mutation précore déstabiliserait le signal d'encapsidation, rendant le virus mutant moins viable (Carman et al., 1989; Tong et al., 1990). Il existe d'autres mutations affectant l'expression de l'AgHBe, telles que des mutations dans le PBC (Kay and Zoulim, 2007).

1.2.6.2.2 Les mutants PréS1 et PréS2

Des mutants des protéines PréS1 et PréS2 ont également été décrits (Melegari et al., 1994; Bock et al., 1997). Les mutations dans ces régions sont généralement des délétions plus ou moins importantes. La région PréS2 n'est pas essentielle pour le cycle viral, et des mutants présentant des délétions dans cette région peuvent être stables. En revanche, la protéine PréS1 (LHBs) est nécessaire pour la morphogenèse et la sécrétion des particules virales et pour la reconnaissance des hépatocytes. De plus, des séquences du promoteur S se trouvent dans cette région PréS1. Des mutants portant des délétions importantes dans la région PréS1 sont donc défectifs et ne peuvent survivre qu'en présence d'autres génomes, non mutés, qui peuvent les complémenter en *trans*.

1.2.6.2.3 Les mutants de l'AgHBs

Il existe également des mutants de l'AgHBs. Un porteur chronique ne développe pas de réponse immune importante contre l'AgHBs, il n'y a donc pas de pression de sélection des mutants de l'AgHBs. Ce n'est pas le cas chez les personnes vaccinées ou recevant des IgG anti-HBs à la suite d'une greffe du foie par exemple. Dans ces deux cas, des mutants échappant à la vaccination ont émergé (Carman et al., 1990). La plupart des mutations connues se trouve dans la région du déterminant a (Cooreman et al., 2001). On pense donc que c'est l'altération de la conformation du déterminant a qui est responsable de l'échappement.

1.2.6.2.4 Les mutants de la Polymérase

Les mutants de la polymérase, qui permettent au VHB de résister aux traitements antiviraux seront développés dans le paragraphe 1.3.4.3 (page 38).

1.3 L'hépatite B : la pathologie

1.3.1 Symptômes

La plupart des individus ne présentent aucun symptôme pendant la phase d'infection aiguë. Néanmoins, certaines personnes subissent une forme aiguë de la maladie, avec des symptômes qui durent plusieurs semaines, parmi lesquels un jaunissement de la peau et des yeux (ictère), une coloration foncée des urines, une extrême fatigue, des nausées, des vomissements et des douleurs abdominales.

1.3.2 Pathogénie

Les lésions hépatocytaires sont principalement dues à l'attaque des cellules infectées par les lymphocytes T cytotoxiques qui reconnaissent les épitopes de capside et de l'enveloppe virale. Parallèlement, les lymphocytes T cytotoxiques et les cellules Natural Killer produisent des cytokines (*i.e.* interféron-gamma) capables d'inhiber la réplication du VHB. Ces deux mécanismes combinés peuvent être à l'origine du contrôle de l'infection virale par la réponse immune (Dény and Zoulim, 2010).

Lorsque la réponse immune cellulaire médiée par les lymphocytes T CD4 est insuffisante, ceci aboutit à une destruction chronique des hépatocytes insuffisante pour

éliminer toutes les cellules répliquant le génome viral. Ce phénomène est à l'origine de l'hépatite chronique qui peut évoluer vers la cirrhose.

Concernant l'hépatite B aigüe, 90 à 95% des cas évoluent vers la guérison (Mutimer and Oo, 2011). Pour 1% des cas d'hépatites aigües, il y a une évolution vers une hépatite fulminante avec nécrose massive du parenchyme hépatique, qui mène à l'insuffisance hépatique. 5 à 10% des hépatites aigües évoluent vers une hépatite chronique (Dény and Zoulim, 2010). L'évolution de l'infection chronique est, en général, composée de 4 phases (Mutimer and Oo, 2011) :

- la phase I, dite de tolérance immune durant laquelle il y une forte réplication du virus et un état de tolérance du système immunitaire vis à vis des cellules infectées. Cette phase est caractérisée par la présence de l'antigène HBe (HBe +), qui pourrait favoriser cet état de tolérance, et par une quantité d'ADN viral sérique élevée. Du fait de la tolérance immunitaire, les transaminases sériques ont un taux normal et on observe une quasi absence de lésions hépatiques.
- La phase II, dite de clairance immune, durant laquelle le système immunitaire s'active. Il y a alors un conflit entre la réplication virale et la réponse immunitaire, ce qui conduit à la constitution des lésions chroniques du foie. Du fait de la destruction des cellules infectées, la quantité d'ADN virale est plus faible. Dans le sérum, on retrouve l'antigène HBe + et une quantité plus élevée des transaminases. C'est à cette phase qu'il est nécessaire de commencer le traitement antiviral pour bloquer la réplication du VHB et d'induire le contrôle de l'infection par le système immunitaire.
- La phase III, dite de portage inactif, durant laquelle les hépatocytes infectés réplique le virus à minima. Il y a une faible expression des antigènes viraux, conduisant à une réduction de l'attaque des cellules infectées. Il y a également une négativation de l'antigène HBe, une apparition d'anticorps anti-HBe, ainsi qu'une diminution de la quantité d'ADN viral sérique et une normalisation du taux de transaminases. Mais à ce stade, des cellules contenant de l'ADN viral super-enroulé peuvent persister (Omata et al., 1986; Ruiz-Opazo et al., 1982). Ceci peut conduire à une réactivation virale et à des cellules comportant le génome viral intégré dans le génome de l'hôte, ce qui peut être à l'origine de l'oncogenèse viro-induite.

- La phase IV, dite de clairance ou élimination de l'antigène HBs, durant laquelle il a une négativation de l'antigène HBs avec éventuellement l'apparition d'anticorps anti-HBs. A ce stade, des cellules comportant de l'ADN viral super-enroulé persistent dans le tissu infecté ainsi que des cellules comportant le génome viral intégré.

La cirrhose virale B est une forme sévère de l'hépatite B chronique. L'hépatite fulminante est une forme rare d'hépatite B. Elle correspond à la destruction massive des hépatocytes infectés par les lymphocytes T cytotoxiques. Les transaminases sont alors très élevées et les marqueurs viraux sont quasi-indétectables du fait de la destruction des hépatocytes.

1.3.3 Diagnostic

Le diagnostic passe d'abord par une évaluation biochimique de la fonction hépatique. Le diagnostic est ensuite confirmé par une détection d'antigènes et/ou d'anticorps spécifique du VHB, dans le sérum. Trois systèmes antigènes/anticorps ont été identifiés pour détecter l'hépatite B :

- l'antigène de surface AgHBs et l'anticorps anti-HBs
- l'antigène de capside AgHBc et les anticorps IgM anti-HBc et IgG anti-HBc
- l'antigène AgHBe et l'anticorps anti-HBe

Il est également possible de détecter l'ADN viral par PCR.

L'hépatite B aiguë est caractérisée par la présence dans le sérum d'AgHBs, d'AgHBe et d'IgM anti-HBc. Au cours de la phase de guérison, les AgHBs et HBe disparaissent et les anticorps anti-HBe et anti-HBs apparaissent (Figure 14). Une personne négative pour l'AgHBs, mais positive pour les anticorps anti-HBs a, soit guéri d'une infection antérieure, soit été vaccinée auparavant (Dény and Zoulim, 2010).

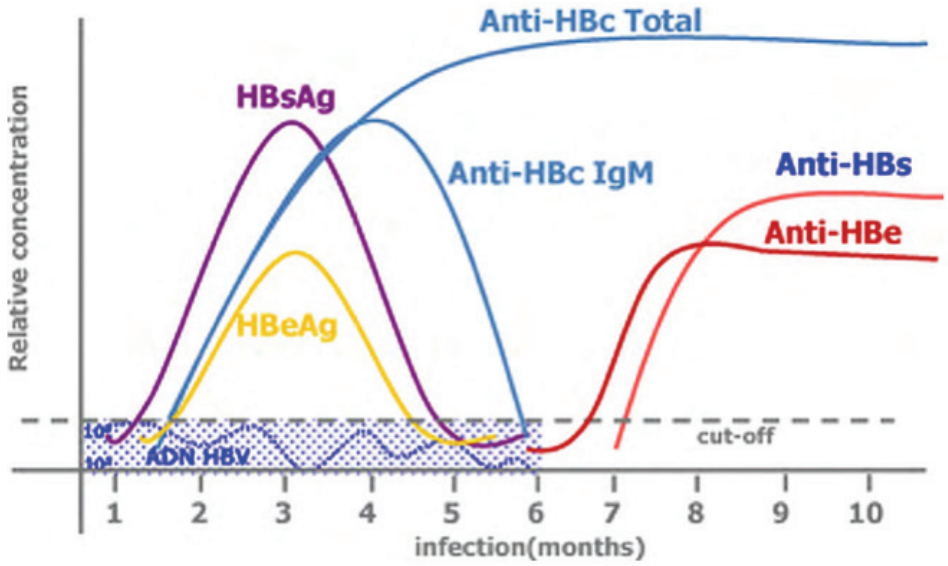


Figure 14 : Evolution de la concentration des marqueur sérologiques au cours de l'infection aigue par le VHB.
(Dény and Zoulim, 2010)

Les patients atteints d'hépatite aiguë, qui maintiennent une concentration constante d'AgHBs sérique, ou dont l'AgHBe persiste dans le sérum 8 à 10 semaines après la disparition des symptômes, sont susceptibles de devenir porteur du VHB, avec le risque de développer une hépatite chronique. Les tests de laboratoire pour mieux cerner le stade de la maladie reposent sur les enzymes hépatiques, la numération formule plaquettaire, le taux de prothrombine et l'électrophorèse des protéines. Les tests pour déterminer le niveau de réplication virale reposent sur la détection de l'AgHBe, des anticorps anti-HBe et de l'ADN viral. Le dosage des IgM anti-HBc, normalement caractéristiques de l'hépatite aiguë, peut être positif lors des phases de réactivation virale ainsi que pour les infections par un mutant précore du VHB. La détection d'antigène pré-S1 peut être utile pour mettre en évidence une réplication virale chez des patients ayant un AgHBe négatif (Dény and Zoulim, 2010).

La biopsie hépatique est recommandée chez les patients présentant des critères d'hépatite chronique : élévation des transaminases et présence d'une réplication virale. Le but de la biopsie hépatique est de déterminer le degré d'atteinte hépatique. Le diagnostic histologique d'une hépatite chronique doit inclure l'étiologie, le niveau d'activité nécro-inflammatoire, et la détermination du stade de la fibrose.

1.3.4 Les traitements contre l'hépatite B

Le petit génome du VHB et l'utilisation des enzymes cellulaires de l'hôte pour de nombreux stades du cycle de vie du virus suggèrent qu'un nombre relativement faible de cibles sont disponibles pour le développement d'antiviraux. Du fait que la polymérase du VHB exerce la fonction enzymatique primaire de la réplication virale, elle a été la principale cible du développement de médicaments anti-VHB. Les analogues de nucléo(t)ides (NA) ont été la principale classe d'agents antiviraux développés à cet effet. La plupart des traitements actuellement approuvés pour l'hépatite B chronique (CHB) appartiennent à cette classe de composés. Dans la plupart des traitements, les NA sont également accompagnés par l'administration de l'interféron alpha (IFN-α) (Ghany and Liang, 2007).

1.3.4.1 L'interféron alpha

L'interféron alpha est une cytokine qui présente plusieurs types d'activité :

- antivirale directe par inhibition des ARN viraux en activant une cascade enzymatique,
- immunostimulante en augmentant l'expression des antigènes d'histocompatibilité de classe I et en stimulant l'activité des lymphocytes auxiliaires helper ainsi que les cellules Natural Killer.

En clinique, l'effet immuno-modulateur de l'IFN-α se traduit par une poussée cytolytique qui correspond à la clairance immunologique des hépatocytes infectés.

Des interférons pegylés ont été développés afin d'allonger la demi-vie de l'IFN-α. Ils se caractérisent par un groupement polyéthylène glycol lié à l'IFN. La pegylation de l'IFN a optimisé sa pharmacocinétique. Comme l'IFN-α, l'IFN pegylé a la double action antivirale et immunostimulante. (Craxi and Cooksley, 2003; Marcellin et al., 2005)

1.3.4.2 Les inhibiteurs de polymérase : analogues de nucléos(t)ides (NA)

Ce sont des inhibiteurs de la transcriptase inverse. Ils se comportent comme de faux substrats pour cette enzyme.

Les 5 analogues de nucléos(t)ides approuvés et utilisés à l'heure actuelle sont : la lamivudine (LMV ou 3TC), l'adéfovir (ADV), l'entecavir (ETV), la telbivudine (LdT), le ténofovir (TDF) (European Association For The Study Of The Liver, 2012).

Ces 5 molécules appartiennent à 3 familles de drogues antivirales ou groupes structuraux : les L-nucléosides, les cyclopentanes (analogues de nucléosides), et les phosphonates acycliques (analogues de nucléotides) (Ghany and Liang, 2007) (Figure 15).

La lamivudine (LMV ou 3TC) est un L-nucléoside analogue de la didéoxycytidine. Elle a une activité inhibitrice sur la transcriptase inverse du VIH et également sur celle du VHB. Son mode d'action est une inhibition compétitive de l'incorporation de déoxycytidine, se traduisant par l'arrêt de la synthèse du brin d'ADN viral (Lai et al., 1998).

L'adéfovir (9-(2-phosphonylméthoxyéthyl) adénine) et sa pro-drogue adéfovir-dipivoxil appartiennent à la famille des phosphonates acycliques. La forme active diphosphorylée est un inhibiteur de virus à ADN et de rétrovirus (Marcellin et al., 2003). Chez l'homme, l'administration par voie orale est limitée par une faible biodisponibilité et une mauvaise absorption cellulaire. L'utilisation d'un dérivé estérifié, l'adéfovir-dipivoxyl, augmente la biodisponibilité orale et l'absorption cellulaire. Une fois dans la cellule, la pro-drogue est rapidement transformée par des estérases cellulaires. L'adéfovir a une grande similarité structurale avec le substrat naturel (dATP), et il est un inhibiteur compétitif par rapport au dATP.

L'entécavir appartient à la classe des composés de cyclopentane et a une activité puissante et sélective contre le VHB. L'entécavir est un analogue carbocyclique de 2-désoxy-guanosine, dans lequel l'oxygène dans le cycle furanose est remplacé par un groupe vinyle. L'entécavir affecte les fonctions multiples de la polymérase, y compris l'amorçage, la transcription inverse et élongation de l'ADN (Innaimo et al., 1997).

La telbivudine (β-L-2'-désoxythymidine) est un L-nucléoside et semble être plus puissant que la lamivudine dans la suppression des niveaux d'ADN VHB *in vivo* (Lai et al., 2005).

Le ténofovir disoproxil fumarate est une molécule proche de l'ADV, puisqu'il appartient à la classe des phosphonates acycliques, étant un analogue de didéoxi-adénosine. Il a un mécanisme d'action et une puissance similaire à l'adéfovir. La dose prescrite est environ 30 fois plus grande que l'adéfovir, ce qui peut expliquer sa plus grande puissance *in vivo* (Clercq, 2007).

Ces drogues bloquent l'étape de réplication du VHB mais n'éliminent pas le pool d'ADNccc présent dans les cellules. Le problème majeur de ces analogues de

nucléos(t)ides à l'heure actuelle, c'est que leur utilisation a abouti à l'émergence de populations virales de mutants résistants à ces drogues.

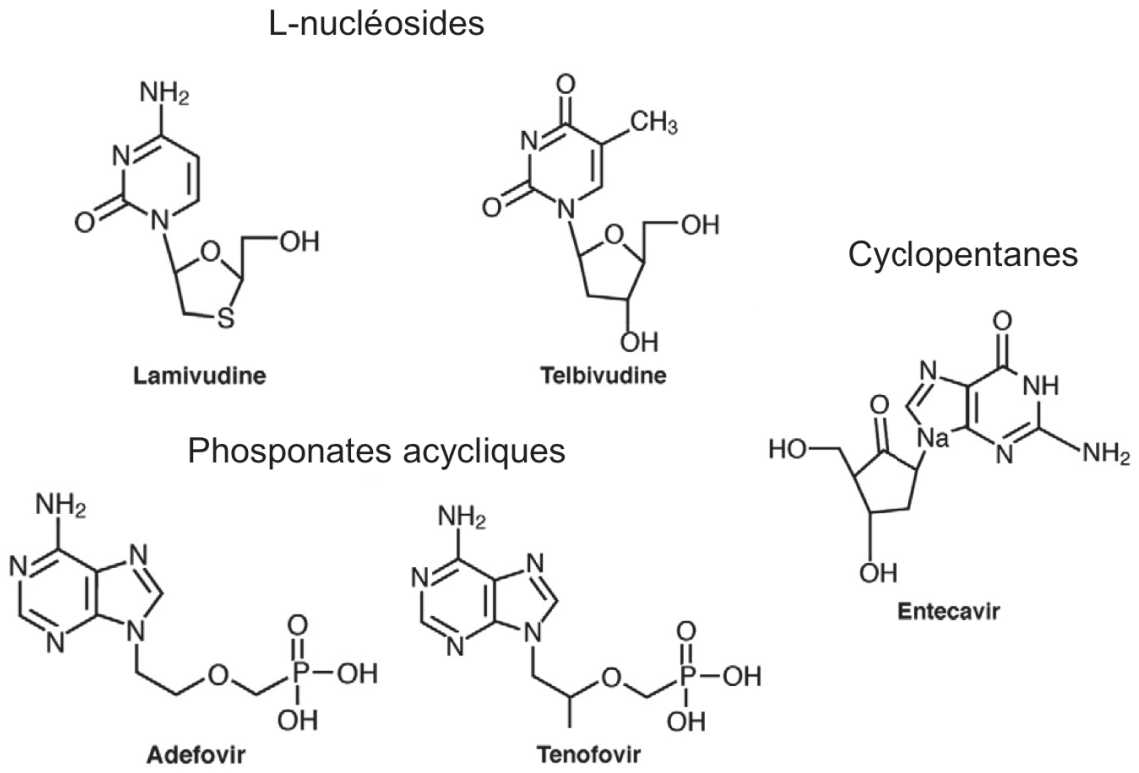


Figure 15 : Formules chimiques des analogues de nucléos(t)ides.
(Ghany and Liang, 2007)

1.3.4.3 La résistance aux NA

La transcriptase inverse du VHB n'a pas de fonction de relecture (proof-reading). Par conséquent, les mutations peuvent survenir rapidement.

Avant le traitement, il peut exister un ensemble diversifié de virus, les quasi-espèces, comprenant des mutants avec une ou plusieurs mutations potentiellement associées à la résistance aux médicaments. Différents mécanismes sont impliqués dans la sélection de mutants résistants au cours de la thérapie antivirale (Zoulim, 2004; Pawlotsky, 2005). Un variant ayant une mutation lui conférant un avantage sélectif va générer un virus de descendance, qui se réplique mieux en présence du médicament et peut se propager plus rapidement dans le foie, ce qui permet au mutant correspondant de s'accumuler et de devenir l'espèce dominante dans le foie en présence de la drogue antivirale (Zoulim and Locarnini, 2009).

La probabilité qu'une mutation soit sélectionnée au cours du traitement dépend de la capacité du médicament à supprimer la réplication virale. Une monothérapie exerçant une activité antivirale modeste et dirigée vers un seul site cible se traduirait par une haute probabilité de sélection de résistance. Le schéma de traitement idéal doit avoir des activités antivirales ciblées sur différents sites afin de réduire le risque de sélectionner des espèces résistantes aux NA. La résistance apparaît lorsque la réplication a lieu en présence de la pression de sélection de la drogue. Par conséquent, si nous pouvions obtenir une suppression complète de la réplication, la résistance ne serait pas un problème. Les autres facteurs contribuant à l'émergence de la résistance aux drogues sont les barrières génétiques au développement de mutations, le mécanisme de résistance et les facteurs de l'hôte impliqués dans le contrôle de réplication virale (Ghany and Liang, 2007).

Certaines mutations de résistance sont spécifiques à une classe d'analogues de nucléos(t)ides.

1.3.4.3.1 Résistance associée aux L-nucléosides

La résistance à la lamivudine se développe à un taux de 14% -24% par an et est d'environ 70% pour 4 à 5 ans (Lok et al., 2003).

La résistance à LMV et LdT a été cartographiée au locus YMDD dans le catalyseur de la RT (domaine C, voir Figure 6 page 13) (Stuyver et al., 2001). Elle est essentiellement médiée par les mutations rtM204I/V (domaine C) ± rtL180M (domaine B) et rtA181T/V (Yeh et al., 2000). Des mutations compensatoires qui augmentent le taux de réplication virale peuvent être trouvées dans d'autres domaines de la RT, comme rtL80V/I, 58 rtI169T, rtV173L, rtT184S/G, rtS202I et rtQ215S (Tenney et al., 2004; Delaney et al., 2003; Shaw et al., 2006).

1.3.4.3.2 Résistance associée aux cyclopentanes

Des mutations associées à l'apparition d'une résistance à l'entecavir ont été cartographiées dans le domaine B (rtI169T, L180M, et/ou rtS184G), dans le domaine C (rtS202I et M204V), et dans le domaine E (rtM250V) du domaine RT. En l'absence des mutations de résistance à la lamivudine L180M et M204V/I, la mutation rtM250V augmente l'EC_{50} (concentration efficace) de l'ETV de 10 fois, alors que les mutations rtI169T, rtT184G ou rtS202I n'ont qu'un effet modeste sur les valeurs d'IC_{50} (Shaw et al., 2006). Une ou plusieurs autres mutations en plus de rtL180M et rtM204V :

rtT184G/S et/ou rtS202I/G et/ou rtM250V, sont nécessaires pour développer une résistance à l'ETV.

1.3.4.3.3 Résistance associée aux phosphonates acycliques

La résistance à l'adéfovir a été initialement associée à des mutations dans les domaines B (rtA181T/V) et D (rtN236T) de la RT (Villet et al., 2008; Angus et al., 2003; Villeneuve et al., 2003). Ces substitutions résultent en une modeste (3 à 8 fois) augmentation de la concentration nécessaire de drogue pour inhiber 50% de la réplication virale *in vitro*. Ces mutations confèrent une partielle résistance croisée au ténofovir, certainement en raison de sa structure chimique similaire à l'ADV (Shaw et al., 2006). La mutation rtN236T n'affecte pas significativement la sensibilité à LMV, LdT ou ETV (Angus et al., 2003; Villeneuve et al., 2003) mais diminue l'efficacité du TDF *in vitro (Brunelle et al., 2005)*. Les mutations rtA181T/V confèrent une sensibilité diminuée à ADV et TDF et une résistance croisée partielle à LMV et LdT (Villet et al., 2008).

Une résistance au TDF aurait été détectée chez plusieurs patients ayant une coinfection VIH-VHB, la substitution rtA194T (plus rtL180M et M204V) a été associée à une résistance au TDF (Sheldon et al., 2005), mais un rapport ultérieur n'a pas confirmé cette résistance (Delaney et al., 2006). A l'heure actuelle, les études n'indiquent pas de mutation de résistance au ténofovir (Marcellin et al., 2012; European Association For The Study Of The Liver, 2012).

1.3.4.3.4 Résistance croisée

Huit codons du domaine RT de Pol sont donc associés à la pharmaco-résistance aux analogues de nucléos(t)ides : 169, 180, 181, 184, 202, 204, 236 et 250. Ces 8 codons ont été décrits comme étant impliqués dans la résistance aux antiviraux par 4 voies d'évolution virale (Locarnini, 2008) :

- la voie M204V/I pour les L-nucléosides,
- la voie rtN236T pour les phosphonates acycliques,
- la voie rtA181T/V, qui est partagée par les L-nucléosides et les phosphonates acycliques,
- la voie cyclopentane/entecavir (L180M + M204V/I ± I169T ± T184S/G/C ± S202C/G/I ± M250I/V).

Les 3 premières voies sont associées à une seule mutation, tandis que la quatrième nécessite au moins 3 mutations pour la résistance. Cette étude des voies d'évolution facilite la compréhension de l'évolution du VHB au cours de la thérapie aux NA et peut être utilisée pour prédire les résultats des patients et améliorer notre compréhension des profils de résistance croisée. Les profils de résistance aux NA sont présentés dans le Tableau 3.

Mutations dans le domaine RT de la polymérase du VHB	**Niveau de susceptibilité**				
	LMV	**LdT**	**ETV**	**ADV**	**TFV**
Virus Sauvage	S	S	S	S	S
M204V	R	S	I	S	S
M204I	R	R	I	S	S
L180M + M204V	R	R	I	S	S
A181T/V	R	R	S	R	I
N236T	S	S	S	R	I
A181T/V + N236T	R	R	S	R	I/R
L180M + M204V/I ± I169T ± V173L ± M250V	R	R	R	S	S
L180M + M204V/I ± T184G ± S202I/G	R	R	R	S	S

Tableau 3 : Profils de substitutions et résistance aux analogues de nucléos(t)ides.
Numérotation RT, niveaux de susceptibilité : S (sensible), I (intermediaire), R (résistant). (European Association For The Study Of The Liver, 2012)

La monothérapie séquentielle peut favoriser la sélection de souches multi-résistantes, en particulier lorsque les patients sont successivement traités avec des médicaments ayant des caractéristiques similaires, comme avec LMV suivie par ETV ou LMV suivie par ADV (Yim et al., 2006; Brunelle et al., 2005). Des analyses clonales ont montré que les souches multi-résistantes apparaissent généralement par l'addition successive de mutations de résistance virale à un même génome. Les mutants qui découlent de ce processus de sélection ont une résistance totale aux multiples drogues (Zoulim and Locarnini, 2009).

Une étude des variants chez un patient avec une souche multi-résistante du VHB après transplantation hépatique a révélé des mutations, dans la région où les gènes de Pol et des protéines de surface se chevauchent, qui confèrent une résistance à la fois à LMV et ADV ainsi qu'une diminution de la reconnaissance du virus par les anticorps anti-HBs (Villet et al., 2006).

1.3.4.3.5 Impact sur les gènes de surface

L'introduction des analogues de nucléos(t)ides dans le traitement de l'hépatite chronique B a vu l'émergence de la résistance aux médicaments antiviraux comme le principal facteur limitant l'efficacité des NA. En outre, en raison du chevauchement des cadres de lecture de la polymérase virale et de l'enveloppe dans le génome du VHB, les mutations associées à la résistance au NA, sélectionnées dans les domaines catalytiques de la polymérase, donnent souvent lieu à des changements importants dans le domaine de liaison aux anticorps neutralisant de l'AgHBS. Ceci peut causer l'émergence de mutants d'échappement potentiel à la vaccination, associés aux drogues antivirales (ADAPVEMs, antiviral drug-associated potential vaccine escape mutants) (Locarnini and Yuen, 2010).

Le changement de nucléotide qui modifie le codon rt204 (rtM204I/V) dans le gène de la polymérase confère une résistance à LMV, LdT, et ETV et entraîne également un changement non synonyme dans le gène codant pour l'antigène de surface du VHB (AgHBs), directement dans la région chevauchante. La mutation M204V se traduit généralement par la substitution sI195M dans AgHBs, alors que le changement rtM204I peut causer sW196S, sW196L, ou un codon terminateur.

Les mutations qui entraînent un codon stop dans le gène de l'enveloppe, comme ceux de LMV, LdT (rtM204I et rtA181T) et ADV (rtA181T), sont généralement trouvées en présence d'un faible pourcentage de VHB de type sauvage, permettant l'encapsidation virale et la libération du variant défectueux (Warner and Locarnini, 2008). Le Tableau 4 présente les mutations dans le domaine RT de la Pol et les substitutions qui sont associées dans l'AgHBs.

Mutation de résistance dans le domaine RT	**Impact correspondant dans l'AgHBs**
rtI169T	sF161H/L
rtA181T	sW172*/L
rtA181V	sL173F
rtT184G	sL176V
rtS202I	sV194F/S
rtS202G	- /sS193L
rtM204V	sI195M
rtM204I	sW196*/S/L

Tableau 4 : Mutations de résistance dans le domaine RT de la Pol et les substitutions correspondantes dans l'AgHBs.
(Zoulim and Locarnini, 2009)

Des études ont montré que des mutations dans la région chevauchant la Pol et les gènes de surface, peuvent nuire à la capacité de réplication, à l'efficacité de sécrétion des virions, et à l'infectiosité. Certains de ces mutants ont échappé à la reconnaissance des anticorps et peuvent donc causer une infection chez des personnes vaccinées et échapper à la détection par les kits de diagnostic commerciaux. L'effet de ces substitutions sur l'infectiosité du VHB a été montré comme étant un facteur déterminant grâce auquel les mutants résistants se propagent plus rapidement dans le foie et dominent les autres variants du VHB (Villet et al., 2009).

1.4 Bases de données biologiques généralistes

L'ère de la génomique a vu le jour avec le premier séquençage d'un génome complet de bactériophage en 1976 (Fiers et al., 1976). Le séquençage de génomes complets a pris une grande ampleur à la fin des années 90 avec le séquençage des premiers génomes à ADN de bactéries, d'archées et d'eucaryotes (Fleischmann et al., 1995; Bult et al., 1996; Goffeau et al., 1996). L'amélioration des techniques de séquençage et la diminution de leur coût ont conduit à une incroyable augmentation du nombre de données de séquences produites. Ceci a contribué au large développement de projets de séquençage de génomes complets, atteignant en septembre 2012 le nombre de 3699. L'évolution quasi-exponentielle du nombre de génomes séquencés est présentée sur la Figure 16.

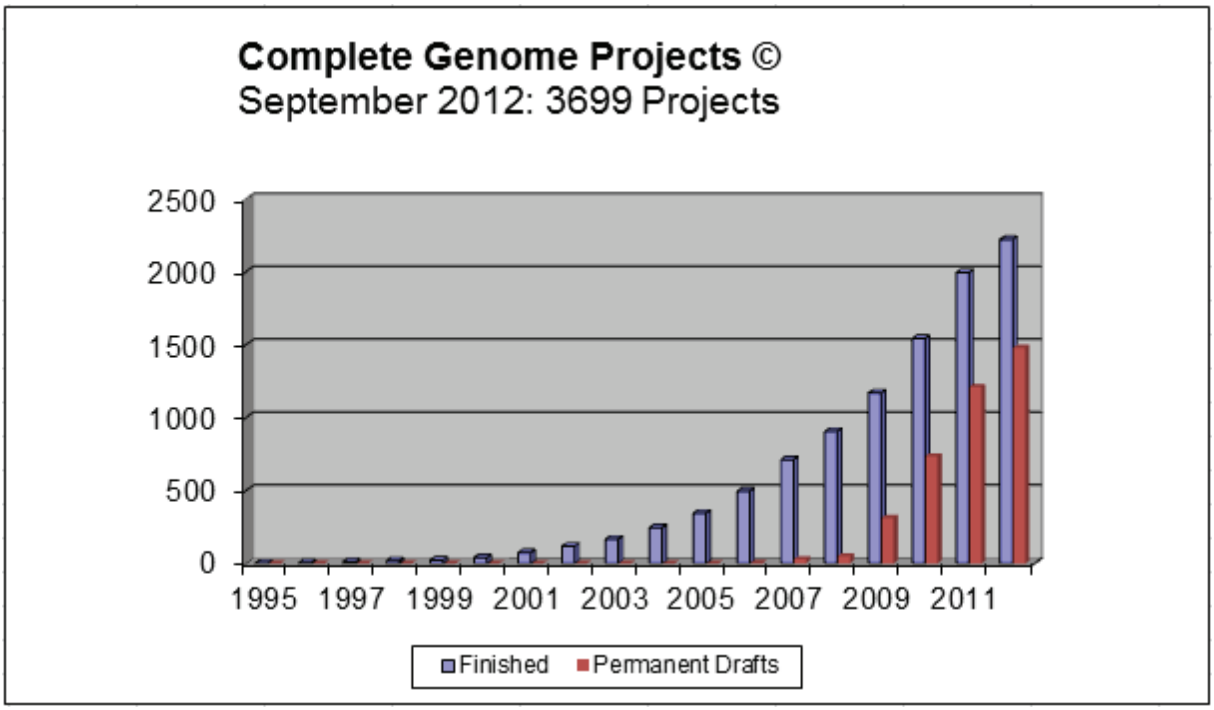


Figure 16 : Evolution du nombre de projets de séquençage de génomes complets.
Source GOLD, Genomes Online Database (Pagani et al., 2012).

A l'heure actuelle, des techniques de séquençage à haut débit sont utilisées, comme le pyroséquençage, et produisent de très grandes quantités de données, toujours en augmentation.

Des systèmes de stockages ont dû être mis en place pour faire face à ce nombre grandissant de séquences. Ainsi, des bases de données publiques ont vu le jour, et des outils informatiques adaptés au traitement et à l'analyse de la grande quantité de séquences disponibles ont été développés, en association avec les bases de données. Il existe différents types de bases : les bases généralistes, couvrant l'ensemble des thématiques biologiques, et les bases spécialisées, fournissant les informations disponibles sur une thématique donnée.

Avec l'émergence de ces bases de données, de nouveaux problèmes sont apparus. Le premier problème correspond à la redondance d'information dans les bases, qui peut biaiser certaines analyses de séquences. Le deuxième problème concerne l'annotation des séquences, qui n'est pas toujours cohérente et standardisée, rendant parfois difficile ou peu efficace la recherche d'information par mots clés dans ces bases.

1.4.1 Bases de séquences nucléotidiques

Les trois banques de séquences nucléotidiques principales dans lesquelles sont déposées toutes les séquences connues sont : DDBJ (DNA Data Bank of Japan) (Kaminuma et al., 2011), ENA (European Nucleotide Archive) (Leinonen et al., 2011; Amid et al., 2012) qui comprend l'EMBL-Bank (European Molecular Biology Laboratory) et GenBank (Etats-Unis) (Benson et al., 2012). Ces trois bases sont réunies au sein de l'International Nucleotide Sequence Database Collaboration (INSDC) (Karsch-Mizrachi et al., 2012), regroupement permettant une harmonisation des formats de données, des échanges et mises à jour des données de manière quotidienne. De ce fait, l'extraction et l'analyse de toutes les séquences disponibles au sein des différentes bases sont possibles à partir de leurs interfaces web respectives.

Dans ces bases généralistes, le niveau d'annotation est faible. L'annotation est en fait principalement réalisée par l'utilisateur qui dépose de nouvelles séquences.

Au mois de novembre 2012, la release 113 de l'EMBL-Bank (septembre 2012) contient 263,6 millions de séquences (Figure 17).

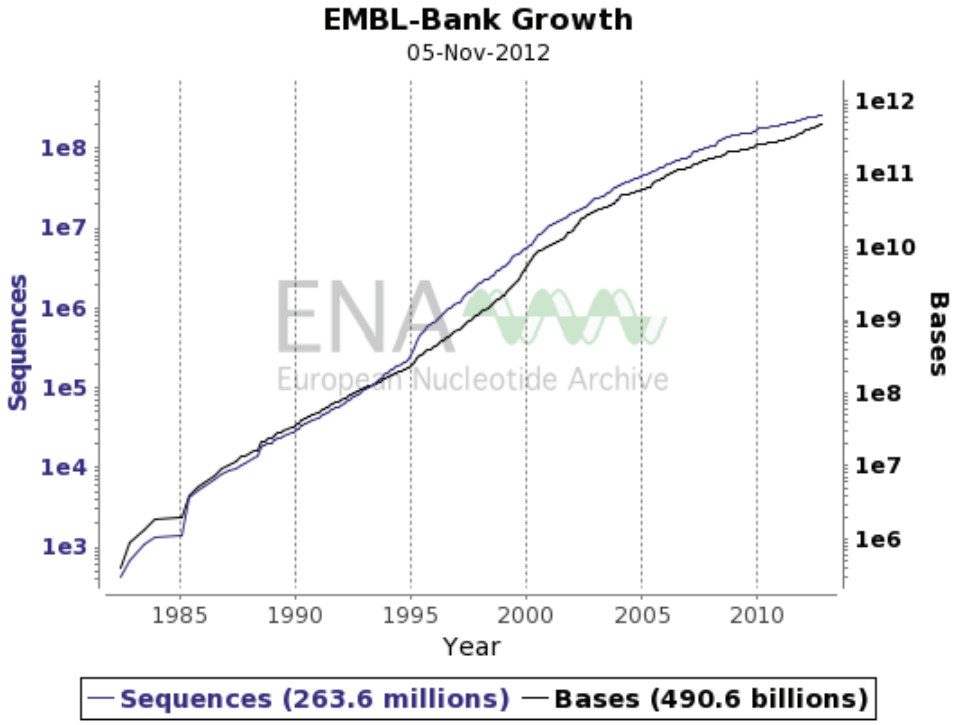


Figure 17 : Croissance de l'EMBL-Bank.

Cette base est divisée en de nombreuses sections de deux types (Tableau 5) : les divisions, qui reposent sur la taxonomie (*e.g.* Homme, plantes, procaryotes, virus, *etc.*) et les classes, qui reposent sur la nature des données qu'elles contiennent (*e.g.* EST : Expressed Sequence Tag, PAT : brevets, HTG : High Throughput Genome sequencing, *etc.*).

Classe	**Nombre d'entrées**
CON :Constructed	7 236 371
EST :Expressed Sequence Tag	73 715 376
GSS :Genome Sequence Scan	34 528 104
HTC :High Throughput CDNA sequencing	491 770
HTG :High Throughput Genome sequencing	152 599
PAT :Patents	24 364 832
STD :Standard	13 920 617
STS :Sequence Tagged Site	1 322 570
TSA :Transcriptome Shotgun Assembly	8 085 693
WGS :Whole Genome Shotgun	88 288 431
Total	252 106 363

Division	**Nombre d'entrées**
ENV :Environmental Samples	30 908 230
FUN :Fungi	65 22 586
HUM :Human	32 094 500
INV :Invertebrates	31 907 138
MAM :Other Mammals	40 012 731
MUS :Mus musculus	11 745 671
PHG :Bacteriophage	8 511
PLN :Plants	52 428 994
PRO :Prokaryotes	2 808 489
ROD :Rodents	6 554 012
SYN :Synthetic	4 045 013
TGN :Transgenic	285 307
UNC :Unclassified	8 617 225
VRL :Viruses	1 358 528
VRT :Other Vertebrates	22 809 428
Total	252 106 363

Tableau 5 : Classes et divisions de la base EMBL et le nombres d'entrées correspondantes.

Chaque entrée de l'EMBL-Bank correspond à une séquence et est subdivisée en champs auxquels sont associées des valeurs. Chaque champ est identifié par deux lettres majuscules en début de ligne. Chaque entrée possède un identifiant (ID) et un numéro d'accession (AC) uniques permettant d'accéder à la séquence. Les données minimales constituant une entrée sont la séquence (champs SQ), les références bibliographiques reliées à cette séquence (champs RN, RT, RF, RA, RX) et la source de la séquence nucléotidique (champs OS, OC et OG). Les informations supplémentaires comme les protéines codées par la séquence, les séquences régulatrices contenues dans la séquence, les sites spécifiques, les mutations, ou toutes autres données à propos de la séquence constituent les annotations et sont présentes au niveau des champs mots clés (KW), commentaires (CC), mais principalement dans les champs caractéristiques (FT, *feature*). Le champ FT suit un format particulier. Chaque *feature* se compose d'un nom, définissant le type de caractéristique (*e.g. CDS*, *mat_peptide*, *promoter*, etc.), d'une « *location* », correspondant aux positions de la séquence définissant la région concernée par le *feature* ; et une liste de « *qualifiers* » correspondant à des qualificatifs ou annotations pour le *feature* concerné. Un exemple d'entrée EMBL est présenté dans l'Annexe 1 et la liste des *qualifiers* du *feature CDS* est présentée Figure 18.

Feature	CDS
Definition	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon); feature includes amino acid conceptual translation.
Optional Qualifiers	/allele="text" /artificial_location="[artificial_location_value]" /citation=[number] /codon_start=<1 or 2 or 3> /db_xref="<database>:<identifier>" /EC_number="text" /exception="[exception_value]" /experiment="[CATEGORY:]text" /function="text" /gene="text" /gene_synonym="text" /inference="[CATEGORY:]TYPE[(same species)][:EVIDENCE_BASIS]" /locus_tag="text" (single token) /map="text" /note="text" /number=unquoted text (single token) /old_locus_tag="text" (single token) /operon="text" /product="text" /protein_id="<identifier>" /pseudogene="TYPE" /ribosomal_slippage /standard_name="text" /translation="text" /transl_except=(pos:<base_range>,aa:<amino_acid>) /transl_table =<integer> /trans_splicing

Figure 18 : *Qualifiers* optionnels proposés par l'EMBL-Bank pour le *feature CDS*.

1.4.2 Bases de séquences protéiques

Tout comme pour les séquences nucléotidiques, il existe des bases de données généralistes contenant les séquences protéiques.

La principale base de données de séquences protéiques est UniProt Knowledge Base (UniProtKB) (UniProt Consortium, 2012).

La base UniProtKB est le point central de la collecte des informations fonctionnelles sur les protéines. De la même manière que pour les séquences nucléotidiques dans l'EMBL-Bank, les données obligatoires sont la séquence d'acides aminés, le nom de la protéine et sa description, les données taxonomiques et bibliographiques. Les entrées peuvent évidemment contenir des annotations supplémentaires. Cela inclut des ontologies biologiques largement utilisées, des classifications et des références croisées et des indications claires quant à la qualité de l'annotation sous la forme de preuves par données expérimentales et informatiques.

La base UniProtKB se compose de deux sections :

- une section contenant des entrées manuellement annotées avec des informations extraites de la littérature et des analyses computationnelles évaluées par des curateurs : « UniProtKB / Swiss-Prot »
- une section avec des entrées de séquences protéiques correspondant aux traductions des CDS (coding sequence ⇔séquence codante) contenues dans la base nucléotidique EMBL-Bank, qui attendent une annotation manuelle : « UniProtKB / TrEMBL ».

La base Swiss-Prot a été créée en 1986. Depuis 1987, elle est maintenue en collaboration avec l'EBI. Aujourd'hui, Swiss-Prot est le résultat d'un partenariat égalitaire entre le SIB (Swiss Institute of Bioinformatics) et l'EBI (European Bioinformatics Institute), antenne de l'EMBL (Grande-Bretagne). C'est une base de séquences protéiques annotées, de faible redondance et qui présente un grand niveau d'intégration avec plus d'une trentaine d'autres bases de données. Ces caractéristiques font la qualité de cette base. Chaque entrée Swiss-Prot correspond à une séquence protéique, et est composée de différents types de lignes. Pour des questions de standardisation, le format Swiss-Prot suit autant que possible le format EMBL (Bairoch and Apweiler, 2000).

La section de TrEMBL UniProtKB a été introduite en 1996 en réponse à l'augmentation de flux de données résultant de projets de génomes. Il était déjà reconnu à l'époque que le processus d'annotation manuelle, demandant beaucoup de temps et de main d'œuvre, et qui est la marque de Swiss-Prot, ne pourrait pas être élargi pour englober toutes les séquences protéiques disponibles. Les séquences protéiques accessibles au public obtenues à partir de la traduction de séquences codantes annotées dans la base de données de séquences nucléotidiques EMBL-Bank, sont automatiquement traitées et entrées dans UniProtKB / TrEMBL où elles sont annotées de manière automatique afin de les rendre rapidement disponibles au public.

1.4.3 Bases de structures de macromolécules biologiques

Les protéines sont définies par leur séquence, suite d'acides aminés, leur fonction, et leur structure tridimensionnelle. En effet, la séquence est la structure primaire de la

protéine. Cette structure primaire peut se replier pour former des éléments de structure secondaire, tels que les brins béta et les hélices alpha. Et enfin, le repliement et l'organisation dans l'espace de ces structures donnent à la protéine sa structure tertiaire ou tridimensionnelle, qui a un rôle primordial pour sa fonction. En biologie, la connaissance de la structure d'une protéine peut donc renseigner sur son rôle et permet de comprendre certains mécanismes moléculaires. Par exemple, pour les interactions hôtes-pathogènes, la connaissance de la structure des protéines impliquées peut permettre de développer des molécules inhibitrices spécifiques.

La structure tertiaire est donc l'arrangement dans l'espace, de tous les atomes composant une protéine. Ces structures peuvent être déterminées par deux techniques principales qui sont la cristallographie par diffraction des rayons X et la Résonance Magnétique Nucléaire (RMN). Les structures protéiques résolues, définies par leurs coordonnées atomiques, sont regroupées au sein d'une base de données : la Protein Data Bank (PDB) (Bernstein et al., 1977). Dans sa version de novembre 2012, la PDB contient 86008 structures, parmi lesquelles 79120 correspondent à des protéines. La composition de la PDB par type de macromolécules biologiques et par type de méthodes est présentée dans le Tableau 6. La PDB contient, à l'heure actuelle, 49,7% de structures de protéines eucaryotes, 36% de protéines bactériennes et 6% de protéines virales et 3,8% de protéines d'archées.

Méthode expérimentale	**Protéines**	**Acides Nucléiques (AN)**	**Complexes Protéine/AN**	**Autres**	**Total**
Rayons-X	70690	1400	3562	3	75655
RMN	8463	1010	191	7	9671
Microscopie Electronique	322	23	120	0	465
Hybride	45	3	2	1	51
Autre	144	4	5	13	166
Total	79664	2440	3880	24	86008

Tableau 6 : Composition de la Protein Data Bank.
Types de structures et méthodes de résolution de la structure.

Le projet de création de la PDB a démarré en 1971, et la base a été opérationnelle en 1973 (Protein Data Bank, 1971; Protein Data Bank, 1973). A cette époque, elle ne contenait qu'une dizaine de structures. Entre 1999 et 2000, elle a atteint le nombre de

10000 structures disponibles. Depuis, le nombre de structures n'a cessé d'augmenter puisqu'il a été multiplié par 8 entre 2000 et 2012. En revanche, le nombre de nouvelles structures par an reste assez stable depuis 2006 (entre 6500 et 8000). La PDB présente une certaine redondance au niveau des structures et des séquences. Les séquences redondantes ont donc été regroupées en fonction du pourcentage d'identité. Il y a ainsi 19762 « clusters » de structures, définis par au moins 30% d'identité entre les séquences d'un même cluster.

Chaque fichier PDB est composé de deux sections principales : l'en-tête et les coordonnées. La section d'en-tête comporte des détails concernant : le nom de la molécule dans la structure, l'auteur, les citations, la ou les séquence(s), les composants chimiques, les structures secondaires et de l'information au sujet de la collecte de données et résolution de la structure. La section des coordonnées comprend les coordonnées atomiques, les résidus et les atomes et leurs numéros, les identificateurs de chaînes pour les polymères, les identificateurs de positions alternatives, les facteurs d'occupation et de température. Grâce aux progrès de la biologie structurale, des méthodes telles que la RMN et la microscopie électronique sont également utilisées pour déterminer la structure. Le format PDB a été modifié pour tenir compte des détails spécifiques à la méthode utilisée pour déterminer la structure, dans la section d'en-tête (Dutta et al., 2009).

Des bases de données exploitant la PDB ont été développées, et elles ont pour principal objectif de classer les structures en familles. Presque toutes les protéines présentent des similitudes structurales avec d'autres protéines et, dans certains cas, une origine évolutive commune. La connaissance de ces relations est essentielle à notre compréhension de l'évolution des protéines. Elle jouera également un rôle important dans l'analyse des données de séquences produites par les projets de séquençage de génomes à travers le monde.

La base SCOP (Structural Classification Of Proteins) établit une classification hiérarchique des protéines, sur la base des domaines, en fonction de leur structure (Murzin et al., 1995). Il y a plusieurs niveaux de hiérarchie (Andreeva et al., 2008) :

- Espèce : les représentants d'une séquence de protéine et ses variants naturels ou artificiels

- Protéines : regroupement de séquences similaires essentiellement de même fonction, qu'elles proviennent d'espèces différentes ou qu'elles représentent des isoformes au sein d'un même organisme.
- Famille : les protéines regroupées dans les familles ont une claire relation évolutive. En général, cela signifie que les identités des résidus par paire entre les protéines sont de 30% et plus. Toutefois, dans certains cas, des fonctions et des structures similaires apportent la preuve définitive d'une origine commune dans l'absence d'identité de séquence élevée, par exemple, de nombreuses globines forment une famille bien que certains membres ont des identités de séquence de 15% seulement.
- Superfamille : les protéines qui ont une identité de séquence faible, mais dont la structure et les caractéristiques fonctionnelles donnent à penser qu'une origine évolutive commune est probable, sont placées ensemble dans des superfamilles.
- Repliement : les protéines sont définies comme ayant un repliement commun si elles ont des éléments de structures secondaires similaires, dans la même disposition et avec les mêmes connexions topologiques. Les protéines placées dans la même catégorie peuvent ne pas avoir une origine évolutive commune : les similitudes structurales pourraient venir de la physique et de la chimie des protéines favorisant certains arrangements d'empaquetage et topologies de chaînes. Le nombre de repliements différents dans la PDB est assez constant depuis 2008, et il s'élève à 1393.
- Classe : les protéines sont regroupées principalement en fonction de leur contenu et organisation en structures secondaires. Il existe 5 classes : la classe tout alpha (composée de protéines essentiellement formées d'hélices α), la classe tout béta (protéines essentiellement composées de brins β), la classe alpha et béta ou α / β (alternance d'hélices α et de brins β entrecoupés), la classe alpha plus béta (α + β, hélices α et brins β bien séparés dans la structure), et enfin la classe multi-domaines (pour le protéines ayant des domaines de repliement différents et dont on ne connaît pas d'homologues).

Dans SCOP, l'unité de classification est généralement le domaine protéique. Les petites protéines, et la plupart de celles de taille moyenne, ont un seul domaine et sont donc traitées comme un tout. Les domaines des protéines de taille importante sont généralement classés séparément.

La base de données CATH est aussi une classification hiérarchique des domaines en familles basées sur la séquence et sur la structure et en groupes de repliement (Pearl et al., 2005). Les 4 niveaux définis au départ étaient : 'CATH' (Class, Architecture, Topology, Homologous superfamilies). Par la suite, des niveaux inférieurs ont été introduits (Greene et al., 2007). Ils correspondent à une hiérarchisation au niveau des séquences, sur la base de pourcentages d'identité.

Dans le niveau S, les séquences sont regroupées en fonction de la similarité de séquence significative (35% d'identité et au-dessus) (Cuff et al., 2009). Aux niveaux supérieurs (CATH), les domaines sont regroupés s'ils partagent une similarité significative de séquence, de structure et/ou de fonction (superfamilles d'homologues, niveau H) ou tout simplement similarité structurale (groupe de repliement ou topologie, niveau T). Les groupes de repliement partageant des architectures similaires (similitudes dans l'arrangement des structures secondaires, indépendamment de leur connectivité) sont ensuite fusionnés dans les architectures communes (niveau A). Au sommet de la hiérarchie, les domaines sont groupés en fonction de leur classe, c'est à dire le pourcentage d'hélices α et/ou de brins β (niveau C).

La base de données Dali est basée sur la comparaison tout-contre-tout des structures 3D des protéines de la PDB. Les voisinages et alignements structuraux sont automatiquement maintenus et mis à jour régulièrement en utilisant l'outil de recherche de structures similaires Dali (Holm and Rosenström, 2010).

1.5 Outils bioinformatiques

Lorsqu'on a une nouvelle séquence dont on ne connait pas la fonction, l'une des premières choses que l'on va faire est de la comparer aux séquences contenues dans les bases de données décrites précédemment. C'est une méthode que l'on appelle recherche de similarité ou recherche d'homologues, qui passe par l'alignement par paire de séquences.

Si on dispose de la séquence protéique, on peut également faire de la recherche de motifs caractéristiques d'une fonction ou d'une famille de protéines. Cette méthode ne sera pas abordée dans cette thèse, car elle n'a pas été utilisée.

Une fois que l'on dispose de plusieurs membres d'une famille de protéines homologues, on peut faire un alignement multiple de leurs séquences pour déterminer les régions conservées et celles qui sont plus variables d'une séquence à l'autre. A partir des alignements multiples, on peut déterminer l'histoire évolutive des séquences, en construisant une phylogénie.

La connaissance de la structure tridimensionnelle est également une donnée importante, qui peut apporter des informations sur la fonction. Si elle n'est pas résolue, des méthodes permettent de prédire, à partir de la séquence protéique, sa composition en structures secondaires, ce qui constitue une première information sur la structure tridimensionnelle (Geourjon et al., 2001). Ces méthodes peuvent également fournir une aide dans la recherche de structure « empreinte » pour la modélisation moléculaire par homologie de la structure tridimensionnelle de la protéine d'intérêt.

L'utilisation combinée de plusieurs méthodes est toujours préférable, et les résultats de prédictions doivent, dans la mesure du possible, être confirmés expérimentalement.

1.5.1 Recherche d'homologie

Deux séquences sont considérées comme homologues si elles ont évolué à partir d'un ancêtre commun. L'homologie est une notion qualitative, elle n'est pas quantifiable.

Par contre, la similarité de séquence est une quantité mesurable qui indique à quel degré deux séquences se ressemblent. La similarité de séquence peut s'exprimer en pourcentage d'identité calculé à partir d'un alignement. L'homologie peut être déduite de la similarité quand celle-ci est statistiquement significative.

1.5.1.1 Alignement de deux séquences

L'alignement de deux séquences vise à identifier les régions communes à ces deux séquences. Il repose sur l'hypothèse de microévolution par mutations ponctuelles qui sont des substitutions, des insertions ou des délétions de résidus. Les insertions et les délétions (indels) sont représentées dans l'alignement par des gaps qui sont le plus souvent symbolisés par un ou plusieurs caractères '-'.

Plusieurs algorithmes ont été développés pour comparer des séquences. Les algorithmes de référence sont celui de Needleman et Wunsch (Needleman and Wunsch, 1970) pour chercher la similarité globale entre deux séquences et celui de Smith et

Waterman (Smith and Waterman, 1981) pour chercher des similarités locales entre deux séquences.

Parmi l'ensemble des alignements possibles, il faut trouver celui ou ceux qui sont optimaux. Mathématiquement, il s'agit de trouver le chemin optimum au sein d'un graphe. La technique de recherche du meilleur chemin (alignement optimal) la plus utilisée est la programmation dynamique.

Appliquées au problème des alignements de séquences, les méthodes de programmation dynamique construisent un alignement optimal de sous-séquences de plus en plus longues en utilisant les scores obtenus pour les sous-séquences.

Needleman et Wunsch ont été les premiers à utiliser la programmation dynamique pour l'alignement de séquences biologiques (Needleman and Wunsch, 1970). Dans leur algorithme, l'alignement optimal recherché devait inclure la totalité des deux séquences à aligner définissant ainsi un alignement global. Ce type d'alignement s'applique bien dans le cas où les séquences présentent une grande similarité sur toute leur longueur. Cependant, ce n'est pas toujours le cas, surtout quand on essaie d'aligner des séquences de protéines multi-domaines. C'est pourquoi cet algorithme a été modifié, par Smith et Waterman (Smith and Waterman, 1981), afin de rechercher l'alignement optimal local qui débute et se termine à l'intérieur du graphe de recherche.

Le but de l'algorithme de Smith et Waterman est de trouver la paire de segments, provenant de deux séquences, telle qu'il n'y ait aucune autre paire de segments ayant plus de similarités entre elles. Pour trouver la paire de segments de plus haut score, il faut chercher l'alignement qui minimise le nombre de substitutions, insertions et délétions.

1.5.1.1.1 Matrices de substitutions

Il existe peu de matrices pour les acides nucléiques car il n'y que 4 lettres dans leur alphabet. La plus fréquemment utilisées est la matrice identité ou unitaire, où une valeur positive est affectée en cas d'identité.

Il existe plusieurs types de matrices protéiques. Les plus anciennes sont les matrices PAM (« Point Accepted Mutation »), qui ont été créées par Magaret Dayhoff et ses collaborateurs, après l'étude de 1572 changements d'aa dans 71 familles de protéines (Dayhoff et al., 1978). Ce type de matrice donne la probabilité que, suite à une mutation par substitution au cours de l'évolution, n'importe quel acide aminé remplace n'importe quel autre acide aminé sans que la fonction de la protéine ne soit altérée.

La première matrice de ce type, appelée 1PAM donne la probabilité qu'une substitution soit acceptée pour 100 acides aminés. Si ces changements étaient purement aléatoires, la fréquence de chaque substitution serait déterminée par la fréquence de chaque acide aminé. Dans les protéines reliées, les fréquences des substitutions sont biaisées par les substitutions qui sont favorables à la fonction de la protéine.

La multiplication X fois de cette matrice par elle-même donne une matrice XPAM qui permet d'analyser des distances d'évolution plus importantes : 2PAM = 1PAM x 1PAM, 3PAM = 2PAM x 1PAM, *etc*.

A chaque matrice XPAM correspond une matrice PAMX. Ce sont les matrices PAMX qui sont utilisées par les algorithmes d'alignement.

Les matrices avec une forte valeur PAM sont utilisées pour l'alignement de séquences fortement divergentes alors que celles avec une faible valeur de PAM sont utilisées pour l'alignement des séquences proches.

Les matrices BLOSUM (BLOcks SUbstitution Matrix) (Henikoff and Henikoff, 1992), développées après les matrices PAM, sont construites à partir de 2000 « blocks » d'alignements provenant de plus de 500 familles de protéines. Les BLOCKS sont des régions conservées de familles de protéines ne contenant pas d'insertions ou de délétions, et sont maintenus à jour dans la base de données BLOCKS.

A partir d'alignements de portions de séquences très conservées, des BLOCKS d'acides aminés sont obtenus. Puis, un sous-ensemble contenant les portions de séquences qui révèlent un pourcentage donné d'identité est constitué. Ainsi, pour la matrice BLOSUM62, les séquences présentant plus de 62% d'identité de séquence ont été rassemblées en une seule famille. Une série de matrices BLOSUM existe. Elles ont été calculées entre 30% et 90% d'identité. La matrice BLOSUM30 est adaptée à la comparaison de séquences fortement divergentes alors que la matrice BLOSUM90 est utilisée pour la comparaison de séquences proches.

Les matrices PAM250 et BLOSUM62 constituent le meilleur compromis pour comparer les séquences protéiques.

1.5.1.1.2 Pénalités pour les insertions et délétions (pénalités de gaps)

Pour ne pas obtenir des alignements trop segmentés, les notions de pénalités d'ouverture et d'extension de gaps ont été introduites. En effet, chaque ouverture de gap

résulte en une diminution du score de l'alignement. Généralement, la pénalité d'extension d'un gap est moins coûteuse pour le score qu'une pénalité d'ouverture.

Ainsi, la valeur d'une pénalisation d'une délétion de longueur k est représentée par la fonction affine : $W_k = r + kt$, avec *r* la pénalité d'ouverture et *t* la pénalité d'extension. Ces valeurs ont été déterminées de manière empirique, et elles sont proposées comme valeurs par défaut par les auteurs des programmes.

1.5.1.1.3 Signification statistique

A l'issue de l'alignement des deux séquences, un score de similarité est calculé. Il faut alors estimer la signification statistique de ce score afin de savoir s'il reflète une homologie ou non.

Pour les alignements locaux non gappés (High-scoring Segment Pair, HSP), Karlin et Altschul ont introduit une fonction de densité de probabilité qui suit une distribution de valeur extrême qui permet l'estimation de la distribution des scores (Karlin and Altschul, 1990). Dans cette distribution, caractérisée par deux valeurs λ et K qui sont estimées à partir de la matrice de substitution et de la composition en acides aminés des séquences, et pour des longueurs de séquences n et m, le nombre attendu de HSP avec un score d'au moins S est donnée par la formule

$$E = Kmne^{-\lambda S}$$

avec S le score de similarité nominal. On l'appelle « E-value » (Expected value) pour le score S.

Si l'on ne connaît pas le système de calcul du score, il est difficile d'interpréter sa valeur. C'est pourquoi, les scores normalisés ont été introduits. Ils sont calculés par :

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

C'est un score exprimé en bits. La E-value ou valeur attendue correspondant à ce score normalisé est calculée par la formule :

$$E = mn2^{-S'}$$

La valeur E représente le nombre d'alignements de score égal ou supérieur attendu par le hasard. Cette valeur E, utilisée par les programmes de recherche d'homologie, permet d'inférer l'homologie qui est vraie pour les valeurs E faibles. Cependant, elle ne permet pas de réfuter l'homologie pour des séquences ayant une valeur E forte qui peuvent être homologues.

1.5.1.2 Algorithmes de recherche de séquences similaires dans une banque de séquences

Des méthodes heuristiques ont été développées pour chercher rapidement des similarités entre une séquence et une banque de séquences.

Les programmes efficaces de recherche de similarité sont ceux qui présentent le meilleur compromis entre sensibilité, spécificité et vitesse d'exécution. La sensibilité est la capacité de détection de paire de séquences reliées (vrais positifs) même si elles sont éloignées. La spécificité est la capacité à ne pas inclure des séquences qui ne sont pas reliées (faux positifs) parmi celles qui le sont.

L'utilisation d'heuristiques peut engendrer le risque d'une diminution de sensibilité. Une des méthodes heuristiques la plus largement utilisée est la recherche de mots communs (successions de plusieurs lettres identiques) aux deux séquences. Le principe étant que dans un alignement de séquences reliées, il existe au moins un mot commun à ces séquences (Wilbur and Lipman, 1983).

Les méthodes heuristiques les plus utilisées sont FASTA (Pearson and Lipman, 1988) et BLAST (Altschul et al., 1990).

1.5.1.2.1 FASTA

Le programme FASTA suit une méthode largement heuristique qui contribue à la grande vitesse de son exécution. Il procède en 4 étapes principales :

1. FASTA réalise une grande partie de sa vitesse et de la sélectivité dans la première étape, en utilisant une table de consultation pour trouver toutes les identités ou groupes d'identités entre deux séquences d'ADN ou de protéines au cours de la première étape de la comparaison. Le paramètre *ktup* détermine combien d'identités consécutives sont nécessaires dans une correspondance. Par exemple, si *ktup* = 4 pour une comparaison de séquences d'ADN, seules les séries de 4 identités consécutives sont examinées. Dans la première étape, les 10 meilleures régions diagonales sont trouvés en utilisant une formule simple basée sur le nombre de correspondances *ktup* et la distance entre les correspondances sans tenir compte identités plus courtes, des substitutions conservatrice, des insertions ou délétions. Un score (*init1*) leur est attribué.
2. La deuxième étape est l'évaluation des 10 régions ayant les plus hauts scores. Les scores sont recalculés (initn) à l'aide d'une matrice de substitution. Pour chaque

diagonale, une sous-région avec un score maximal est identifiée, et appelée « région initiale ».

3. FASTA vérifie si plusieurs régions initiales peuvent être reliées entre elles. Les régions initiales sont réunies à chaque fois que leur score, diminué d'une pénalité de jonction, est supérieur ou égal au score *init1*. Ensuite, un alignement local optimisé est calculé dans une bande étroite du graphe de recherche englobant ces régions et donne lieu au score *opt*.
4. Enfin, dans la quatrième étape de la comparaison, les séquences avec les scores les plus élevés sont alignées en utilisant une modification de la méthode d'optimisation décrite par Smith et Waterman.

Le programme calcule un Z-score qui est déduit de la valeur Z :

$$Z = \frac{(score - moyenne(scores))}{\sigma(scores)}$$

avec σ l'écart-type. La distribution de ce score suit la distribution de la valeur extrême (Pearson, 1998). FASTA fournit également la valeur E attendue (E-value).

Le programme FASTA est une méthode d'alignement local car sa méthode de recherche s'appuie sur la recherche de segments.

1.5.1.2.2 BLAST

L'algorithme BLAST (Basic Local Alignment Search Tool) a amélioré la vitesse de recherche et a permis d'évaluer la signification statistique des alignements obtenus (Altschul et al., 1990). Par la suite, la prise en compte explicite des gaps a été introduite (Altschul et al., 1997).

Le principe du programme est de découper la séquence requête en mots élémentaires (3 acides aminés, ou 11 nucléotides par exemple) et de rechercher, comme FASTA, tous les mots de la base de données qui s'alignent avec ce mot au dessus d'un score seuil. Cependant, une innovation a été apportée qui est le calcul de mots voisins. En effet, à l'aide de la matrice de substitution (BLOSUM 62 pour les protéines), des mots voisins de ceux de la séquence requête sont retenus si leur score est supérieur à un seuil T.

Dans le programme BLAST original, chacun de ces «hit» était ensuite étendu, pour vérifier s'il était contenu dans un alignement de score élevé. Pour la valeur par défaut T, cette étape consomme la majorité du temps de traitement. La deuxième méthode, dite à deux hits, suppose l'existence de plus d'une paire de mots au sein d'une HSP (Altschul et

al., 1997). Ainsi, dans une fenêtre de A résidus, si deux hits non chevauchants sont présents, l'extension de l'alignement est effectuée pour le second hit. Le temps d'extension des alignements représentant 90% du temps de traitement de BLAST, en diminuant le nombre de segments à étendre (au moins deux hits) on augmente la vitesse de traitement. Si l'HSP obtenue après l'extension possède un score nominal supérieur à une valeur seuil Sg, un alignement gappé est alors recherché.

Pour construire l'alignement gappé, une graine (seed) est choisie en fonction de la longueur de l'HSP obtenue. Si la longueur de l'HSP est supérieure à 11 résidus, le segment de 11 résidus de score le plus élevé est cherché et la paire de résidus au centre constitue la graine. Sinon, une paire centrale de résidus est choisie. A partir de cette graine, un processus de programmation dynamique étend l'alignement à gauche et à droite de celle-ci. Dans cette procédure, les cellules utilisées sont celles qui ne font pas chuter le score de l'alignement local d'une valeur supérieure à Xg par rapport au meilleur score trouvé jusque-là.

Ainsi, contrairement à FASTA, la région explorée est adaptée en fonction des données. De plus, cette procédure pouvant être répétée pour d'autres HSP, de nouveaux alignements gappés seront présentés s'ils ne recouvrent pas ceux précédemment calculés, conférant un avantage supplémentaire sur FASTA qui ne fournit que le meilleur alignement.

Une valeur E attendue (E-value) est attribuée à chaque alignement. Elle représente le nombre d'alignements de séquences de longueurs égales (m et n) et de score égal ou supérieur, attendus par le hasard, et dépend également de la taille de la banque de séquences utilisée.

1.5.1.3 Méthodes des profils

D'autres méthodes permettent de chercher des séquences homologues, en utilisant des profils. Ces méthodes nécessitent d'avoir généré préalablement un alignement multiple de séquences homologues pour lesquelles on cherche d'autres homologues. Pour certaines de ces méthodes, la phase de recherche d'homologues et leur alignement pour construire le profil, est incluse dans le processus. Nous décrirons des méthodes utilisant deux types de profils : les matrices PSSM (« position-specific score matrix ») et les profils HMM (« Hidden Markov Model »).

1.5.1.3.1 L'algorithme PSI-BLAST

Le programme PSI-BLAST (Position-Specific Iterative BLAST) est utilisé pour trouver des homologues distants d'une séquence protéique. Tout d'abord, un BLAST est fait à partir de la séquence requête pour récupérer les séquences étroitement apparentées. Les séquences similaires significatives sont alignées sur la séquence requête. La distribution des résidus et des insertions dans les colonnes de l'alignement, permet de calculer une fréquence d'occurrence des résidus à chaque position, qui peut être traduite en score. La matrice des scores résultante est appelée profil ou matrice PSSM.

Le profil est utilisé pour chercher de nouvelles séquences dans la base de données s'alignant avec ce profil. De manière itérative, quand de nouvelles séquences sont trouvées de manière significative, elles sont utilisées pour construire un nouveau profil qui sera utilisé pour une nouvelle recherche, et le processus est répété.

Le fait d'utiliser des séquences homologues à la recherche, rend PSI-BLAST beaucoup plus sensible pour détecter des relations évolutives lointaines, contrairement au BLAST classique (Altschul et al., 1997).

1.5.1.3.2 Le logiciel HMMER

HMMER est utilisé pour la recherche de séquences homologues à des séquences protéiques dans les bases de données, et pour la construction d'alignements de séquences protéiques. Il met en œuvre des méthodes utilisant des modèles probabilistes appelés profils modèles de Markov cachés (profil HMM).

Les HMM permettent de créer un modèle statistique d'un alignement multiple, dans lequel l'état n ne dépend que de l'état n-1. On peut alors modéliser un alignement par une chaîne d'éléments à 3 états :

- M est l'état qui correspond à une position alignée
- I correspond à une insertion
- D correspond à une délétion

Des probabilités d'émission et de transition sont attribuées à ces 3 états et entre eux.

Les probabilités qui sous-tendent un alignement sont inconnues, il faut les estimer à partir des fréquences d'occurrence observées à chaque position. Cette information est enrichie par la connaissance *a priori* des probabilités d'occurrence des acides aminés.

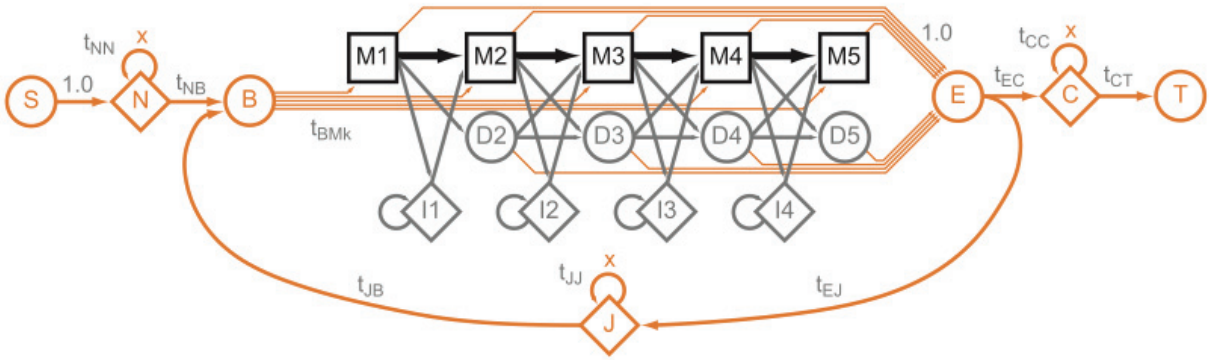


Figure 19 : Schéma représentant la recherche à l'aide d'un profil HMM.
(Eddy, 2011)

La Figure 19 représente la recherche de séquences avec le profil. Les régions alignées à la requête sont représentées par un modèle de base linéaire consistant en M positions consensus (dans cet exemple, M ~ 5), chacune avec un état de match, de délétion, et d'insertion (représentées par des carrés marqués M, cercles marqués D et losanges marqués I), reliés par des probabilités de transition entre les états (flèches). Les états de match portent des probabilités d'émission position-spécifiques pour le calcul du score des résidus à chaque position consensus. Les états d'insertion émettent des résidus avec des probabilités d'émission identiques à une distribution d'arrière-plan. Les états supplémentaires marqués N, C, et J, émettent zéro ou plusieurs résidus de la distribution d'arrière-plan, modélisant des régions non homologues précédant, suivant ou se joignant à des régions homologues alignées sur le modèle de base. Les cercles marqués S (start), B (begin), E (end), et T (termination) correspondent respectivement au départ de la recherche, au début et à la fin du profil, et à l'état de terminaison, et ces états n'émettent pas de résidus (Eddy, 2011).

Lorsque les paramètres sont des probabilités plutôt que les scores arbitraires, ils sont plus facilement optimisés par des critères mathématiques objectifs. Cela permet de construire des modèles plus complexes, biologiquement réalistes avec un grand nombre de paramètres. En effet, un profil HMM utilise les probabilités d'insertion / délétion position spécifique, plutôt qu'un score arbitraire tel qu'une pénalité de gap, comme l'utilise BLAST. Ceci permet au profil HMM de modéliser le fait que les indels se produisent plus fréquemment dans certaines parties d'une protéine que dans d'autres (par exemple, dans la surface de boucles par opposition à cœur enfoui) (Eddy, 2008).

L'algorithme de Viterbi (Viterbi, 1967) permet de trouver des séquences reliées aux séquences de l'alignement, en identifiant la trajectoire la plus probable au sein du modèle HMM.

En réalité, on calcule la probabilité qu'un profil HMM donné génère la séquence alignée de façon optimale.

Le programme *hmmbuild* permet de construire des profils HMM à partir d'alignements multiples. Le programme *hmmsearch* permet de chercher dans une banque de séquences, celles qui « s'alignent » avec le profil HMM. Quant au programme *hmmalign*, il permet de générer un alignement multiple de séquences selon un profil HMM donné.

1.5.2 Alignements multiples

Une fois que l'on a récupéré plusieurs séquences homologues par l'une (ou plusieurs) des différentes méthodes citées précédemment, on peut les aligner afin d'établir les relations d'évolution et de mettre en évidence des motifs communs, jouant un rôle dans la fonction ou la structure des protéines homologues.

On ne pourrait pas calculer un alignement multiple par la méthode de programmation dynamique, car cela demanderait trop de temps de calcul et d'espace mémoire. D'autres stratégies ont alors été développées pour aligner plusieurs séquences. Les premières méthodes apparues sont des méthodes progressives. Dans ces méthodes, les séquences sont comparées par paires et l'alignement multiple est construit progressivement en partant de la paire la plus proche.

L'un des programmes d'alignement multiple progressif les plus utilisés est Clustal W.

1.5.2.1 Clustal W

On peut décomposer le fonctionnement de Clustal W en 3 grandes étapes (Thompson et al., 1994). Dans la première étape, les alignements de toutes les paires de séquences sont réalisés (alignement global), afin de générer une matrice de distances entre toutes les séquences. Ensuite, un arbre guide est construit à partir de cette matrice de distance, grâce à un algorithme de neighbor-joining. Enfin, les séquences sont alignées en utilisant l'arbre guide. Chaque paire de séquences situées sur une même branche extérieure est alignée par programmation dynamique. Les alignements permettent de créer des profils, avec la fréquence observée de chaque acide aminé à

chaque position. Ensuite, les profils associés par un même nœud de l'arbre sont alignés. Cet alignement de séquences puis de profils se poursuit de manière récursive jusqu'à l'alignement final, en procédant des feuilles de l'arbre vers la racine.

Au cours de l'alignement des paires, des matrices de substitution et des pénalités d'ouverture et d'extension de gaps sont impliquées.

Au fur et à mesure de la construction de l'alignement, des gaps sont créés. A partir des pénalités d'ouverture et d'extension des gaps fournies par l'utilisateur, de nouvelles pénalités de gaps spécifiques de la position sont calculées pour étendre les gaps existants plutôt que d'en créer de nouveaux.

Les algorithmes d'alignements progressifs comportent des limitations, et peuvent conduire à des erreurs lorsque les séquences sont trop divergentes. C'est pour cette raison que d'autres stratégies ont été développées. Parmi elles, ont retrouve des approches itératives. C'est le cas de l'algorithme MUSCLE.

1.5.2.2 MUSCLE

Les éléments de l'algorithme comprennent l'estimation rapide des distances en utilisant un comptage *kmer*, un alignement progressif en utilisant une nouvelle fonction de profil, et un affinement utilisant un partitionnement restreint dépendant de l'arbre (Edgar, 2004).

Les séquences sont converties en un alphabet réduit qui permet d'identifier rapidement les segments similaires (kmer) et de construire l'arbre guide en un minimum de temps.

Un kmer est une sous-séquence contiguë de longueur k, également désigné par mot ou k-tuple. Les séquences apparentées ont tendance à avoir plus de kmers en commun que ce qu'on peut attendre par le hasard.

Après la construction de l'arbre guide, les séquences sont alignées de façon progressive. Puis, l'algorithme réaligne les séquences entre elles de manière itérative, en extrayant deux sous alignements multiples et en les réalignant par des méthodes plus précises (comparaison profil-profil).

1.5.2.3 Entropie de Shannon et logos de séquences

A partir d'un alignement on peut déterminer des motifs conservés, on peut aussi étudier les fréquences des résidus à chaque position. Par exemple, la méthode des logos,

permet de présenter une séquence consensus de l'alignement (Schneider and Stephens, 1990). Ceci en montrant les acides aminés les plus représentés à une position de l'alignement avec des lettres de plus grande taille. En fait, la méthode des logos concentre plusieurs informations sur un même graphique :

- le consensus général des séquences
- l'ordre de prédominance des résidus à toutes les positions
- les fréquences relatives de tous les résidus à toutes les positions
- la quantité d'information présente à chaque position, mesurée en bits
- un point d'initiation, un point de coupure

Si les fréquences de bases ne sont pas exactement équiprobables, alors un calcul plus complexe est nécessaire pour trouver l'information moyenne à une position. En 1948, Shannon a montré comment faire ceci (Shannon et al., 1948). A partir des études de Shannon, on peut définir la mesure de l'incertitude par cette formule :

$$H(l) = -\sum_{b=a}^{t} f(b,l)\log_2 f(b,l) \quad \text{(bits)}$$

où *H(l)* est l'incertitude à la position *l*, *b* est l'un des résidus, et *f(l,b)* est la fréquence du résidu *b* à la position *l*. On peut déduire de cette formule la quantité d'information à la position *l* :

$R(l) = 2 - (H(l) + e(n))$ (Séquences nucléotidiques),

$R(l) = \log_2 20 - (H(l) + e(n))$ (Séquences protéiques),

où *e(n)* est un facteur correctif, requis lorsque le nombre de séquences *n* est faible.

La taille du résidu *b* dans le logo peut ainsi être déduite par le produit : $f(b,l)R(l)$.

1.5.3 Prédiction de la structure secondaire des protéines

Les méthodes de prédiction de la structure secondaire des protéines peuvent servir à connaître la classe structurale à laquelle la protéine appartient, et peuvent aider pour la modélisation de sa structure tridimensionnelle. La qualité de prédiction est notamment estimée par le paramètre Q_3, qui représente le nombre de résidus correctement prédits sur le nombre total de résidus dans une prédiction en trois états (hélice, brin et apériodique).

Les premières méthodes de prédiction de structures secondaires utilisaient une seule séquence pour la prédiction. En effet, dans les années 70, Chou et Fasman (Chou

and Fasman, 1978) développent une méthode qui utilise les préférences de chaque acide aminé pour chaque état conformationnel, associées à un ensemble de règles empiriques. Sa qualité de prédiction Q_3 est évaluée à 52%.

Par la suite, d'autres méthodes reposant sur des alignements de séquences par familles de protéines ont été développées. Dans un premier temps, des séquences homologues à la séquence requête sont récupérées. Les régions conservées chez les membres d'une famille de protéines sont supposées avoir un repliement similaire et ainsi les mêmes structures secondaires. Il existe plusieurs types de méthodes. Je présenterai une méthode utilisant des statistiques linéaires, une méthode basée sur l'homologie, et une basée sur les méthodes d'apprentissage (réseaux neuronaux).

1.5.3.1 La méthode DSC

DSC est une méthode de prédiction de structures secondaires à partir d'un alignement multiple de protéines homologues avec un Q_3 de 70,1%. Elle a deux objectifs : obtenir une grande précision grâce à l'identification d'un ensemble de concepts importants pour la prédiction ainsi par l'utilisation d'une fonction de discrimination à partir de ces concepts (statistiques linéaires), et de donner un aperçu du repliement.

Les concepts importants dans la prédiction de structures secondaires sont identifiés comme : les tendances de conformation des résidus, les effets de bord des séquences, les régions d'hydrophobie, la position des insertions et délétions dans l'alignement des séquences homologues, les régions de conservation, l'autocorrélation des résidus, les ratios des résidus, des effets de rétroaction des structures secondaires, et le filtrage (King and Sternberg, 1996).

1.5.3.2 Les méthodes SOPM et SOPMA

Une méthode appelée méthode de prédiction auto-optimisée (SOPM) a été décrite pour améliorer le taux de succès dans la prédiction des structures secondaires des protéines. Elle est basée sur la méthode de Levin (Levin and Garnier, 1988) qui repose sur l'idée que de courtes séquences peptidiques similaires adoptent un repliement identique. Ainsi, dans cette méthode, la séquence est découpée en heptapeptides qui sont comparés à des heptapeptides répertoriés dans une base de données.

Dans la méthode SOPM, les paramètres sont optimisés à partir d'un sous-ensemble de la base de données de référence, qui est construit avec les séquences similaires à la séquence requête (Geourjon and Deléage, 1994).

La méthode SOPMA apporte des améliorations à la méthode SOPM par la prédiction de toutes les séquences d'un ensemble de protéines alignées appartenant à la même famille. Le Q_3 est de 72,5% (Geourjon and Deléage, 1995).

1.5.3.3 La méthode PHD

Cette méthode combine les résultats de plusieurs réseaux neuronaux. Chacun des réseaux prédit la structure secondaire sur la base du contexte de séquence local et sur les caractéristiques globales. La prédiction finale est une moyenne arithmétique du résultat de chacun des réseaux de neurones. La prédiction atteint une qualité Q_3 de 72,5% (Rost and Sander, 1993).

L'idéal pour l'analyse des structures secondaires prédites, c'est de combiner plusieurs méthodes et de vérifier l'accord entre ces méthodes, ou le consensus. Ceci est possible sur le serveur d'analyse NPS@ (Combet et al., 2000).

1.5.4 Prédiction de la structure tertiaire des protéines

Il existe plusieurs méthodes de prédiction de la structure tridimensionnelle des protéines :

- la modélisation par homologie
- la reconnaissance de repliement ou « threading », qui utilise des structures disponibles pour identifier le repliement le plus adapté à une séquence
- la modélisation *de novo* (ou *ab-initio*)

Dans cette thèse, nous nous intéresserons uniquement à la modélisation par homologie. Elle consiste à prédire la structure tridimensionnelle d'une protéine à partir d'une structure résolue d'une protéine homologue, c'est l'empreinte. L'homologie est souvent inférée à partir du pourcentage d'identité entre les séquences. Au delà de 25% d'identité entre 2 séquences, on peut affirmer qu'elles sont homologues (Geourjon et al., 2001). En dessous de ce seuil, il y a toujours possibilité d'homologie, mais lointaine et donc difficile à prouver. Dans ce cas, le choix de l'empreinte doit être guidé par d'autres éléments tels que les structures secondaires (Geourjon et al., 2001) ou la transitivité de

la propriété d'homologie. Une fois l'empreinte sélectionnée, l'alignement entre les deux séquences est une étape très importante de la modélisation. La construction du modèle peut se faire par des méthodes de substitution moléculaire ou en mesurant des contraintes d'angles et de distances sur la structure empreinte.

1.6 Bases de données et outils spécialisés pour le virus de l'hépatite B

Afin de permettre aux chercheurs d'étudier la variabilité génétique des séquences du VHB et la résistance virale au traitement, plusieurs bases de données et dépôts de séquences ont été publiés à ce jour (HepSEQ au Royaume Uni, Hepatitis Virus Database au Japon, HBVRegDB en Nouvelle Zélande, HBVrtDB aux Etats Unis) (Gnaneshan et al., 2007; Shin-I et al., 2008; Panjaworayan et al., 2007; Rhee et al., 2010). Certaines de ces bases proposent des outils d'analyse de séquences associés. En outre, des outils de génotypage des séquences du VHB, indépendants de ces bases, sont également disponibles pour les virologues. La plupart des bases sont accessibles librement et d'autres, telles que SeqHepB (Australie), nécessitent un enregistrement (Yuen et al., 2007).

1.6.1 Bases de données VHB et outils associés

1.6.1.1 HepSEQ (Royaume-Uni)

HepSEQ été développée comme une base de données avec contrôle qualité manuel pour agir comme un outil pour la surveillance, pour la gestion des cas d'infection par le VHB et pour la recherche (Gnaneshan et al., 2007).

Le format de la base de données permet le dépôt de données moléculaires, cliniques et épidémiologiques détaillées, de rechercher et de manipuler les données stockées. Il permet également d'extraire et de visualiser les informations épidémiologiques, virologiques, cliniques, les séquences nucléotidiques et les mutations liées à l'infection par le VHB, via une interface web.

Des outils spécifiques, intégrés à la base de données, peuvent être utilisés pour analyser les données déposées et fournir de l'information sur le génotype du VHB,

identifier les mutations qui ont une importance clinique connue (par exemple échappement à la vaccination, mutations précore et de résistance aux antiviraux) et d'effectuer des recherches d'homologie de séquence par rapport aux autres souches déposées.

En octobre 2012, HepSEQ contient 3668 séquences virales dont 28 génomes complets et 2875 entrées de patients (Figure 20). L'annotation et le contrôle qualité des données sont effectués manuellement par les personnes les soumettant, et sont vérifiées par des curateurs qui valident l'entrée dans la base. Les données contenues dans HepSEQ proviennent majoritairement du Royaume Uni (Myers et al., 2008).

Il faut être enregistré pour pouvoir accéder aux données (l'enregistrement est gratuit pour les académiques), mais les outils d'analyse de séquences sont accessibles sans enregistrement.

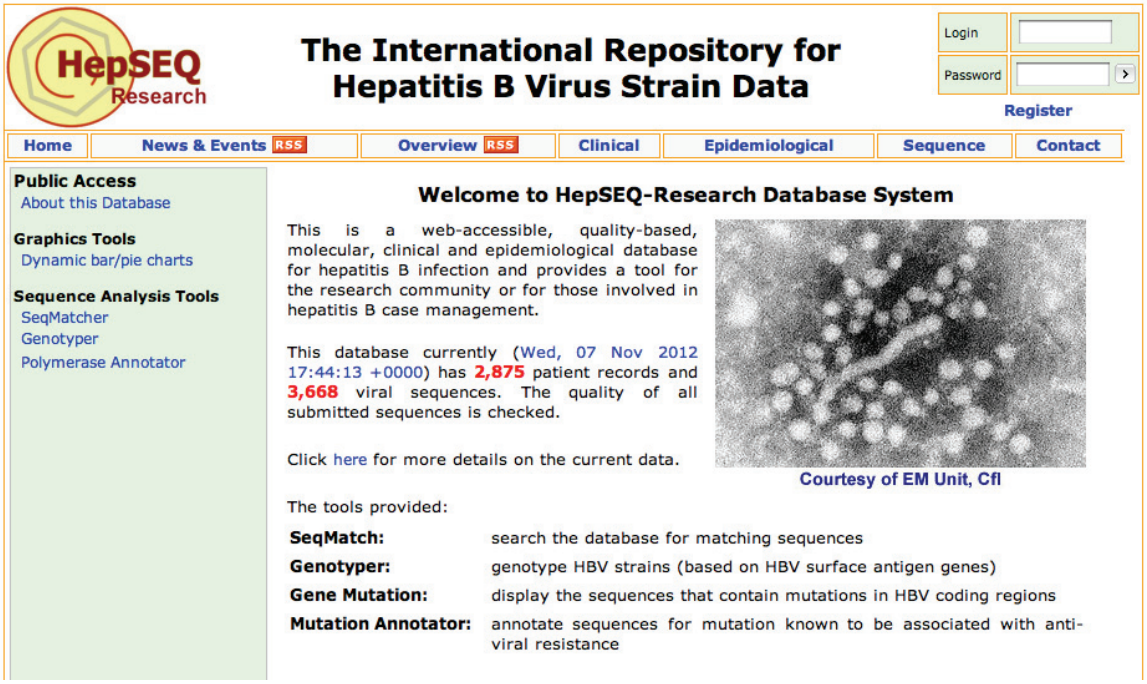


Figure 20 : Page d'accueil de la base HepSEQ.

1.6.1.2 Hepatitis Virus Database (Japon)

La base de données sur les hépatites virales ou « Hepatitis Virus Database » (HVDB), est accessible par le web et contient toutes les séquences du VHC (virus de d'hépatite C), du VHB, et les séquences de VHE (virus de l'hépatite E) disponibles à l'INSDC (Shin-I et al., 2008). Dans HVDB, toutes les séquences obtenues à partir de l'INSDC sont ordonnées en fonction du génome de chaque virus. HVDB donne également les relations phylogénétiques entre chaque locus du génome et les variants de chaque virus.

Les utilisateurs de la base de données peuvent facilement récupérer les entrées (séquences avec des annotations) d'un génotype spécifique en se référant aux relations phylogénétiques ou aux loci spécifiques présentés sur la carte du génome. HVDB fournit aux utilisateurs un outil pour l'analyse phylogénétique qui peut être utilisé en combinaison avec des outils d'extraction de données.

Il y a des divisions pour chaque locus du génome ; ces divisions peuvent contenir des alignements de séquences nucléotidiques ou des alignements des séquences protéiques traduites. Les annotations sont extraites des entrées INSDC.

HVDB fournit une « carte d'information » qui peut être utilisée pour localiser une entrée donnée sur la séquence de référence. Les utilisateurs peuvent récupérer toutes les entrées qui couvrent une région spécifique en se référant à cette « carte d'information ». Les utilisateurs peuvent également récupérer les entrées qui appartiennent à un cluster spécifique à partir d'un arbre phylogénétique.

La base HVDB contient 38336 entrées pour le VHB, en octobre 2012.

1.6.1.3 HBVRegDB (Nouvelle-Zélande)

HBVRegDB est un outil d'analyse génomique comparative intégré avec une base de données de séquences. La base de données contient des séquences génomiques de virus représentatifs (Panjaworayan et al., 2007). Elle fournit des ressources de séquences régulatrices des génomes d'*Hepadnaviridae* avec des annotations et des liens vers des ressources connexes. En plus des annotations de l'INSDC et de RefSeq (base non redondante de génomes, transcrits et séquences protéiques manuellement annotés) (Pruitt et al., 2007), HBVRegDB contient aussi des annotations systématiquement calculées (par exemple les promoteurs) et des résultats d'analyses comparatives de génomes. Elle contient également des analyses basées sur des alignements vérifiés de séquences du VHB. Des informations sur les régions conservées et des prédictions de structures secondaires d'ARN sont intégrées dans la base de données.

1.6.1.4 HBVrtDB (Etats-Unis)

HBVrtDB est une base de données des variants de la transcriptase inverse (RT) du VHB (Rhee et al., 2010). Elle a pour but de caractériser les associations entre les mutations dans la RT du VHB et les traitements aux analogues de nucléos(t)ides (NA) chez des individus pour lesquels des séquences ont été obtenues. La base de données présente ces associations dans le cadre du génotype viral et de l'origine géographique.

23871 séquences de RT du VHB ont été récupérées dans GenBank et triées (grâce aux publications) afin de connaître le nombre de personnes pour lesquelles des séquences ont été obtenues, les traitements reçus, ainsi que l'année et la région de l'échantillonnage virus. Ces données ont été utilisées pour remplir la base de données relationnelle HBVrtDB. En juillet 2010, HBVrtDB contenait 6811 séquences de 3869 individus. Parmi ces 3869 personnes, 73% étaient des patients naïfs de traitement aux NA et 27% avaient reçu un ou plusieurs NA.

HBVrtDB a été construite par curation et annotation de plus de 250 études dans GenBank et a été complétée par la contribution des séquences du VHB bien caractérisées de deux grandes populations cliniques (443 individus venant d'une clinique aux Etats-Unis et d'une autre en Allemagne : numéros d'accession HM173808 à HM174250). HBVrtDB fournit de nouvelles données sur l'ampleur du polymorphisme à chaque position de RT selon le génotype, et sur la prévalence relative de chacune des mutations bien caractérisées de résistance aux NA.

Un outil d'analyse de séquences de RT, HBVseq, est intégré à la base HBVrtDB.

HBVseq permet aux utilisateurs d'identifier des mutations dans les séquences qu'ils soumettent et de récupérer la prévalence de ces mutations dans HBVrtDB en fonction du génotype et du traitement aux NA.

1.6.1.5 SeqHepB (Australie)

SeqHepB est la combinaison d'un programme d'analyse de séquences génomiques du VHB et d'une base de données relationnelle qui héberge les données recueillies à partir de plusieurs sources (Yuen et al., 2007). Son utilisation requiert un enregistrement. Les utilisateurs enregistrés peuvent accéder à la composante d'analyse de séquences en ligne. Sa fonction principale est de déterminer le génotype du VHB, d'identifier des mutations clés associées à la résistance aux antiviraux, et d'identifier les mutants du VHB cliniquement importants. Toutes les informations produites sont chargées dans une base de données et intégrées aux dossiers médicaux des patients, aux tests de pathologie et aux résultats supplémentaires de virologie. Il est possible d'extraire et de corréler les données cliniques, virologiques et des phénotypiques.

Les principaux objectifs de SeqHepB sont de prévenir les mauvaises prescriptions de traitement, d'améliorer l'efficacité de la surveillance du traitement et de l'apparition de mutants du VHB échappant à la vaccination parmi la population vaccinée, et d'améliorer

les lignes directrices de pratique professionnelle pour le traitement des patients atteints d'hépatite B chronique.

SeqHepB peut permettre aux virologues et aux médecins d'individualiser la gestion des patients, face à l'émergence des mutations de résistance aux antiviraux associées au VHB, et de mener des études transversales rétrospectives ou prospectives sur les individus infectés par le VHB pendant le traitement.

1.6.2 Outils de génotypage pour les séquences du VHB

1.6.2.1 NCBI Genotyping Tool

Le NCBI fournit un outil pour le génotypage de séquences virales, notamment pour le Virus de l'Immunodéficience Humaine (VIH), pour le Virus de l'Hépatite C (VHC) et aussi pour le VHB (Rozanov et al., 2004).

Il fonctionne en utilisant BLAST pour comparer une séquence requête à un ensemble de séquences de référence de génotypes connus. L'utilisateur peut également fournir son propre jeu de séquences de référence pour l'analyse. La séquence requête est divisée en segments pour la comparaison avec la référence afin que l'organisation en mosaïque de certaines séquences recombinantes puisse être révélé. L'algorithme procède donc par fenêtres glissantes et chevauchantes, dont les paramètres sont fixés à 300 nt pour la longueur de la fenêtre et 100 nt le pas d'incrémentation. Pour chaque fenêtre un BLAST est réalisé et le génotype de la séquence de référence la plus similaire est attribué à la fenêtre. Les résultats sont affichés graphiquement à l'aide d'une couleur par génotype. La Figure 21 présente un résultat de génotypage d'une séquence recombinante, voire mosaïque.

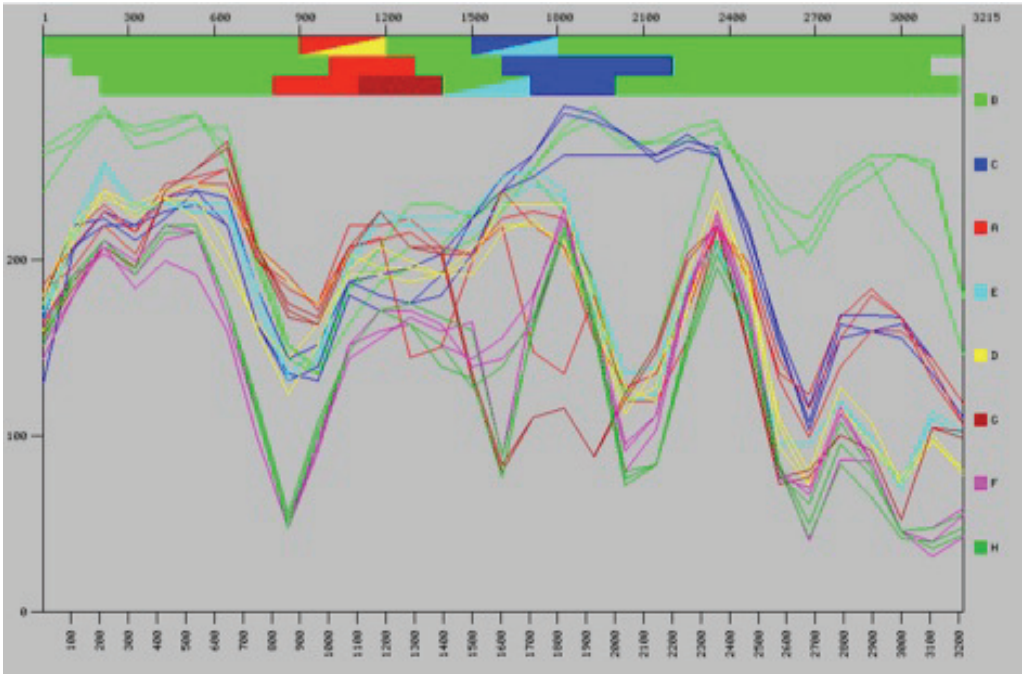


Figure 21 : Graphique de résultat du génotypage par l'outil du NBCI.
Le graphique montre un génotype B majoritaire (vert) avec des régions potentielles de recombinaison pour lesquelles la séquence est plus similaire à des séquences de génotype A (en rouge, entre 500 et 1200) et à des séquences de génotype B (en bleu, entre 1500 et 2100). Séquence de l'entrée ENA : AB231909.

1.6.2.2 Oxford Subtyping Tool

Il s'agit d'un système rapide et à haut débit de génotypage de séquences virales (Alcantara et al., 2009). Le procédé implique l'alignement d'une séquence requête avec un ensemble des séquences de référence sélectionnées. Ensuite, une analyse phylogénétique de plusieurs segments de l'alignement qui se chevauchent, est réalisée en utilisant un système de fenêtres glissantes (longueur de la fenêtre : 400 nt, longueur du pas d'incrémentation : 40 nt). Chaque segment de la séquence requête est associé au génotype de la séquence de référence avec le plus haut bootstrap (> 70%) et le plus haut score de bootscanning (> 90%). Les résultats de toutes les fenêtres sont combinés et affichés graphiquement à l'aide d'un code couleur par génotype.

1.6.2.3 jp-HMM HBV

L'outil jp-HMM (jumping HMM) avait déjà été utilisé pour génotyper les séquences du VIH (Schultz et al., 2006). Il a été adapté aux génomes circulaires pour pouvoir génotyper les séquences du VHB (Schultz et al., 2012).

Jumping HMM est un modèle probabiliste qui permet de comparer une séquence de nucléotide à un alignement multiple d'une famille de séquences. Compte tenu du

partitionnement de l'alignement en sous-classes, chaque sous-type est modélisé comme un profil HMM. En plus des transitions d'états habituelles dans un profil HMM, des transitions appelées «sauts» entre les différents profils HMM sont autorisées.

L'alignement de la séquence requête avec l'alignement multiple est alors défini par le chemin le plus probable à travers le modèle permettant de générer la séquence, le chemin de Viterbi (Viterbi, 1967), autorisant des sauts entre différents sous-types. Les positions des sauts entre les différents sous-types définissent les points de recombinaison, qui sont décrits par des intervalles.

Le problème pour le génotypage du VHB, réside dans le fait qu'il possède un génome circulaire. Pour prendre cette circularité en compte, et pour permettre de prédire des recombinaisons aux extrémités, jp-HMM crée une extension de la séquence requête au niveau de ces deux extrémités.

Les résultats sont présentés sous forme circulaire, on visualise le génotype le long de la séquence requête, sans les extrémités ajoutées pour la prédiction. Jp-HMM est très efficace pour la détection de recombinants (Figure 22).

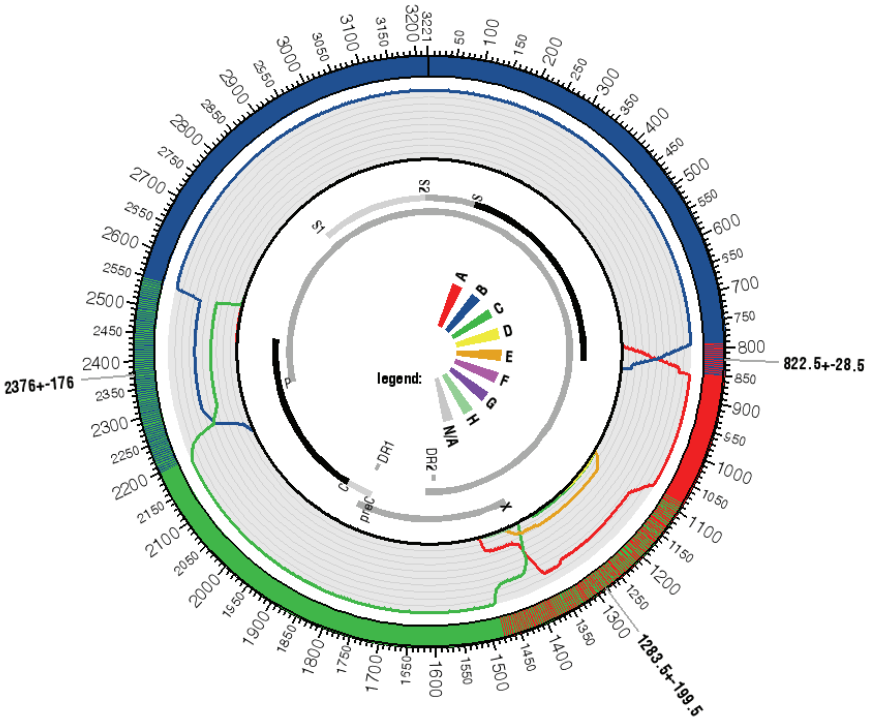


Figure 22 : Résultat du génotypage par jpHMM.
La figure montre qu'une partie de la séquence correspond à du génotype B (bleu) et que l'autre partie est plus similaire aux génotypes C (vert) puis A (rouge). Séquence de l'entrée AB231909.

1.6.2.4 HBV STAR

HBV STAR est un outil qui attribue le génotype sur la base d'un modèle de score défini de manière statistique. C'est une adaptation de l'outil précédemment développé pour génotyper les séquences du VIH : Subtype Analyser (STAR), aux séquences du VHB (Myers et al., 2006).

Cette méthode convertit les alignements spécifiques d'un génotype en matrices PSSM pour chaque génotype. Ensuite, elle compare la séquence requête de génotype inconnu à chaque PSSM. Les scores qui sont générés par cette comparaison (huit scores dans le cas du VHB), sont transformés en Z-scores, donnant la répartition en huit points d'une moyenne de zéro et d'un écart type de 1. La PSSM génotype qui génère le meilleur Z-score et la valeur de ce Z-score sont utilisés pour prédire le génotype de la séquence requête et pour donner un indice de confiance à la prédiction. Les Z-scores> 2,0 indiquent une prédiction significative du génotype.

Une fois la séquence requête génotypée, un processus séparé est exécuté pour l'analyse des séquences recombinantes.

La détection de recombinaisons est réalisée à l'aide de la différence dans l'identité de séquence par rapport au génotype attribué le long d'une fenêtre glissante de 150 nt avec un pas d'incrémentation de une base. Les séquences contenant un segment de plus de 150 nt où l'identité de séquence moyenne est plus semblable à un génotype différent et qui diverge du génotype attribué par plus de 1%, sont considérés comme des recombinants potentiels.

Chapitre 2 :

La base de connaissances HBVdb

Ce chapitre, structuré sous la forme d'un article, présente les travaux réalisés pour le développement de la base de connaissances HBVdb.

2.1 Introduction

Afin de permettre aux chercheurs d'étudier la variabilité génétique des séquences du VHB et la résistance virale aux traitements, nous avons développé une base de données spécifique au VHB, HBVdb (http://hbvdb.ibcp.fr), accessible librement par le web. Cette base doit contenir toutes les séquences du VHB humain issues de la banque de séquences nucléotidiques EMBL-Bank, avec une mise à jour automatique mensuelle.

L'objectif principal de cette base est de ré-annoter toutes ces séquences de manière automatique, et ce, avec un vocabulaire standardisé afin de rendre la recherche par mots clés beaucoup plus puissante et efficace. Ce processus permet d'annoter les séquences codantes ainsi que les protéines codées.

Cette procédure d'annotation automatique fait également intervenir une étape de génotypage, quand la longueur de la séquence le permet. Elle permet aussi, si la séquence de la transcriptase inverse (RT) est contenue dans la séquence à annoter, de détecter les mutations de résistance aux NA et ainsi de dresser le profil de résistance de la séquence.

Un autre objectif, en plus d'offrir l'accès à ces données, est de proposer à l'utilisateur des outils spécialisés pour l'analyse des séquences du VHB, via l'interface web de la base HBVdb.

Ce chapitre va tout d'abord exposer le matériel et les langages informatiques utilisés pour le développement de HBVdb. Il va décrire le processus de génération de la base de données ainsi que la procédure d'annotation automatique des entrées. Puis, l'interface web sera présentée en détail, avec une description des outils proposés à l'utilisateur pour l'analyse de ses propres séquences. Après quelques statistiques sur les données répertoriées dans HBVdb, je discuterai ces résultats et donnerai des perspectives quant à l'évolution de HBVdb.

Ces travaux ont fait l'objet d'une publication dans Nucleic Acid Research : Database Issue (article 1).

2.2 Matériel et méthodes

2.2.1 Cluster de calcul et gestionnaire de ressources

L'ensemble des calculs a été fait sur un cluster du laboratoire composé d'un nœud maître et d'un ensemble de nœuds de calcul, d'un serveur de base de données, d'un serveur pour les connexions utilisateur et d'un serveur hébergeant le serveur web. Le serveur web et le serveur de base de données sont doublés et déclinés en deux groupes identiques mais distincts : le premier pour la production et le deuxième pour le développement. Ceci permet le développement aisé sans impact sur la production laissée aux utilisateurs de part le monde pour les calculs nécessaires à leurs productions scientifiques.

Un gestionnaire de ressource est installé sur ce cluster : PBS Pro. Chaque calcul lui est soumis sous la forme d'une tâche (*job*) et il gère l'allocation des ressources au *job*, son exécution et le nettoyage après le *job*. Celui-ci assure l'isolation des tâches et gère la priorité des différents *jobs* en fonction de leur type.

Le système d'exploitation installé sur le cluster est Linux, plus précisément Red Hat Linux. Ce choix a été effectué lors de l'installation en faisant le compromis entre stabilité, support matériel et maintenance.

Pour le gestionnaire de ressource, un *job* est un ensemble de commandes sous la forme d'un script shell. Celui-ci nous permet aussi de spécifier au gestionnaire de ressources les ressources en temps, logiciel et disque nécessaire au *job*. Notre utilisation se basant essentiellement sur la librairie ISA (développée plus loin), c'est en partie elle qui crée les scripts en fonction des besoins.

2.2.2 PostgreSQL

PostgreSQL est un système de gestion de base de données relationnelles (SGBDR). C'est le logiciel open source qui s'approche le plus des 12 règles énoncées par Edgar Franck Codd pour définir le modèle relationnel. Il permet de stocker et d'accéder aux données de manière uniforme grâce au langage SQL (Structured Query Langage). Ce langage informatique est normalisé et il sert à effectuer des opérations sur une base de données relationnelle.

Pour gérer la base de données relationnelle HBVdb, notre choix s'est porté sur le logiciel PostgreSQL. Nous avons écarté deux autres logiciels de type SGBDR que sont MySQL et Oracle. Le premier pour la qualité de son support du modèle relationnel. En effet, le moteur de base ne supporte ni les transactions ni les clefs étrangères, ce qui est bloquant dans notre cas. Et le deuxième pour son coût de possession (achat, complexité d'installation et maintenance) qui se chiffre très rapidement en millier d'euros ainsi qu'un investissement humain important. PostgreSQL est un compromis entre performance, possibilités de configuration, facilité d'installation et de maintenance ainsi que respect des normes.

2.2.3 Schémas relationnels des bases de données

La base de données *metadb* est la base nous permettant de répertorier nos bases de données ainsi que toutes les informations les concernant (Annexe 2). Nos bases qui sont au format EMBL, comme HBVdb, ont le même schéma relationnel. Celui-ci est présenté dans la Figure 23.

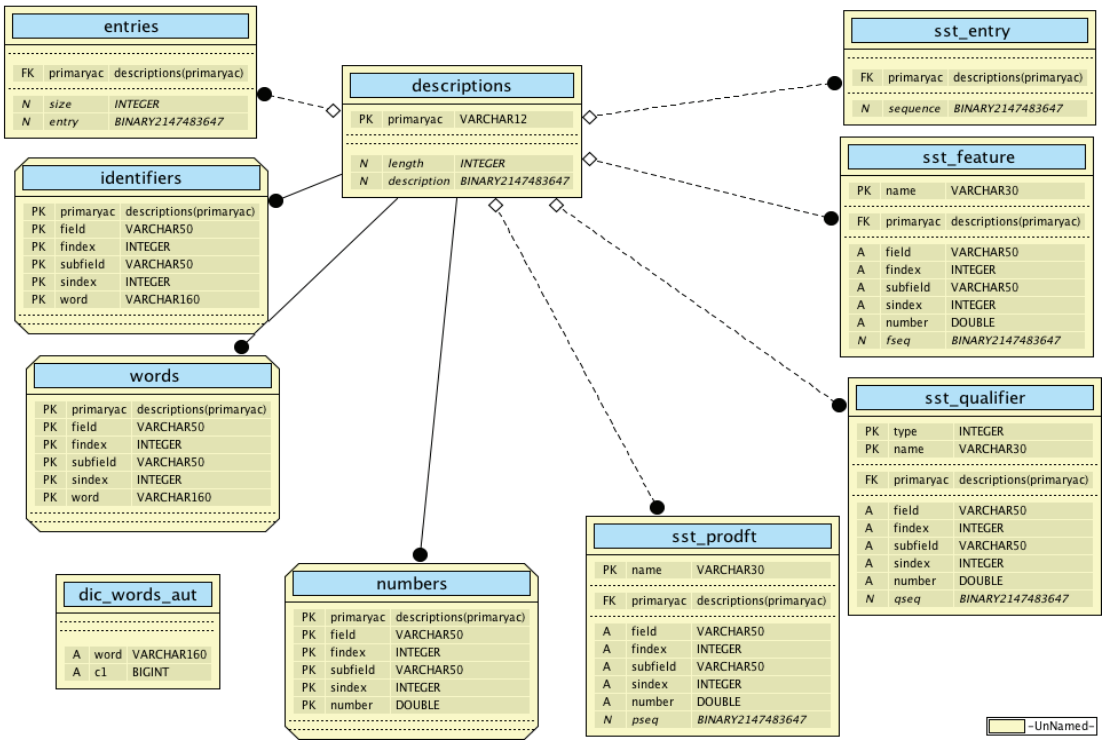


Figure 23 : Schéma relationnel d'une base de données au format EMBL.
La table *descriptions* est la table principale qui contient les numéros d'accession (*primaryac*) de toutes les entrées de la base ainsi que la séquence correspondante et sa longueur. Les autres tables contiennent toutes les autres informations relatives aux entrées et sont reliées à descriptions par la clé *primaryac*.

2.2.4 Java

Tous les programmes développés et utilisés dans les processus de gestion des bases et d'annotation sont développés avec le langage JAVA (version 1.6). Java est un langage orienté objet et il est multiplateforme. Un code développé pour Windows fonctionne sans reprogrammation sous Linux, MacOS ainsi que iOS et Android. Il a été conçu par Sun Microsystems et fonctionne en mode interprété *just in time*. Celui ci n'est pas compilé en avance mais au moment de l'exécution.

A cause de ce mode de fonctionnement, certains prêtent une lenteur à JAVA. Ceci est vrai en comparaison des langages bas niveau que peuvent être le C ou le fortran. Ce désavantage est largement compensé par un apprentissage aisé et une utilisation facilitée par son approche objet et l'ensemble des modules préexistants. Un module est un ensemble fonctionnel regroupé au sein d'une classe implémentant les méthodes nécessaires pour interagir avec les objets de la classe.

Par exemple, une classe pour travailler sur l'objet géométrique carré. L'instanciation de la classe sera l'objet carré, avec au moins 2 attributs : la longueur des côtés et les coordonnées du centre. Quatre méthodes au moins seront définies : en lecture, deux méthodes pour obtenir la longueur d'un côté et les coordonnées du centre ; et en écriture deux autres méthodes définissant la longueur d'un côté et les coordonnées du centre. D'autres méthodes seront nécessaires pour créer et détruire l'objet carré. Pour élargir, autant de méthodes que nécessaire peuvent être ajoutées, tant qu'elles sont en rapport avec l'objet.

Il existe une classe pour chaque besoin, par exemple toutes les interactions avec les SGBDR sont faites à travers les classes java.sql et javax.sql. Une autre classe permet d'envoyer des mails, une autre d'interagir avec le système de fichiers, d'autres sont encore dédiées à la communication avec le serveur web, *etc*.

2.2.5 HTML

Toutes les interfaces utilisateurs sont de type web et utilisent le langage HTML pour la mise en forme et l'accessibilité aisée depuis un simple navigateur web. Les pages sont générées par le serveur Tomcat à partir du code écrit avec l'utilisation de la librairie ISA développée ci-dessous.

2.2.6 Servlets Java

L'utilisation du langage Java, nous fait passer par un mécanisme de servlet pour produire les pages web. Servlet est le nom donné à une classe Java produisant du contenu au sein d'un serveur web de manière dynamique. Le fonctionnement est un peu plus complexe qu'un serveur web classique : la requête est reçue par le serveur HTTP (Apache HTTPd 2) qui la transmet au conteneur de servlets (Tomcat). Celui-ci la traite et renvoie le résultat au serveur web qui est ensuite transmis au client léger (navigateur web) de l'utilisateur.

2.2.7 La librairie ISA

Tous les développements en Java sont basés sur une bibliothèque logicielle développée au laboratoire, ISA (Integrated Sequence Analysis). Cette librairie, initialement créée par Christophe Combet, est développée de manière collaborative au sein du laboratoire, grâce à un système de gestion des versions : subversion (SVN). Le code source de la librairie est donc partagé par les membres l'utilisant et la développant, et de nouvelles classes ou fonctionnalités (méthodes) peuvent y être ajoutées puis partagées via le SVN. Ainsi, les autres membres peuvent mettre à jour la librairie à partir du SVN et disposer des fonctionnalités nouvellement développées.

La librairie ISA est une librairie très complète qui comporte 6 types de packages majeurs :

- ISA A : toutes les classes qui permettent de manipuler des objets biologiques allant de l'acide aminé à des listes d'alignements multiples de séquences, en passant par des classes intermédiaires de séquences, listes de séquences, *etc*.
- ISA B : ces classes contiennent tous les objets et outils utilisés pour les bases de données dans leur format « fichier plat ». Ce sont toutes les classes définissant, par exemple, les objets d'une fiche ENA (EMBL) ou UniProt.
- ISA G : regroupe les classes graphiques, notamment utilisées pour l'affichage des alignements.
- ISA I : regroupe toutes les classes permettant de générer des pages HTML et les CSS (feuilles de style, Cascading Style Sheets) associées.

- ISA R : contient toutes les classes qui permettent de gérer les bases de données relationnelles. Ce sont les classes qui permettent de communiquer avec PostgreSQL.
- ISA U : ce sont les classes utilitaires, celles qui permettent de manipuler des fichiers, de communiquer avec le système d'exploitation et le système de distribution des calculs (PBS).

La librairie ISA contient donc des packages « source » organisés en 3 types de classes (objets, outils, constantes) :

- les classes objets : elles définissent l'objet avec tous ses attributs, ainsi que les méthodes « get » et « set » qui permettent de récupérer ou de remplir les attributs de l'objet.
- Les classes « tools » (outils) : elles définissent toutes les méthodes complexes permettant de manipuler les objets auxquels elles se rapportent.
- Les classes constantes : elles permettent de définir toutes les constantes utilisées dans les classes sources de ISA A, B, I, R, U, ou dans les applications.

Sur la base de la librairie ISA, des applications ont été développées : des outils génériques (apps), des outils pour les bases de données spécialisées (db) et applications web (web). Les packages « apps » contiennent des programmes permettant par exemple de traiter des alignements (les convertir, les éditer, *etc*.), d'autres permettant de gérer des bases de données (création, chargement d'une base relationnelle à partir d'un fichier plat, formatage, *etc*.). Les packages « db » permettent de gérer les bases de données spécialisées. Les packages « web » contiennent tous les servlets nécessaires au déploiement d'un service web, par exemple HBVdb.

2.3 Résultats

2.3.1 Processus de génération de HBVdb

La génération de la base HBVdb passe par plusieurs étapes, dont la génération de plusieurs bases de références. L'ensemble du processus régissant ces étapes est exécuté par l'application *ManageHBVdb*.

2.3.1.1 La base hbvdbembl

La toute première étape est la récupération des entrées correspondant au virus de l'hépatite B dans la section STD VRL (standard, viral) de la base de données relationnelle EMBL. Cette étape fait appel à l'application *QueryDatabase*, basée sur la librairie ISA, à laquelle on passe en argument un fichier contenant la requête SQL à effectuer et la liste des sections de la base de données à interroger. L'application exécute alors la requête et les résultats, qui correspondent aux entrées, sont écrits au format texte EMBL dans un fichier de sortie portant l'extension « .dat ».

Ce fichier est alors utilisé pour générer une base de données relationnelle de ces entrées, appelée : *hbvdbembl*. Cette base est générée grâce à l'application *ManageDatabase* qui crée la base, sur le même modèle relationnel que la base EMBL, et qui charge ensuite les données du fichier plat (.dat) dans les tables de la base nouvellement créée.

2.3.1.2 Les bases maîtresses

L'entrée portant le numéro d'accession X02763 correspond à un génome complet de génotype A, de longueur 3221 nt. J'ai choisi cette entrée comme référence maîtresse (« master ») pour avoir un génome de référence le plus long possible, sans choisir un génotype G (3248 nt mais trop particulier), et dont le génotype est non recombinant et sûr. J'ai réorganisé sa séquence pour qu'elle soit dans la numérotation standard EcoR1, et je l'ai manuellement annotée.

Cette entrée manuellement annotée, au format ENA, correspond à l'entrée master. Cette entrée est utilisée par l'application *ManageDatabase* pour créer la base de données *hbvdbmastm* (master manuelle).

Une fois cette base créée, elle va servir de base de référence pour créer la base *hbvdbmast* (master non manuelle) qui correspond à la base master, annotée par le processus automatique d'annotation. Pour créer cette base, le processus va interroger la base *hbvdbembl* pour récupérer l'entrée X02763 (application *QueryDatabase*). Une fois récupérée, le fichier texte de l'entrée est donné au processus automatique d'annotation : *AnnotateHBV* (décrit plus loin dans la partie 2.3.2), et la base *hbvdbmastm* est spécifiée comme base de référence pour l'annotation. L'entrée nouvellement annotée est utilisée par *ManageDatabase* pour construire la base *hbvdbmast*.

2.3.1.3 La base de référence hbvdbref

J'ai choisi 16 génomes de référence (2 par génotype), qui sont présentés dans l'Annexe 3. Ces 16 génomes ont été sélectionnés selon plusieurs critères. Ils devaient représenter un génome complet du génotype à représenter, avoir les protéines principales complètes, et être des séquences de génotypes non recombinants. Afin de vérifier que ces génomes sont de génotypes non recombinants, nous avons combiné plusieurs méthodes de génotypage. De plus, il fallait 2 représentants pour chacun des génotypes, et ces deux représentants devaient être le plus divergent possible.

De la même manière que pour la base master non manuelle, une requête est faite sur *hbvdbembl* pour récupérer les 16 entrées de référence. Une fois récupérées, elles sont données au processus *AnnotateHBV*, la base de référence spécifiée étant *hbvdbmast*.

Une fois les 16 entrées de référence annotées, leurs annotations sont vérifiées et comparées (*AnnotateHBV*) à des objets d'annotation dont les valeurs ont été remplies manuellement pour chaque génome de référence. Si les annotations ajoutées automatiquement sont correctes en comparaison des manuelles, elles sont écrites dans le fichier de sortie .dat. Si pour une entrée il y a un conflit entre les annotations manuelles et automatiques, l'entrée est écartée et n'est pas écrite dans le fichier .dat. Enfin, le fichier .dat contenant les entrées est passé à l'application *ManageDatabase* qui construit alors la base de données *hbvdbref*.

2.3.1.4 La base hbvdb

Pour créer la base *hbvdb*, toutes les entrées de *hbvdbembl* sont récupérées grâce à l'application *QueryDatabase*. Elles sont ensuite annotées par le processus *AnnotateHBV* utilisant *hbvdbref* comme base de référence, et enfin, les entrées nouvellement annotées sont chargées dans la base ***hbvdb*** par *ManageDatabase* (Figure 24).

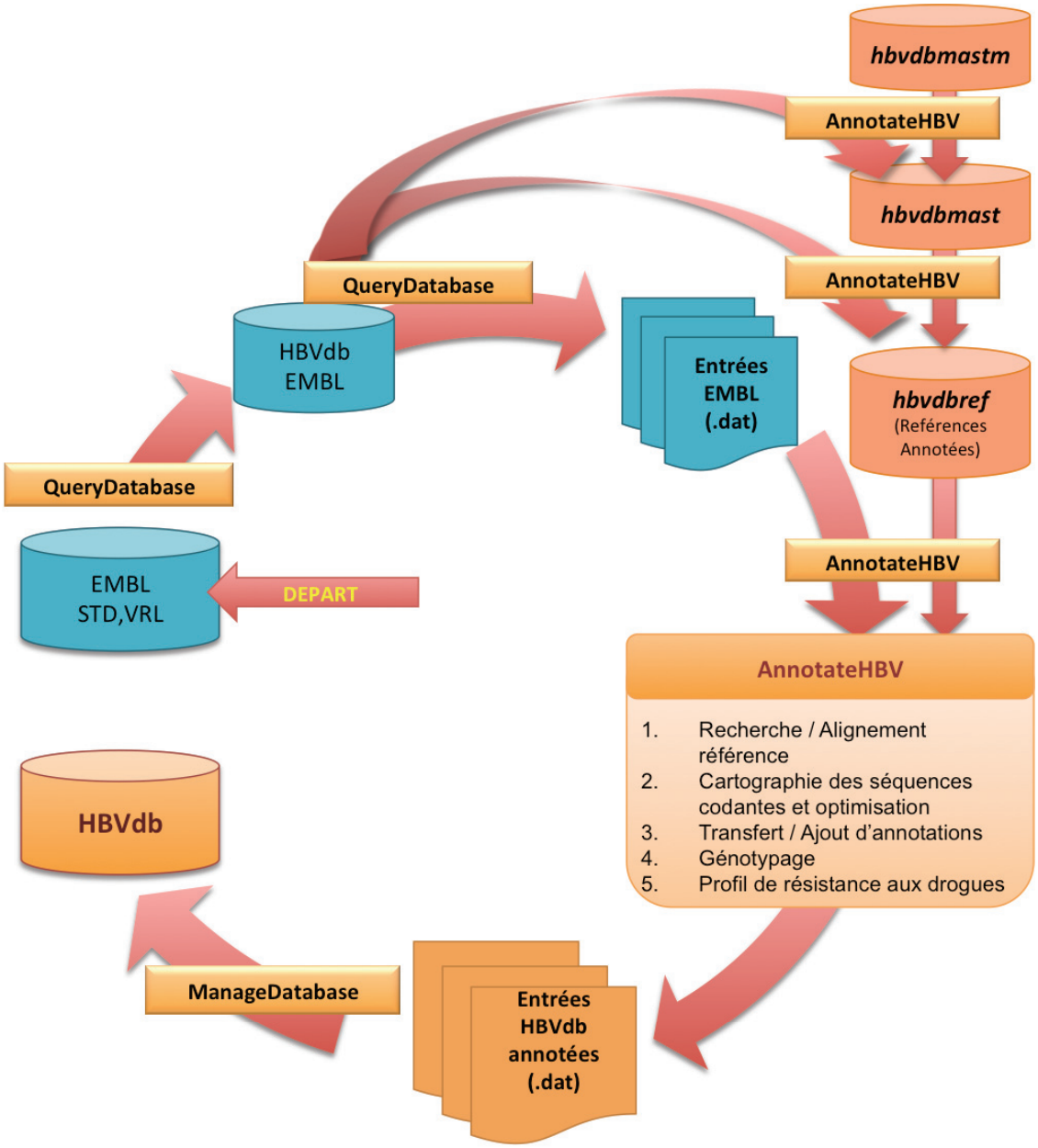


Figure 24 : Processus de génération de la base de données HBVdb.

2.3.1.5 Parallélisation

Pour diminuer le temps de génération de la base, les fichiers texte contenant les entrées (.dat) sont divisés en plusieurs fichiers (24) et l'application *AnnotateHBV* est lancée par PBS, en utilisant un « Job Array » qui permet de lancer un *job* par fichier, soit 24 jobs simultanés distribués sur 24 cœurs.

2.3.1.6 Contrôle des étapes

A chaque étape du processus, des contrôles ont lieu. Le processus s'arrête si :

- *hbvdbembl* ne contient pas les 16 entrées de référence
- *hbvdbmast* n'est pas créée (à cause d'un problème lors de l'annotation par exemple)
- *hbvdbref* ne contient pas les 16 entrées de référence
- *hbvdb* ne contient pas les 16 entrées de référence et ne contient pas au minimum 90% des entrées de *hbvdbembl.*

Pour les bases *hbvdbmastm*, *hbvdbmast*, et *hbvdbref*, une fois les contrôles passés, les séquences nucléotidiques sources de toutes les entrées de chaque base sont récupérées (*QueryDatabase*), elles sont dupliquées et écrites dans un fichier au format fasta (un fichier par base). Ces fichiers de séquences sont utilisés comme bases de séquences de référence pour le processus d'annotation qui suit.

2.3.1.7 Mise à disposition de la base

Si toutes les étapes se sont passées correctement, des requêtes sont faites sur les différentes bases, notamment par l'application *GenerateHBVDataset* qui génère tous les jeux de données pré-calculés disponibles sur le site web d'HBVdb (voir paragraphe 2.3.5.2).

Les fichiers générés sont placés dans le répertoire adéquat pour être accessibles via le site web.

Tous les fichiers de données générés sont classés dans une architecture de répertoires, puis une archive (.tar.gz) du répertoire parent est créée. Ce processus permet l'archivage de toutes les versions de la base HBVdb.

Toutes les bases construites depuis le début du processus sont en réalité nommées avec le suffixe *_r* (pour run). Si tout le processus s'est déroulé correctement, s'il existe des bases portant un suffixe *_old (e.g. hbvdbref_old*), elles sont supprimées. Puis, toutes les bases ayant le nom de production (*e.g. hbvdbref*) sont renommées avec le suffixe *_old* et enfin, les bases nouvellement créées portant le suffixe *_r* (*e.g. hbvdb*_r) sont renommées avec leurs noms de production (sans le *_r*). Cette étape constitue la mise à jour des données dans la base *metadb* et permet la mise en production de la base HBVdb.

Le processus *ManageHBVdb* est lancé une fois par mois de manière automatique, afin de créer la release la plus à jour possible en fonction des données de l'ENA/EMBL-Bank, qui sont mises à jour seulement quelques jours avant.

2.3.2 Processus d'annotation automatique

2.3.2.1 Problèmes liés à la circularité du génome du VHB

Un système de numérotation standard des génomes du VHB existe. Il est défini par le site (parfois hypothétique) de restriction EcoR1 comme origine du génome (Ono et al., 1983). Cependant, la circularité du génome du VHB conduit au dépôt de séquences, dans les bases de données publiques génériques, qui ne suivent pas ce système. Les séquences ne sont donc pas forcément toutes « découpées » ou ordonnées de la même façon.

Ces séquences nucléiques, qui ne sont par ordonnées selon un standard, vont poser des problèmes d'alignements, principalement lors de la recherche de séquences nucléotidiques similaires. En effet, si on considère une séquence codant la polymérase et que l'on veut faire une recherche de similarité avec cette séquence dans la base des génomes de référence, organisés de manière standard, la recherche va aboutir à un ou plusieurs alignements partiels. La séquence ne sera jamais alignée de manière complète avec un génome de la base.

Afin de contourner ce problème, nous avons modélisé la circularité du génome du VHB en dupliquant les séquences de chaque génome de référence (de ~ 3,2 kb à ~ 6,4kb). Ainsi, les séquences organisées de manière non standard pourront s'aligner dans leur totalité aux génomes de référence, chevauchant le génome initial et sa duplication (Figure 25).

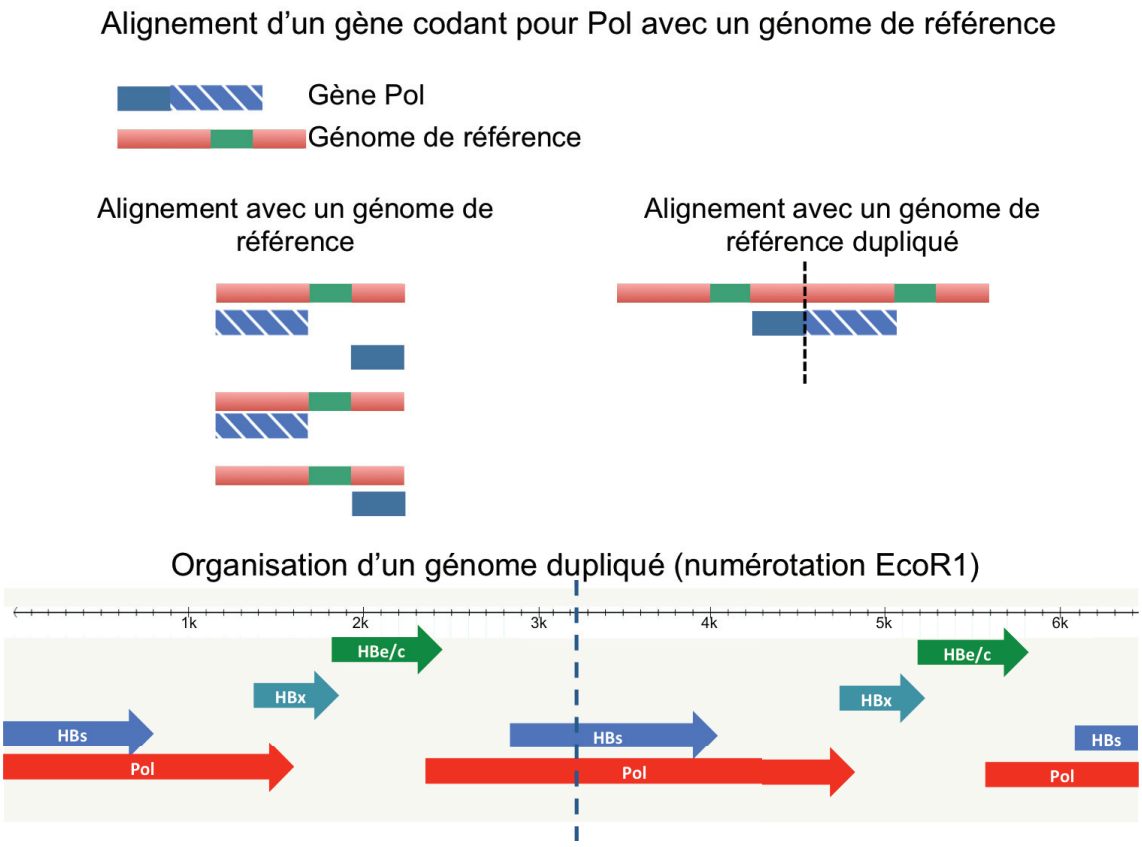


Figure 25 : Problèmes liés à la circularité du génome et organisation d'un génome dupliqué. Les différents alignements possibles d'un gène codant la Pol avec un génome de référence sont représentés. Aucun ne permet d'aligner le gène sur toute sa longueur. La solution de duplication des génomes de référence est présentée en bas du schéma.

2.3.2.2 Recherche du génome de référence le plus proche

L'entrée à annoter sera appelée entrée requête. La toute première étape de l'annotation consiste à cloner l'objet stockant les informations de l'entrée ENA/EMBL-Bank (*EMBLEntry*) requête, et à effacer les attributs du clone pour obtenir un objet HBVdb vierge (*HBVdbEntry*). Les attributs enlevés sont : tous les *features*, les lignes de commentaires, la ligne de description (DE), la ligne de mots clés (KW), la ligne d'en-tête de la séquence (SQ) et la séquence. Tous ces éléments seront mis à jour ou ré-annotés par le processus *AnnotateHBV*. La nouvelle entrée conserve le numéro d'accession initial de l'ENA/EMBL-Bank. L'objet *EMBLEntry* initial sera utilisé lors du processus de transfert d'annotations vers l'objet *HBVdbEntry*.

L'étape suivante récupère la séquence source depuis l'objet *EMBLEntry*, par la suite appelée séquence requête. Dans l'objet *EMBLEntry*, la « *location* » (positions / localisation) de cette séquence se trouve dans le *feature* (ligne FT) nommé *source*. Cette *location* doit couvrir toute la séquence contenue dans la section SQ. Si l'objet *EMBLEntry*

contient plusieurs *feature source* ou si la séquence récupérée est trop courte (inférieure à 27 nt) ou plus longue que les plus longs génomes du VHB (supérieure à 3500 nt), l'entrée est écartée sans être annotée.

Sinon, une recherche de similarité est effectuée, par le programme FASTA, avec la séquence requête contre la base des séquences des génomes de référence dupliquées.

Si des résultats significatifs sont trouvés, avec une valeur E inférieure à 10^{-6}, le premier alignement par paire est récupéré. Il correspond à l'alignement de la séquence requête avec le génome de référence dupliqué le plus similaire. Si aucun résultat n'est trouvé, une nouvelle recherche est faite en utilisant la séquence inversée complémentée de la séquence requête. En cas de nouvel échec de la recherche de références similaires, l'entrée est écartée sans être annotée.

2.3.2.3 Réarrangement de la séquence requête

Une fois le premier alignement par paire récupéré, l'entrée correspondant au génome de référence le plus proche est également récupérée et stockée dans un objet. L'alignement est vérifié et si la séquence requête ou reverse complémentée chevauche le génome initial de référence et sa duplication, l'alignement va être réarrangé par l'algorithme de « *cut&paste* » (coupé-collé).

L'algorithme de *cut&paste* va couper toute la partie de la séquence requête qui est alignée sur la duplication, et va la translater au niveau de la région de la référence qui est identique et qui se trouve à l'origine du génome initial. Dans ce cas de figure, si après le *cut&paste* il y a une zone de gaps entre la partie 3' et la partie 5' de la séquence requête (séquence plus courte qu'un génome complet), ces gaps sont remplacés par une série de « n » de nombre égal au nombre de gaps (Figure 26).

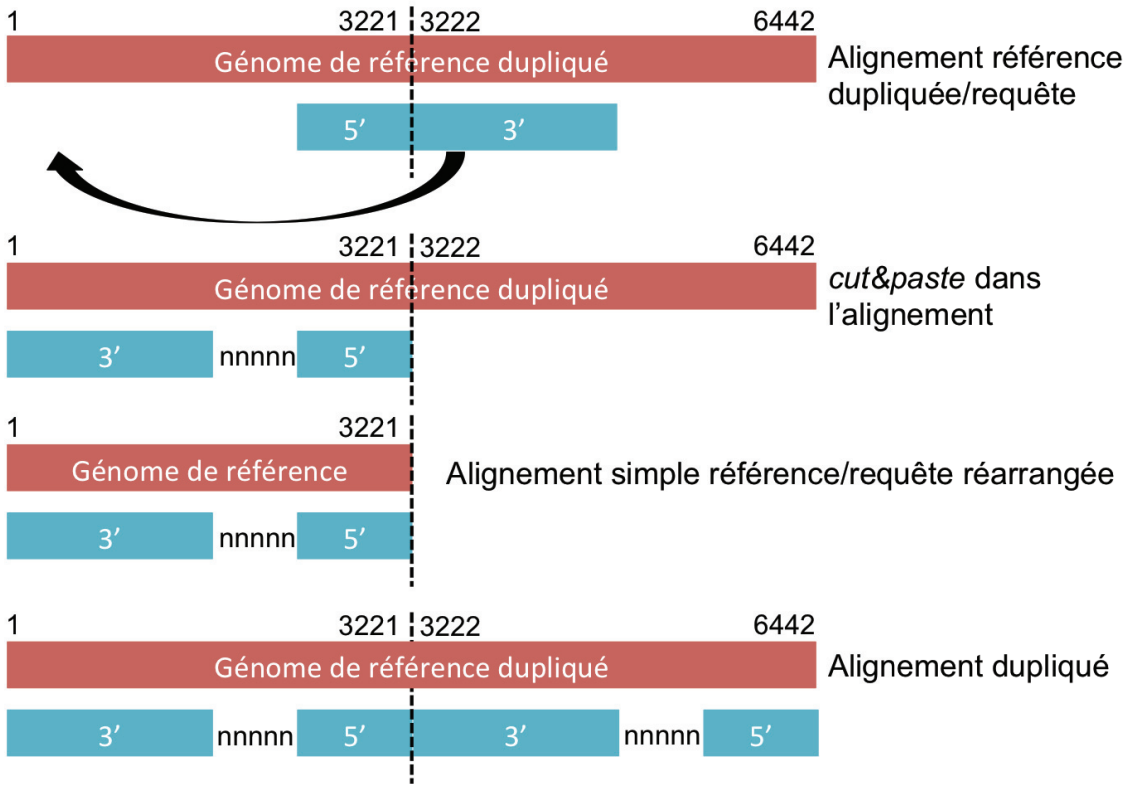


Figure 26 : Fonctionnement de l'algorithme *cut&paste*

Qu'il y ait eu *cut&paste* ou pas, l'alignement est coupé en deux, puisque à ce stade, la séquence requête n'est plus alignée sur la duplication de la séquence de référence. Le résultat est un alignement simple, qui va être dupliqué pour modéliser la circularité des deux séquences, la requête et la référence.

La séquence requête, éventuellement réarrangée par le *cut&paste* (avec suite de « n » ou pas), est extraite de l'alignement simple et stockée dans l'objet cloné (*HBVdbEntry*), comme séquence source (sans les gaps).

2.3.2.4 Annotation du feature source

Comme précisé dans la partie 1.4.1, les *features* sont composés d'un nom, d'une *location* (localisation / positions), et d'une liste de *qualifiers*. Pour le *feature source*, la *location* correspond aux positions qui délimitent la séquence de l'entrée requête (de 1 à la longueur de la séquence). La liste de *qualifiers* ajoutés au *feature source* contient deux types de *qualifiers*. Le premier type existe dans le format EMBL, et est nommé *db_xref*. Ces *qualifiers* sont des liens vers des bases de données et leurs valeurs correspondent aux identifiants dans ces bases. Le processus *AnnotateHBV* ajoute 2 *db_xref* dans le *feature Source* :

- le lien ENA (EMBL) avec le numéro d'accession de l'entrée de départ
- le lien HBVdb avec le numéro d'accession de l'entrée de référence utilisée pour l'annotation

L'autre type de *qualifier* ajouté est un *qualifier* spécifique à HBVdb, désigné par *PRABI_genotype*. Sa valeur permet de renseigner 3 informations (séparées par « : ») concernant le génotype du VHB de la séquence requête :

- le génotype déposé dans l'entrée ENA s'il a pu être extrait
- le génotype prédit par l'algorithme de génotypage (développé dans le paragraphe 2.3.3)
- le génotype confirmé, qui n'est rempli que pour les génomes de référence dont les génotypes ont été vérifiés par plusieurs méthodes.

Quand une de ces informations n'est pas connue, sa valeur est : « n.a. ».

2.3.2.5 Cartographie des séquences codantes (CDS) et optimisation des alignements

Dans le format ENA, les *features* représentant les séquences codantes sont désignés par le nom *CDS*. Le processus d'annotation parcourt l'entrée de référence pour récupérer tous les *features CDS*. Pour chaque *CDS* de référence récupérée, les *locations* sont cartographiées sur l'alignement dupliqué. Les positions correspondantes dans la séquence requête sont repérées sur l'alignement et sont stockées pour générer, plus tard, la *location* de la *CDS*, si elle est présente dans la séquence requête.

Il faut bien noter que, le VHB ayant un génome circulaire, certaines *locations* de *CDS* sont notées avec un opérateur de jointure, comme ceci : **join(2307..3221,1..1623)**.

Cette notation indique que la séquence codante comprend la région allant des positions 2307 à 3221 en 5' et la région allant des positions 1 à 1623 en 3'. Lorsqu'il s'agit de *locations* notées ainsi, le processus d'annotation va traduire les *locations* de la partie 3' (début du génome) pour qu'elles correspondent à la région identique sur le duplicata du génome de référence, afin que les positions se suivent physiquement sur l'alignement dupliqué. Ainsi, ces *locations* peuvent être décrites sans opérateur de jointure et la *CDS* est donc délimitée par 2 positions dans la référence : **2307..4844** (4844=3221+1623).

Ces 2 positions sont utilisées pour couper l'alignement et ne récupérer que la partie correspondant à la *CDS* : c'est l'alignement de la *CDS*.

Le processus génère tous les alignements par *CDS* en utilisant les *locations* des CDS de référence. L'étape suivante consiste à vérifier et à optimiser si besoin les alignements requête / référence de chaque *CDS* pour éviter une traduction erronée ou pour détecter des séquences non-fonctionnelles. Si un alignement de *CDS* ne contient que des gaps au niveau de la séquence requête, c'est que la séquence source ne contient pas cette *CDS*. L'alignement de cette *CDS* n'est donc pas conservé et elle ne sera pas annotée dans la nouvelle entrée.

Si, au contraire, la séquence requête est présente dans l'alignement de la *CDS* en cours de traitement, l'alignement est divisé en fenêtres non-chevauchantes de 3 nucléotides (codons). Si une fenêtre contient un ou deux gaps, le processus tente d'optimiser les gaps afin d'avoir seulement 3 nucléotides ou trois gaps (délétion d'un codon) dans la fenêtre. Dans certains cas, l'optimisation n'est pas possible car le nombre de gaps présents dans la zone à optimiser n'est pas divisible par 3.

Si l'optimisation échoue, le processus d'annotation s'arrête pour cette entrée, et elle ne sera pas intégrée dans la base.

2.3.2.6 Transfert d'annotations

2.3.2.6.1 Les séquences codantes

Pour chaque *CDS* correcte et présente dans la séquence stockée dans l'objet *HBVdbEntry*, le *feature* correspondant est créé avec la *location* stockée au départ. Certains *qualifiers* (*PRABI_name*, *locus_tag* ; voir tableau Annexe 4) sont récupérés dans la liste de *qualifiers* de la *CDS* de la référence. Les *qualifiers* nommés *translation*, *dbxref* et *codon_start* sont créés indépendamment de la référence.

Pour remplir le qualifier *translation*, une traduction de la séquence requête est effectuée à partir de l'alignement de la *CDS*, éventuellement optimisé. Si la traduction contient un codon stop avant le dernier codon, le *qualifier note* est ajouté, avec comme valeur : « non-functional ». Sinon, le *qualifier translation* est ajouté avec comme valeur la chaine de caractères correspondant à la traduction de la CDS.

La traduction est comparée avec les traductions stockées à partir de l'entrée ENA de départ. Si une traduction identique est retrouvée, les *qualifiers dbxref* sont récupérés du *feature CDS* correspondant dans l'entrée ENA de départ, et ils sont ajoutés à la liste des *qualifiers* de la *CDS* en cours d'annotation de la requête. Les *qualifiers dbxref* étant des

liens vers d'autres bases de données, ils ont ici pour valeurs les identifiants de la protéine traduite dans des bases telles que UniProtKB.

La valeur du *qualifier codon_start* est tirée de l'alignement de la CDS éventuellement optimisée. Il a une valeur valide de 1 ou 2 ou 3, ce qui indique, le nucléotide à partir duquel on trouve le premier codon complet de la CDS. Le numéro du nucléotide est donné par rapport à la première base du *feature CDS* indiquée dans la *location*. En général la valeur est 1, sauf quand le premier codon de la CDS contient 1 ou 2 gap(s) (dans la séquence requête au niveau de l'alignement de la CDS).

Dans le cas où la CDS correspond à la polymérase, la séquence protéique traduite est fournie au programme *FindHBVRMut*, qui va chercher si la séquence présente des mutations associées à la résistance aux traitements, et ainsi définir le profil de résistance de la séquence.

2.3.2.6.2 Les protéines

Lorsqu'une *CDS* est annotée et qu'elle possède une traduction, le *feature mat_peptide* correspondant est créé. Ses *locations* sont cartographiées à partir de la référence sur la séquence requête. Les *qualifiers function*, *locus_tag*, *product* et *PRABI_name* sont récupérés de la référence et ajoutés au nouveau *mat_peptide* de la requête.

Des *qualifiers* particuliers sont ajoutés, s'ils existent dans le *mat_peptide* de référence : les *PRABI_prodft*. Ce sont des annotations de la séquence de la protéine traduite qui suivent le format des *features* de la base UniProtKB. Les *qualifiers PRABI_prodft* décrivent les chaînes des protéines, les domaines protéiques, certains sites tels que les sites actifs (*act_site*) qui désignent les résidus catalytiques des enzymes. Ils sont également utilisés, dans le processus d'annotation, pour décrire les mutations associées à des résistances aux traitements, avec la drogue concernée et le statut de résistance (voir paragraphe 2.3.4).

La Figure 27 résume les différentes étapes du processus *AnnotateHBV*.

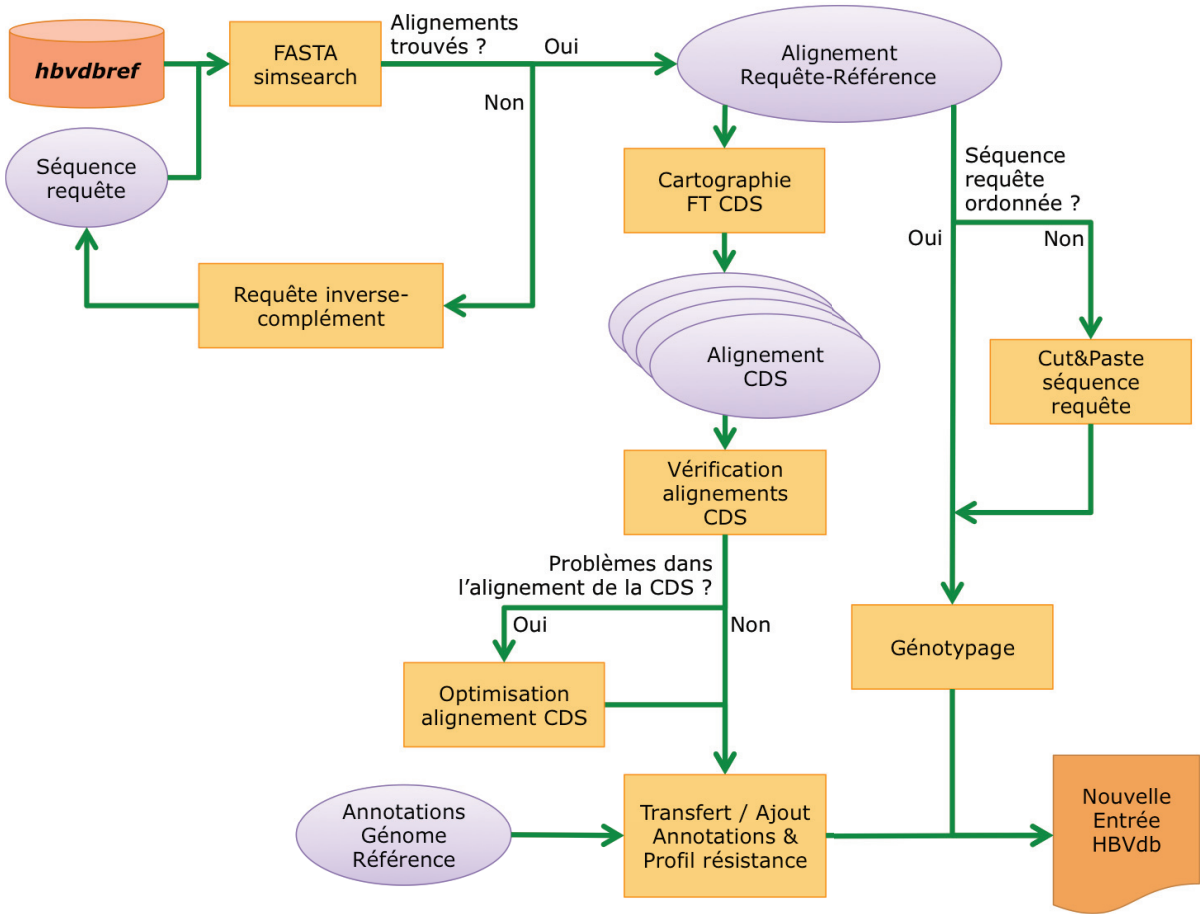


Figure 27 : Schéma descriptif du processus d'annotation automatique.

La Figure 28 correspond à une partie de l'entrée annotée HBVdb X02763 (« master »), qui est présentée dans sa totalité dans l'Annexe 5. Cette partie est l'annotation des *features CDS* et *mat_peptide* pour la polymérase.

```
FT   CDS             join(2307..3221,1..1623)
FT                   /PRABI_name="P"
FT                   /locus_tag="HBVORF13"
FT                   /codon_start="1"
FT                   /db_xref="UniProtKB/Swiss-Prot:P03159"
FT                   /translation="MPLSYQHFRKLLLLDDGTEAGPLEEELPRLADADLHRRVAEDLNL
FT                   GNLNVSIPWTHKVGNFTGLYSSTVPIFNPEWQTPSFPKIHLQEDIINRCQQFVGPLTVN
FT                   EKRRLKLIMPARFYPTHTKYLPLDKGIKPYYPDQVVNHYFQTRHYLHTLWKAGILYKRE
FT                   TTRSASFCGSPYSWEQELQHGRLVIKTSQRHGDESFCSQSSGILSRSSVGPCIRSQLKQ
FT                   SRLGLQPRQGRLASSQPSRSGSIRAKAHPSTRRYFGVEPSGSGHIDHSVNNSSSCLHQS
FT                   AVRKAAYSHLSTSKRQSSSGHAVEFHCLPPNSAGSQSQGSVSSCWWLQFRNSKPCSEYC
FT                   LSHLVNLREDWGPCDEHGEHHIRIPRTPARVTGGVFLVDKNPHNTAESRLVVDFSQFSR
FT                   GISRVSWPKFAVPNLQSLTNLLSSNLSWLSLDVSAAFYHIPLHPAAMPHLLIGSSGLSR
FT                   YVARLSSNSRINNNQYGTMQNLHDSCSRQLYVSLMLLYKTYGWKLHLYSHPIVLGFRKI
FT                   PMGVGLSPFLLAQFTSAICSVVRRAFPHCLAFSYMDDVVLGAKSVQHRESLYTAVTNFL
FT                   LSLGIHLNPNKTKRWGYSLNFMGYIIGSWGTLPQDHIVQKIKHCFRKLPVNRPIDWKVC
FT                   QRIVGLLGFAAPFTQCGYPALMPLYACIQAKQAFTFSPTYKAFLSKQYMNLYPVARQRP
FT                   GLCQVFADATPTGWGLAIGHQRMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSV
FT                   VLSRKYTSFPWLLGCTANWILRGTSFVYVPSALNPADDPSRGRLGLSRPLLRLPFQPTT
FT                   GRTSLYAVSPSVPSHLPVRVHFASPLHVAWRPP"
FT   mat_peptide     join(2307..3221,1..1623)
FT                   /function="DNA-polymerase/Reverse Transcriptase coding
```

```
FT                   sequence"
FT                   /locus_tag="HBVORF13"
FT                   /product="Polymerase/Reverse Transcriptase"
FT                   /PRABI_name="Pol"
FT                   /PRABI_prodft=(pos:1..845, chain, "Polymerase/Reverse
FT                   transcriptase")
FT                   /PRABI_prodft=(pos:1..183, domain, "Terminal Protein (TP)/
FT                   Primase domain")
FT                   /PRABI_prodft=(pos:1..183, domain, "Terminal Protein (TP)/
FT                   Primase domain")
FT                   /PRABI_prodft=(pos:184..348, domain, "Spacer")
FT                   /PRABI_prodft=(pos:349..692, domain, "Reverse Transcriptase
FT                   (RT) domain")
FT                   /PRABI_prodft=(pos:431..431, act_site, "RT catalytic Asp")
FT                   /PRABI_prodft=(pos:553..553, act_site, "RT catalytic Asp")
FT                   /PRABI_prodft=(pos:554..554, act_site, "RT catalytic Asp")
FT                   /PRABI_prodft=(pos:693..845, domain, "Ribonuclease H
FT                   (RNaseH) domain")
FT                   /PRABI_prodft=(pos:702..702, act_site, "RNaseH catalytic
FT                   Asp")
FT                   /PRABI_prodft=(pos:731..731, act_site, "RNaseH catalytic
FT                   Glu")
FT                   /PRABI_prodft=(pos:750..750, act_site, "RNaseH catalytic
FT                   Asp")
```

Figure 28 : Une partie de l'entrée HBVdb annotée X02763.
Les *features CDS* et *mat_peptides* correspondant à la polymérase sont présentés.

2.3.3 Processus de génotypage

Le processus *GenotypeHBV* commence par une recherche de similarité, exécutée par le programme FASTA, dans la base de référence avec la séquence requête éventuellement réarrangée par le processus *cut&paste* durant l'annotation.

À partir de tous les alignements par paires requête/référence générés par le programme FASTA, l'algorithme calcule une matrice contenant le pourcentage d'identité moyen par génotype à chaque position entre la requête et les références, et ce, pour chaque génotype.

Les deux dimensions de la matrice d'identité sont respectivement : la longueur de la séquence requête, et le nombre de génotypes (8). Chaque alignement est parcouru par des fenêtres glissantes chevauchantes de 301 nucléotides (nt) de longueur, avec un décalage 1 nt. Le pourcentage d'identité de chaque fenêtre est ajouté dans la matrice, à la position médiane de la fenêtre, au niveau du génotype correspondant à celui du génome de référence qui est aligné avec la requête, dans l'alignement en cours de traitement.

Une autre matrice de dimensions égales est créée pour stocker le nombre de valeurs ajoutées dans chaque case de la matrice d'identité. Ceci permet de faire une moyenne des pourcentages d'identité par position pour chaque génotype.

Puis, pour chaque position, la lettre du génotype correspondant à la valeur maximale est insérée dans un tableau de la taille de la séquence. Si l'identité maximale est inférieure à 90%, un « ? » est ajouté à la place d'une lettre.

Le tableau des génotypes prédits par position est utilisé pour écrire une chaîne de caractères représentant le génotype le plus probable à chaque position de la séquence requête.

Le génotype global est alors déduit de cette chaine de caractères. *GenotypeHBV* génère un fichier de sortie qui contient :

- le génotype global prédit
- le pourcentage de positions informatives utilisées dans le calcul du génotype
- la longueur de la séquence requête
- la séquence requête
- la chaine de caractère des génotypes prédits à chaque position
- les informations sur les paramètres utilisés dans le programme.

Le fichier de sortie (présenté en Annexe 6) est lu par le programme *AnnotateHBV*, et le génotype global prédit est récupéré pour annoter le *qualifier PRABI_genotype* du *feature source*.

L'algorithme traite uniquement des séquences de longueur égale ou supérieure à la taille de la fenêtre. Il permet l'identification de séquences de génotypes «purs» ainsi que des formes recombinantes (Kramvis et al., 2008; Simmonds and Midgley, 2005).

La Figure 29 résume les étapes du processus *GenotypeHBV*.

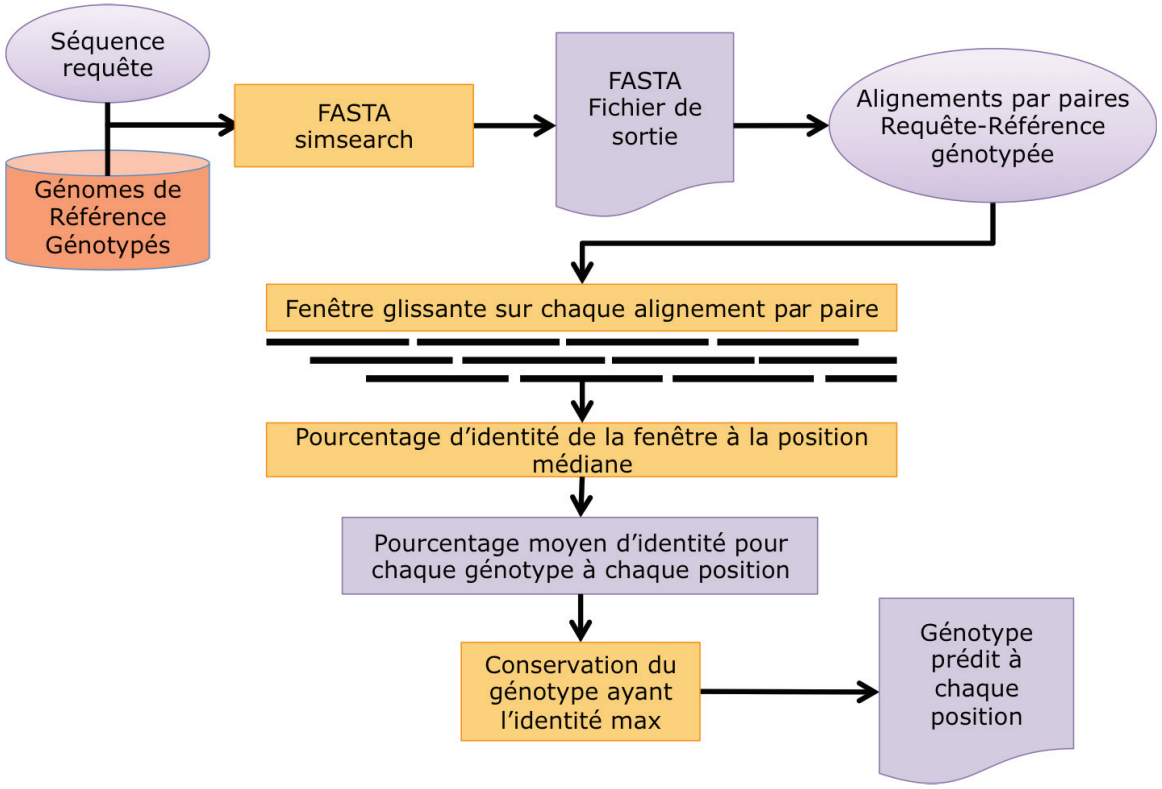


Figure 29 : Schéma descriptif du processus de génotypage.

J'ai effectué une comparaison de l'outil *GenotypeHBV* avec d'autres outils disponibles permettant de génotyper des séquences du VHB, sur un jeu de données d'une vingtaine de séquences comprenant des génotypes « purs », des recombinants (démontrés dans des publications (Simmonds and Midgley, 2005)), des génomes mosaïques, et des séquences courtes. Cette étude, dont les résultats sont présentés dans le Tableau 7, montre que l'efficacité de notre outil de génotypage est comparable ou équivalente à celle de l'outil « Oxford subtyping tool » (Alcantara et al., 2009; de Oliveira et al., 2005), de jpHMM (Schultz et al., 2012), et de l'outil « Genotyping » du NCBI (Rozanov et al., 2004). Tous ces outils permettent de détecter des recombinants et présentent les génotypes prédits sur la longueur de la séquence. L'outil de génotypage de la base de données HepSEQ (Gnaneshan et al., 2007) n'utilise que le gène de la polymérase (chevauchant le gène des protéines de surface) pour le calcul du génotype, et ne permet visiblement pas de détecter les formes recombinantes.

Accession	Longueur	HBVdb (IP*)	HepSEQ (C**)	jpHMM	NCBI	Oxford
AB064313	3248	G (100,00)	G (HC)	G	G	G
AB073853	3215	B (100,00)	B (LC)	B	B	B
AB106564	3212	E (100,00)	E (HC)	E	E	E
AB231908	3215	AGC (95,55)	G (n.a.)	AC	AGCBE	AGEBC
AB241117	3215	BC (99,78)	B (HC)	BC	B	BC
AB330367	62	n.a.	E (LC)	E	n.a.	n.a.
AB330367	3182	D (100,00)	D (HC)	D	D	D
AF233236 (fragment)	240	n.a.	D (n.a.)	C	BC	C
AF233236	3068	CB (90,19)	C (HC)	BC	CB	n.a.
AF461043	3215	DC (100,00)	D (LC)	DC	DC	DC
AY090454	3215	H (100,00)	H (HC)	H	H	H
AY090458	3215	F (100,00)	F (LC)	F	F	F
AY161141	3221	AD (99,78)	A (HC)	AD	ADE	AD
EU833890	3248	GA (100,00)	G (n.a.)	GA	GA	G
EU939678	3215	B (96,42)	B (HC)	BC	B	BC
GU456651	318	n.a.	D (HC)	D	D	D
Referee	681	D (100,00)	D (HC)	D	D	n.a.
X02496	3182	DE (100,00)	D (HC)	D	DE	D
X02763	3221	A (99,97)	A (HC)	A	A	A
X65258	318	DA (100,00)	D (LC)	DA	DA	DA
X75665	3215	C (100,00)	C (HC)	C	C	C

Tableau 7 : Comparaison des résultats des outils de génotypage du VHB.
Comparaison sur un jeu de données de 21 séquences (IP* : informative positions, C** : confidence, HC : high confidence, LC : low confidence, n.a. : not available)

2.3.4 Processus de détection de profils de résistance aux drogues

L'application *FindHBVRMut* utilise 2 paramètres importants :

- un fichier contenant une séquence de référence de transcriptase inverse (RT)
- un fichier contenant les profils de résistance connus.

Les profils de résistance ont été définis à partir des mutations de résistance répertoriées par l'EASL dans le rapport de 2012 (guidelines) (European Association For The Study Of The Liver, 2012). Ces profils définissent les résistances connues aux NA suivants : lamivudine, telbivudine, entécavir, adéfovir et ténofovir. A partir du Tableau 3 page 41 (tiré du rapport EASL 2012), toutes les combinaisons possibles de mutations ont été calculées et répertoriées dans un fichier. Dans ce fichier, chaque ligne correspond à un profil de résistance, et les colonnes représentent :

- le nom du profil (drogue concernée + numéro)
- le statut de résistance : R (résistant), I (intermédiaire) ou S (sensible)
- le nombre de mutations impliquées dans la résistance
- la liste des mutations, décrites avec le système de numérotation RT (Stuyver et al., 2001).

La première étape de l'application *FindHBVRMut* est le calcul d'un alignement entre la séquence protéique requête et une séquence de référence de transcriptase inverse (RT).

Ensuite, l'algorithme repère, dans la séquence requête, toutes les positions listées dans les différents profils. Pour chacune de ces positions, il vérifie quel est l'acide aminé présent. S'il s'agit d'un aa différent de l'aa « sauvage », il détermine si l'aa présent correspond à une mutation connue. Si ce n'est pas une mutation connue, le profil sera noté « other ». Si c'est une mutation connue, tous les profils auxquels elle participe sont stockés.

Une fois toutes les positions de mutations connues vérifiées, pour chaque profil stocké, le programme vérifie que toutes les mutations définissant le profil ont été trouvées dans la séquence. Si elles ne sont pas toutes présentes, le profil est éliminé. Si elles sont toutes retrouvées, le profil est conservé (Figure 30).

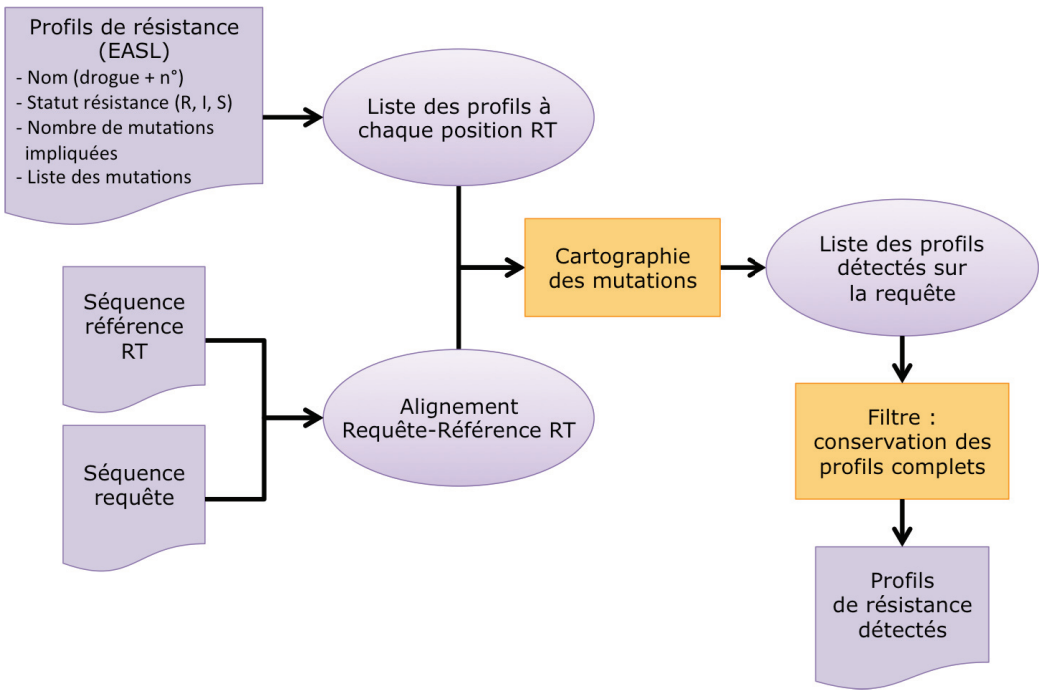


Figure 30 : Schéma descriptif du processus de détection des profils de résistance aux NA.

Enfin, *FindHBVRMut* crée un fichier de sortie (présenté en Annexe 7) avec la liste des profils détectés. Le format de cette liste est presque identique au format du fichier d'entrée qui définit les profils à chercher, mais avec une colonne en plus donnant la liste des mutations notées selon leur position dans la séquence requête. Les profils « other », s'il y en a, sont indiqués mais ne présentent pas de statut de résistance, et le nombre de mutations participant au profil est noté -1 (car il n'est pas connu), quant à la liste de mutations, elle ne contient que la mutation détectée comme n'étant pas connue.

La première ligne du fichier donne le statut de résistance déduit des profils détectés, et la séquence requête utilisée est affichée à la suite de la liste de profils. Si aucun profil de résistance n'est détecté, le fichier de sortie donne le statut « Sensitive » (sensible) et la séquence requête.

Le processus *AnnotateHBV* lit le fichier de sortie de *FindHBVRMut* pour annoter les *PRABI_prodft* « *res_mut* » dans le *feature mat_peptide* correspondant à la polymérase, si des mutations de résistance ont été détectées.

2.3.5 Interface web

La base HBVdb est accessible librement et sans enregistrement via un site web : http://hbvdb.ibcp.fr.

Son interface est divisée en plusieurs menus et sous-menus qui sont présentés sous forme de tableau sur la page d'accueil, accessible par le menu *Home*.

2.3.5.1 Menu HBV

Le menu HBV donne accès à des pages d'information et d'introduction sur le virus de l'hépatite B. Il se compose de 3 sous-menus : *Genome*, *Nomenclature*, et *Proteins*.

Le sous-menu *Genome* présente une brève introduction sur l'épidémiologie, la pathologie et les traitements. Puis, il y a un paragraphe qui détaille l'organisation du génome du VHB, avec des liens vers d'autres bases (ViralZone (Hulo et al., 2011), NCBI Taxonomy (NCBI Resource Coordinators, 2013), *etc.*). Le troisième paragraphe décrit les étapes principales du cycle de réplication.

Le sous-menu *Nomenclature* présente les différents génotypes du VHB et leur répartition géographique. Cette page donne également accès à des informations sur les génomes de référence utilisés pour construire HBVdb, sous forme d'un tableau avec pour chaque référence : le génotype, le numéro d'accession, la longueur du génome,

l'isolat et les références littéraires. Ce menu présente également la nomenclature standard utilisée pour les annotations des CDS et des protéines (*mat_peptide*), et explique la numérotation standard avec le site EcoR1 qui est utilisée dans HBVdb.

Le sous-menu *Proteins* est composé de 4 autres sous-menus :

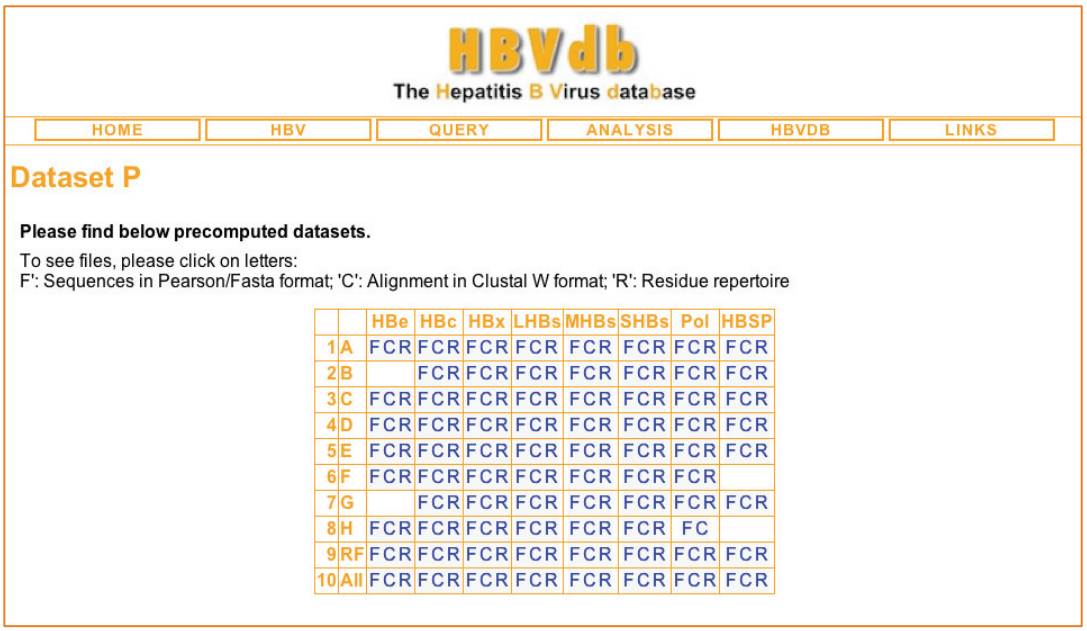
- *Core* : présente les informations concernant la protéine de capside et la protéine HBe
- *HBx* : donne des informations sur la protéine X du VHB
- *Surface* : montre les caractéristiques des 3 protéines de surface LHBs, MHBs et SHBs
- *Polymerase* : décrit les différents domaines de la polymérase du VHB, ainsi que ses inhibiteurs, les analogues de nucléos(t)ides qui sont utilisés dans le traitement contre l'hépatite B.

2.3.5.2 Menu Query

Ce menu a pour objectif principal de permettre à l'utilisateur d'accéder aux données de la base HBVdb. A l'heure actuelle, il comporte un sous-menu intitulé *Dataset*. Ce menu propose à l'utilisateur des données précalculées à partir de la base, qui ont été préparées par l'application *GenerateHBVDataset*, durant le processus de génération de la base HBVdb (*ManageHBVdb*). Deux types de jeux de données sont disponibles : des jeux de séquences nucléotidiques (sous-menu *Nucleotide*) et des jeux de séquences protéiques (sous-menu *Proteins*). Pour chacun de ces sous-menus, la page présente un tableau dont les lignes correspondent aux différents génotypes et les colonnes représentent les CDS (pour les jeux nucléotidiques) ou les protéines (pour les jeux protéiques). L'utilisateur peut donc récupérer les séquences complètes d'une protéine donnée, ou d'une CDS donnée, pour un génotype donné. Il y a 10 lignes qui correspondent aux génotypes : 8 pour les génotypes de A à H, une pour les formes recombinantes (RF), et une ligne « All » pour tous les génotypes confondus (Figure 31). Chaque case du tableau contient généralement 3 lettres qui sont des liens vers les pages présentant le jeu de données correspondant. Ces 3 liens sont :

- F : le fichier des séquences au format Fasta
- C : l'alignement multiple au format ClustalW
- R : les répertoires de résidus et leurs fréquences, calculés à partir de l'alignement multiple.

Pour chacune des pages, l'utilisateur peut télécharger les fichiers de données pour faire des analyses supplémentaires, à partir du menu *Analysis* par exemple.



		HBe	HBc	HBx	LHBs	MHBs	SHBs	Pol	HBSP
1	A	FCR	FCR	FCR	FCR	FCR	FCR	FCR	FCR
2	B		FCR	FCR	FCR	FCR	FCR	FCR	FCR
3	C	FCR	FCR	FCR	FCR	FCR	FCR	FCR	FCR
4	D	FCR	FCR	FCR	FCR	FCR	FCR	FCR	FCR
5	E	FCR	FCR	FCR	FCR	FCR	FCR	FCR	FCR
6	F	FCR	FCR	FCR	FCR	FCR	FCR	FCR	
7	G		FCR	FCR	FCR	FCR	FCR	FCR	FCR
8	H	FCR	FCR	FCR	FCR	FCR	FCR	FC	
9	RF	FCR	FCR	FCR	FCR	FCR	FCR	FCR	FCR
10	All	FCR	FCR	FCR	FCR	FCR	FCR	FCR	FCR

Figure 31 : Page web HBVdb des jeux de données protéiques.
(Menu *Query* => *Dataset* => *Protein*)

2.3.5.3 Menu Analysis

Ce menu donne accès à des outils d'analyses de séquences génériques et spécialisés. Le menu *Analysis* est donc divisé en 3 sous-menus :

- *Generic N* : donne accès à des outils d'analyses de séquences nucléotidiques tels que BlastN, FASTA pour séquences nucléotidiques, et Clustal W.
- *Generic P* : donne accès aux mêmes outils, mais pour des séquences protéiques.
- *Specialized* : donne accès à 3 outils spécialisés pour l'analyse des séquences du VHB : *Annotate*, *Genotype*, *Resistance*.

2.3.5.3.1 Les outils génériques

Les outils génériques sont en fait regroupés sur le serveur d'analyse NPS@ (Combet et al., 2000). Pour les outils génériques de recherche de séquences similaires, il est possible de choisir comme banques de séquences à interroger celles qui contiennent les séquences extraites depuis HBVdb. En effet, l'application *GenerateHBVDataset*, exécutée durant le processus *ManageHBVdb*, génère un fichier de toutes les séquences nucléotidiques et un fichier de toutes les séquences protéiques, extraites de HBVdb. Ces fichiers sont ensuite formatés pour pouvoir être utilisés comme base de données par les

programmes BLAST et FASTA. Les interfaces web de ces programmes, lorsqu'ils sont utilisés avec les banques de séquences de HBVdb, donnent accès, via des liens, aux entrées annotées HBVdb correspondant aux séquences détectées comme similaires.

2.3.5.3.2 Les outils spécialisés pour l'analyse des séquences du VHB

Ils permettent à l'utilisateur d'annoter, de génotyper, ou de détecter des profils de résistance aux drogues dans ses propres séquences.

Ces 3 menus vont conduire l'utilisateur à un formulaire permettant de coller une séquence ou de choisir un fichier de séquences au format Fasta à télécharger. Pour les outils *Annotate* et *Resistance*, le formulaire contient également une liste déroulante permettant de choisir le type de séquences à analyser : nucléotidiques ou protéiques. L'outil *Genotype* ne permet de soumettre que des séquences nucléotidiques. Enfin, il y a un bouton « Submit » pour lancer l'analyse, et un bouton « Reset » pour effacer les données saisies.

Une fois soumises, les séquences vont passer dans le processus d'annotation, *AnnotateHBV*, auquel il est spécifié que le point d'entrée est un fichier de séquences au format Fasta et leur type (nucléotidiques ou protéiques), à la différence du mode génération de base de données dans lequel le point d'entrée est un fichier d'entrées texte EMBL. *AnnotateHBV* est donc lancé ici en mode « servlet » (pour le web).

Dans le cas de séquences nucléotidiques, *AnnotateHBV* prend en charge les séquences et crée des entrées à partir de ces séquences, puis suit le processus d'annotation normal. A la fin de la procédure, il crée tous les fichiers nécessaires à l'affichage des résultats dans l'interface web.

Dans le cas des séquences protéiques, en mode *servlet*, la première étape de *AnnotateHBV* est de lancer une recherche de similarité avec la séquence protéique requête, dans une base de séquences protéiques de référence (programme FASTA). Cette base des protéines de référence correspond à un fichier Fasta de toutes les protéines annotées des génomes de référence, produit par le processus de génération de HBVdb, *ManageHBVdb*. L'alignement entre la séquence protéique de référence la plus proche et la séquence requête est sélectionnée et conservée. Les positions de début et de fin de la requête sont cartographiées sur l'alignement, et comme pour les séquences nucléotidiques, les fichiers nécessaires à l'affichage des résultats dans l'interface web sont créés.

Tous le fichiers produits par *AnnotateHBV* en mode servlet sont pris en charge par l'application *HBV2HTML* pour générer les fichiers de résultats HTML à afficher dans l'interface web. Pour les 3 outils (servlets), *HBV2HTML* produit une première page de résultats, la page principale ou « master ». Cette page contient un lien permettant à l'utilisateur de télécharger une archive des résultats, et un tableau dont les lignes correspondent à toutes les séquences soumises par l'utilisateur. Le tableau est composé de deux colonnes dont la première est la colonne « Query », présentant le numéro d'accession de la séquence, sous forme de lien qui mène à la page détaillée des résultats pour cette séquence. Cette première colonne, ainsi que la page détaillée, sont identiques pour les 3 outils. La deuxième colonne est spécifique de l'outil utilisé. Pour l'outil *Annotate*, la deuxième colonne affiche les CDS contenues dans la séquence requête nucléotidique, ou la protéine trouvée quand il s'agit d'une séquence requête protéique. Pour l'outil *Genotype*, elle affiche le génotype prédit sous forme de lien menant vers le fichier de sortie de l'application *GenotypeHBV*. Enfin, pour l'outil *Resistance*, la deuxième colonne donne le statut de résistance de la séquence, qui est un lien vers le fichier de sortie de *FindHBVRMut*. La Figure 32 présente un tableau « master » de résultats de l'outil *Genotype* et une page détaillée de résultats d'annotation.

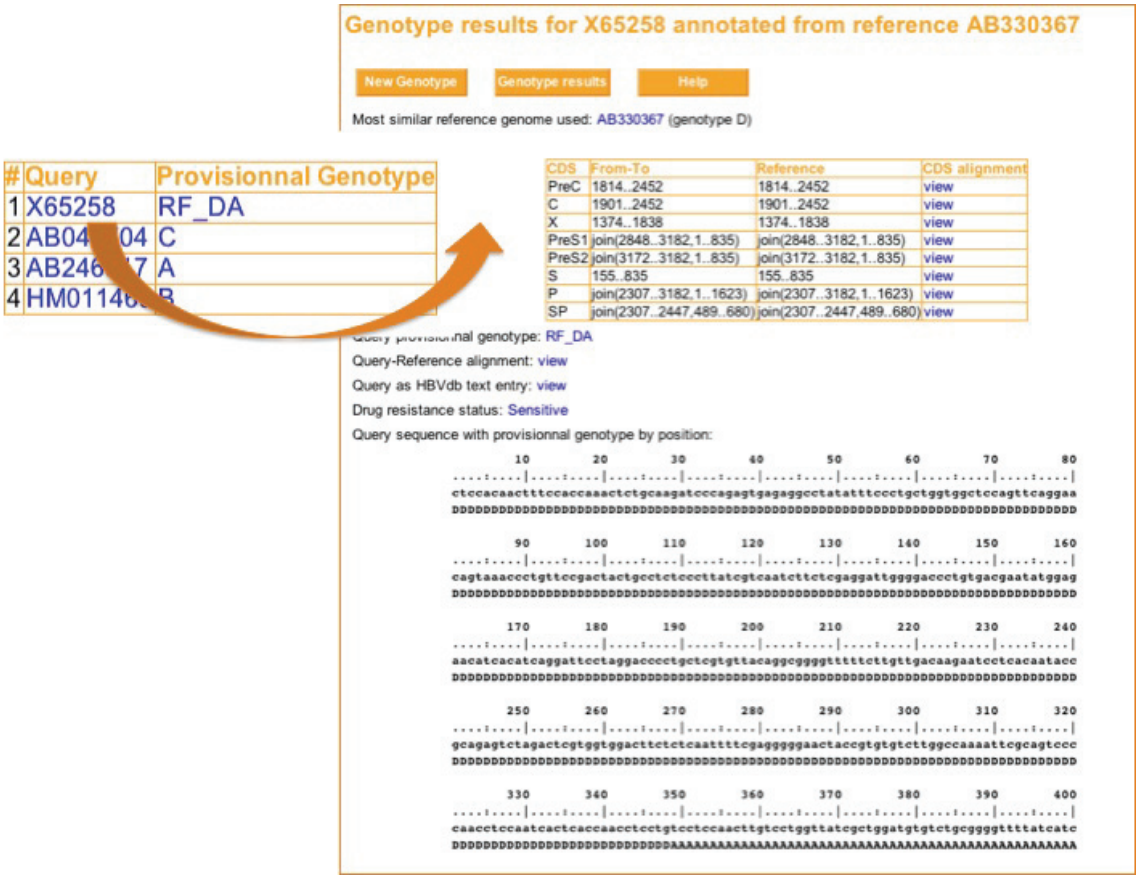


Figure 32 : Table principale de résultats de l'outil *Genotype* et page détaillée des annotations.

Pour les séquences dont l'annotation a échoué, la deuxième colonne contient la valeur « not annotated » pour l'outil *Annotate*, « not genotyped » pour l'outil *Genotype*, et « not profiled » pour l'outil *Resistance*.

La page détaillée de résultats pour une séquence protéique contient :

- le lien vers l'entrée de référence contenant la protéine la plus proche de la séquence requête
- les positions de début et de fin de la requête ainsi que les positions correspondantes sur la séquence référence
- l'alignement de la séquence requête avec la référence
- la statut de résistance pour la protéine détectée (si ce n'est pas une polymérase, le statut est n.a.)
- la séquence requête
- le lien vers le fichier de sortie de FASTA utilisé pour l'annotation.

La page détaillée de résultats pour une séquence nucléotidique contient :

- le lien vers l'entrée correspondant au génome de référence le plus similaire, utilisée pour l'annotation
- un tableau contenant toutes les CDS détectées avec leurs positions (*locations*) dans la requête et les positons correspondantes dans la référence, et les liens vers les alignements requête/référence des CDS
- le génotype prédit avec le lien vers le fichier de sortie de *GenotypeHBV*
- le lien vers l'alignement complet de la séquence requête avec la référence
- le lien vers l'entrée requête annotée en format texte HBVdb
- le statut de résistance avec le lien vers le fichier de sortie de *FindHBVRMut*
- la séquence requête avec, dessous, le génotype prédit à chaque position
- les liens vers les fichiers de sortie du programme FASTA utilisés pour l'annotation et pour le génotypage.

2.3.5.4 Menu HBVdb

Ce menu donne accès à toutes les informations relatives à la base de données HBVdb.

Le sous-menu *About* présente la base, l'équipe, les membres et plateformes impliqués dans son développement.

Le sous-menu *Contact* permet à l'utilisateur de contacter le service administrateur de HBVdb.

Le sous-menu *Help* donne accès à des pages d'aides pour l'utilisateur. Le sous-menu *Home* amène à une page d'aide générale pour le site HBVdb. Les 3 autres sous-menus donnent des aides spécifiques aux outils d'analyse spécialisés pour le VHB : *Annotate*, *Genotype* et *Resistance*.

Le sous-menu *News* informe l'utilisateur des dernières mises à jour ou ajouts de fonctionnalités concernant la base de données et le site.

Le sous-menu *Statistics* présente les statistiques d'utilisation du site web de HBVdb, avec le nombre de visites pour chaque rubrique ou menu, et l'origine géographique des visiteurs.

2.3.5.5 Menu Links

Ce menu donne accès à des liens menant vers d'autres bases de données développées au laboratoire (BYKdb (Jadeau et al., 2012), euHCVdb (Combet et al.,

2007)) ou des bases externes (ENA, UniProtKB, *etc.*), vers d'autres outils bioinformatiques, ou encore vers le site du laboratoire.

2.3.6 Statistiques

La release 4 de HBVdb (28 septembre 2012), contenait 39289 entrées annotées dont 3606 correspondaient à des génomes complets. Dans 3842 entrées, une ou plusieurs mutations de résistance ont été détectées. Concernant le génotype, la Figure 33 présente le nombre (et le pourcentage) d'entrées par génotype, sachant qu'au total, 25910 entrées ont été génotypées.

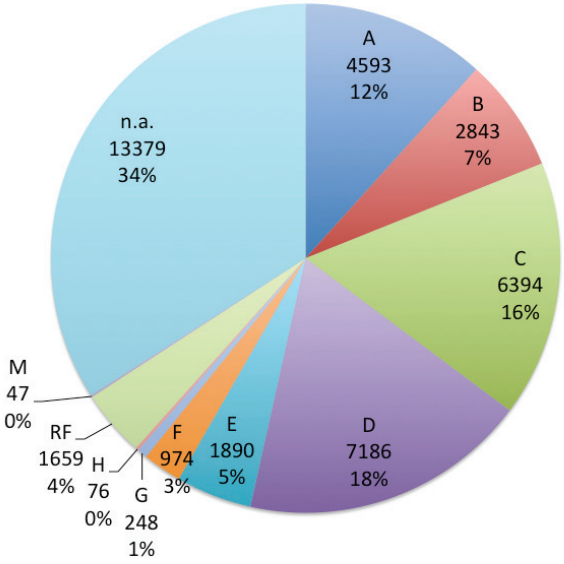


Figure 33 : Nombre d'entrées de chaque génotype.
(RF= forme recombinante, M=mosaïque, n.a. : non génotypée).

Le Tableau 8 présente le nombre de séquences annotées pour chaque *mat_peptide* et chaque *CDS*. Evidemment, ces nombres ne représentent pas que des *CDS* et *mat_peptides* complètes, mais aussi beaucoup de séquences partielles.

mat_peptide	**HBc**	**HBe**	**HBx**	**LHBs**	**MHBs**	**SHBs**	**Pol**	**HBSP**
	10659	9221	12144	29413	29481	28343	33091	27844
CDS	**C**	**PreC**	**X**	**PreS1**	**PreS2**	**S**	**P**	**SP**
	10889	12747	12332	30800	30901	29577	33682	30402

Tableau 8 : Nombre d'entrées présentant les *mat_peptide* et *CDS* spécifiés dans les colonnes.

Les séquences de *mat_peptides* et *CDS* complètes et génotypées sont regroupées dans la partie 'Dataset' du site web de HBVdb. On peut voir sur le tableau de l'Annexe 8, le nombre de séquences protéiques (*mat_peptides*) complètes pour chaque génotype.

L'histogramme de la Figure 34 représente l'évolution du nombre d'entrées dans HBVdb depuis la première release. Des modifications dans le processus d'annotation ont été faites entre la première et la deuxième release, et elles expliquent la diminution du nombre d'entrées. Des problèmes techniques sont survenus lors de la release 5, et ont engendré cette diminution du nombre d'entrées.

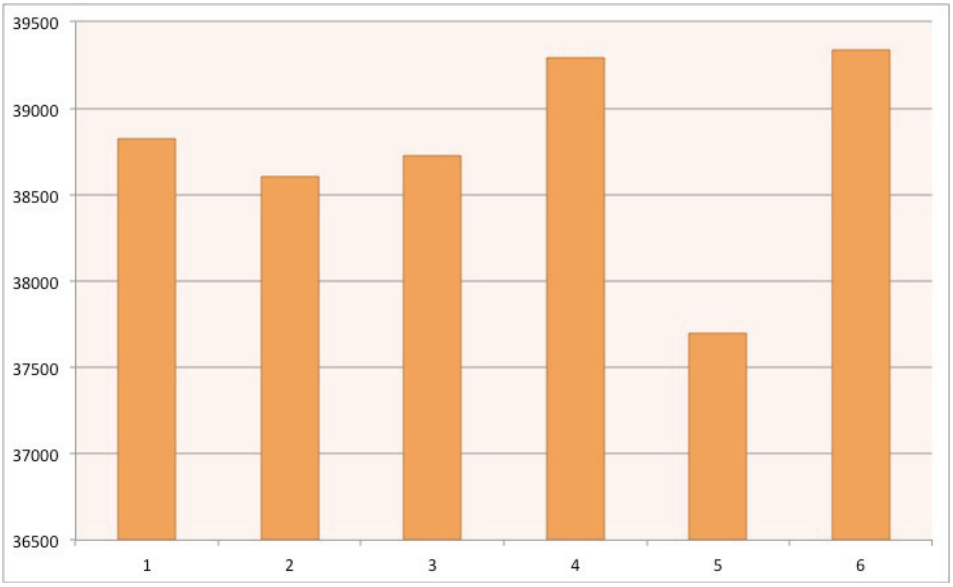


Figure 34 : Nombre d'entrées par release de HBVdb.

Les statistiques d'utilisation du site web de HBVdb depuis sa mise en ligne en juin 2012, sont présentées sur la Figure 35.

Figure 35 : Statistiques d'utilisation du site web HBVdb depuis juin 2012.

2.4 Discussion – Perspectives

La base HBVdb représente un outil très intéressant pour les chercheurs de la communauté du VHB. Elle permet de stocker toutes les séquences du VHB récupérées à partir de l'EMBL-Bank, et de les mettre à disposition des utilisateurs dans le cadre d'une base spécialisée, avec une mise à jour mensuelle des données à partir de l'ENA.

Le processus d'annotation automatique à partir d'une base de référence, assure la standardisation et la cohérence des données. La duplication des génomes de référence et des alignements entre la séquence à annoter et la référence, permet véritablement de modéliser cette circularité, et de s'affranchir des problèmes dans l'alignement des séquences.

La procédure d'annotation permet également d'établir un vocabulaire contrôlé, par exemple pour les noms des protéines et des CDS, qui sont extrêmement variables dans les bases génériques. En effet, une protéine peut être désignée par 4 ou 5 noms différents (core, capside, C, HBc, AgHBc, *etc*.), ce qui va engendrer de la perte d'information lors d'une recherche par mots clés par exemple. Ce problème est résolu dans HBVdb, car chaque protéine, et chaque CDS, est désignée par un nom unique et standard, rendant ainsi la recherche par mots clés bien plus efficace. Evidemment, cela implique de connaitre le vocabulaire standard choisi.

Le fait qu'il y ait un contrôle au niveau des alignements par CDS, et une optimisation de ces alignements si nécessaire, permet l'élimination de séquences qui présenteraient des insertions ou délétions aberrantes, pouvant être dues à des erreurs de séquençage, provoquant des décalages du cadre de lecture, menant à une ou plusieurs protéines non-fonctionnelles.

Le fait d'annoter le génotype et les mutations de résistance de manière automatique apporte des informations qui ne sont que rarement précisées dans les entrées des bases généralistes comme ENA.

La procédure de génotypage permet l'identification de génotypes « purs » (non recombinants), mais également de génotypes recombinants. Elle offre à l'utilisateur de visualiser le génotype le plus probable à chaque position de sa séquence, et donne une valeur de confiance avec le nombre de positions informatives utilisées pour prédire le

génotype. Une comparaison de notre outil avec les autres outils de génotypage disponibles a montré que son efficacité est équivalente à celle de tous les outils qui permettent de détecter des génotypes recombinants.

Au cours de différentes analyses effectuées lors du développement de l'outil de génotypage, j'ai pu remarquer que dans certains cas, la règle des 8% de différence pour déterminer que deux séquences sont de génotypes différents (Schaefer, 2007), ne s'applique pas toujours. Il existe des exceptions à cette règle, et il y a même à l'heure actuelle, des génotypes considérés comme purs, qui sont le résultat de recombinaisons. Par exemple, certains génomes de génotype B, sont en réalité composés d'une partie de séquence se rapprochant plus du génotype C, et pourraient être considérés comme des recombinants BC.

Le programme *FindHBVRMut*, accessible depuis le menu *Resistance* de l'interface web, qui définit les profils de résistance aux analogues de nucléos(t)ides d'une ou plusieurs séquences, peut permettre au clinicien, de connaître de manière claire et rapide les mutations existantes, et peut donc l'aider à prendre une décision quant à la suite d'un traitement. Pour le moment, le programme prend en entrée un fichier des profils connus qui est tenu à jour en fonction des « guidelines » de l'EASL. A l'avenir, une base de données sera développée pour stocker les profils de résistance avec les drogues, mutations et statut associés, et certainement d'autres informations. Un formulaire de saisi y sera associé pour permettre d'entrer dans la base toute nouvelle mutation ou nouveau profil de résistance. Elle sera bien sûr directement en lien avec l'outil *FindHBVRMut* pour qu'il puisse détecter les profils de résistance les plus récents.

Dans son implémentation actuelle, l'outil ne décrit pas les mutations de résistance qui ont également un potentiel d'échappement à la vaccination (ADAPVEM, antiviral drug-associated potential vaccine-escape mutants) (Locarnini and Yuen, 2010).

Dans les perspectives d'évolution de la base HBVdb, la première se situe au niveau des annotations, avec l'ajout d'autres types d'annotations dans les références. Par exemple, en plus des CDS et des protéines, on pourrait ajouter l'annotation des promoteurs et des éléments régulateurs.

Une évolution à plus long terme serait de permettre à l'utilisateur de stocker ses propres séquences dans HBVdb et de pouvoir y associer des données cliniques. Ceci

permettrait de faire de vraies analyses de corrélation entre les données génétiques et phénotypiques.

La base HBVdb peut être utilisée pour faire des études à grande échelle sur les séquences, puisqu'elle donne accès à un très grand nombre de séquences du VHB. Par exemple, on peut analyser la variabilité des protéines, voir quelles sont les positions les plus variables, et ce en fonction des génotypes, ou encore regarder s'il y a des positions présentant des résidus spécifiques pour certains génotypes. HBVdb constitue une ressource intéressante à partir de laquelle on peut faire de nouvelles découvertes concernant le VHB, grâce aux séquences annotées qu'elle contient et à leur analyse.

Chapitre 3 :

Analyses de séquences de virus des hépatites B et C

Ce chapitre expose des travaux d'analyse de séquences réalisés dans le cadre d'études sur les virus des hépatites B et C. Il est présenté sous forme d'article.

3.1 Introduction

Dans une première partie, je présenterai une analyse de séquences que j'ai réalisée dans le cadre d'une étude sur le virus de l'hépatite C (VHC). Cette étude a été effectuée en collaboration avec l'équipe du Professeur Baumert et a fait l'objet d'une publication dans le journal Gastroenterology. Dans les deux parties suivantes, j'exposerai des travaux d'analyse de séquences dans le cadre d'une étude sur la polymérase du virus de l'hépatite B.

Les principaux traitements utilisés contre le VHB sont des analogues de nucléos(t)ides qui ciblent le domaine RT de la polymérase virale. Or, l'utilisation prolongée de ces traitements a sélectionné des populations virales résistantes aux NA, ayant une meilleure réplication que les virus sauvages en présence de drogues.

On pourrait envisager le développement de nouvelles molécules spécifiques, ayant pour cible un autre domaine que la RT. Le domaine RNase H est nécessaire à la réplication virale puisqu'il porte l'activité enzymatique permettant de dégrader le brin d'ARNpg qui forme un hétéro-duplex avec le brin d'ADN moins en cours de synthèse. Pour cette raison, la RNase H est une cible de choix pour le développement de nouveaux médicaments. Or, le développement de molécules inhibitrices spécifiques nécessite de connaître la structure tridimensionnelle de la protéine, ou du domaine protéique à inhiber. Malheureusement, la structure 3D de la polymérase du VHB n'a pas été résolue, ni même les domaines RT et RNase H.

Pour pallier ce manque de structure résolue expérimentalement, nous avons décidé de modéliser par homologie la structure de la RNase H du VHB.

La modélisation par homologie consiste à prédire la structure tridimensionnelle d'une séquence protéique à partir d'une séquence protéique homologue dont la structure est résolue. Cette structure résolue constitue ce que l'on appelle « l'empreinte ». La modélisation par homologie passe par plusieurs étapes majeures :

- l'identification d'une empreinte
- l'alignement de la séquence à modéliser avec la séquence de l'empreinte

- la modélisation en elle même, par un logiciel de modélisation moléculaire.

La première étape de la modélisation par homologie est donc une étape d'analyse de séquences, et plus précisément de recherche de séquences homologues. J'exposerai cette étape dans ce chapitre. Je présenterai ensuite des analyses de séquences réalisées sur le domaine RNase H afin d'en déterminer certaines caractéristiques, notamment à l'aide de HBVdb.

En effet, la rubrique « Dataset » de HBVdb fournit à l'utilisateur des jeux de séquences complètes de CDS ou de protéines, pour chaque génotype. Nous avons utilisé certaines de ces données pour analyser la variabilité des séquences de RNase H du VHB. Ces analyses à partir des données de la base HBVdb ont été complétées par une étude de la variabilité de la RNase H virale, analysée par des techniques de séquençage à haut débit ou pyroséquençage (UDPS, ultra deep pyrosequencing) sur des virus recueillis chez des patients infectés.

Dans un deuxième temps, nous avons positionné la RNase H dans son contexte, la polymérase. Ainsi, nous avons voulu construire un modèle moléculaire comprenant les domaines RT et RNase H. La deuxième partie de ce chapitre présente les analyses de séquences préliminaires à cette modélisation, incluant la recherche de séquences homologues à la polymérase du VHB.

3.2 Matériel et méthodes

3.2.1 Analyse qualitative et quantitative de la variabilité

Les données de départ pour ces analyses sont des alignements multiples de séquences protéiques. Après une conversion des alignements au format Clustal W, les répertoires de résidus ont été calculés pour chaque alignement multiple. Ils donnent les fréquences de chaque acide aminé à chaque position de l'alignement. Les répertoires de résidus présentent tous les résidus qui sont présents dans plus de 1% des séquences à une position donnée de l'alignement. Ces répertoires permettent donc de détecter, de manière qualitative, les positions qui varient.

Une autre approche que nous avons utilisée pour étudier la variabilité des séquences, est le calcul de l'entropie de Shannon (Shannon et al., 1948). Elle donne une mesure quantitative de la variabilité pour chaque position de l'alignement. Nous avons utilisé l'entropie de Shannon normalisée, pour avoir une valeur de variabilité comprise entre 0 et 1.

Les répertoires de résidus et les valeurs d'entropie de Shannon ont été calculés grâce à l'application *computeRepertoire*. Cette application prend en entrée un fichier d'alignement au format Clustal W. Une option permet de spécifier le seuil minimum à partir duquel le répertoire doit présenter un résidu, en effectif ou en fréquence. Une autre option permet de spécifier si l'on souhaite un calcul de l'entropie de Shannon (normalisée ou non) pour toutes les positions de l'alignement. D'autres options permettent de comparer deux répertoires de résidus (donc deux alignements).

Cette application fournit deux fichiers de sortie : un présentant le répertoire de résidus (extension .rep) et un présentant les effectifs ou les fréquences de tous les résidus pour chaque position de l'alignement (extension .num).

3.2.2 Extraction des séquences de HBVdb et alignements multiples

Les séquences protéiques ont été extraites de la base HBVdb (Release 1.0) grâce à des requêtes par mots clés sur les *PRABI_name* des *features mat_peptide*, avec une jointure sur les entrées ayant un *PRABI_genotype* donné dans le *feature source*.

Ensuite, les séquences ont été alignées avec MUSCLE (version 3.8).

3.2.3 Recherche d'empreinte pour la modélisation par homologie

Il s'agit de trouver une séquence homologue dont la structure est connue. Plusieurs outils bioinformatiques ont été utilisés pour cette étape, les principaux sont des outils de recherche de séquences similaires tels que BLAST et PSI-BLAST (Altschul et al., 1997). PSI-BLAST permet de chercher des séquences homologues distantes, grâce à la construction de profils de manière itérative. Nous avons effectué des PSI-BLAST avec une base protéique contenant des séquences de UniProtKB non redondantes (à 60% d'identité maximum entre les séquences), ainsi que les séquences des structures de la PDB. Puis, nous avons cherché si parmi les séquences similaires, il y avait des séquences correspondant à des structures disponibles dans la PDB.

J'ai également utilisé Dali server (Holm and Rosenström, 2010), pour chercher des structures proches d'une protéine de même fonction que celle à modéliser. Une fois quelques structures similaires trouvées, j'ai procédé à des alignements structuraux avec Matt (Multiple Alignment with Translations and Twists) (Menke et al., 2008) ou directement Dali. Matt est un algorithme de chaînage de paire de fragments alignés. Il permet une flexibilité locale entre fragments : des petites translations et rotations sont temporairement autorisées pour rapprocher des fragments alignés. Il est efficace dans l'alignement de structures de protéines ayant une homologie distante, car il est capable de modéliser les distorsions du squelette (backbone) dans les protéines apparentées les plus distantes.

A partir de ces alignements structuraux, j'ai utilisé le package HMMER 3 (Eddy, 2011) pour construire des profils HMM (modèles de Markov cachés) avec *hmmbuild*. J'ai ensuite utilisé ces profils, pour chercher des séquences s'alignant sur ce profil, grâce à *hmmsearch*.

3.3 Résultats et discussion

3.3.1 Analyses de séquences sur la glycoprotéine E2 du virus de l'hépatite C

J'ai réalisé des analyses de séquences du virus de l'hépatite C (VHC), dans le cadre d'une collaboration avec l'équipe du Professeur Thomas Baumert à Strasbourg (Inserm U748, Hôpitaux de Strasbourg). Ces travaux ont fait l'objet d'une publication dans Gastroenterology (Fofana et al., 2012) (article 2).

En raison de l'échappement viral à la réponse immunitaire de l'hôte, la réinfection du greffon est courante. Les mécanismes par lesquels le virus échappe à l'immunité de l'hôte pour réinfecter le greffon ne sont pas connus. Dans une analyse de six patients atteints du VHC qui ont subi une transplantation du foie, l'équipe du Professeur Baumert avait montré que les variants du VHC qui réinfectaient le greffon hépatique étaient caractérisés par une entrée efficace et une neutralisation peu efficace par les anticorps. Ces variants, présents dans le sérum avant la transplantation, étaient majoritaires après

la transplantation par rapport à d'autres variants qui n'étaient plus détectés (Fafi-Kremer et al., 2010).

Pour la deuxième étude, ils ont isolé le cas du patient P01, qui présentait le variant VL comme variant majoritaire après transplantation (Figure 36). Les variants de type VL présentaient une entrée virale très efficace et un échappement à la neutralisation par les anticorps, contrairement aux variants de type VA et VC.

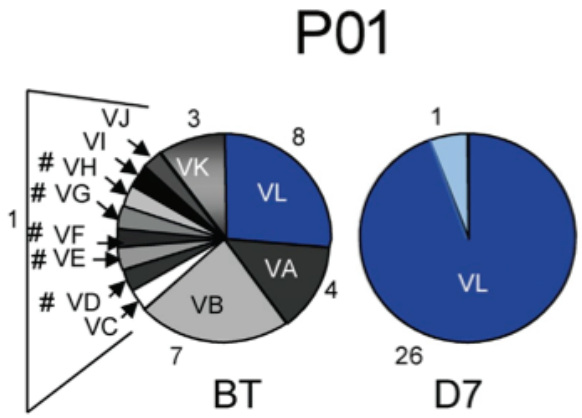


Figure 36 : Population virale des variants du VHC avant transplantation, et 7 jours après.
(Fafi-Kremer et al., 2010)

L'entrée du VHC est un processus qui passe par les glycoprotéines de l'enveloppe virale E1 et E2 et plusieurs facteurs de l'hôte, par exemple le récepteur CD81, la claudine-1, etc. Ils ont donc fait l'hypothèse que les différences entre ces variants se situaient dans les protéines d'enveloppe E1 et/ou E2.

En comparant les séquences de ces 3 variants, ils ont découvert plusieurs positions présentant des variations de résidus entre les variants VL et VC, et VL et VA. Après quelques analyses et expérimentations, 3 de ces positions ont été proposées comme étant impliquées dans l'échappement à la neutralisation et dans l'augmentation de l'efficacité de l'entrée virale.

La position importante qui change entre les variants VA et VL est la position 447 à laquelle VL présente une phénylalanine (F) et VA une Leucine (L).

Les deux positions concernées entre VL et VC sont 458 et 478. Le variant VL porte les résidus S458 (sérine) et R478 (arginine), et le variant VC présente les résidus G458 (glycine) et C478 (cystéine).

Ils ont montré que si l'on introduit les mutations S458G et R478C dans le variant VL, l'efficacité de l'entrée virale est réduite, et la neutralisation par les anticorps est augmentée. Inversement, si l'on introduit les mutations G458S et C478R dans le variant VC, l'efficacité de l'entrée virale et la neutralisation par les anticorps sont similaires à ce qui était observé pour le variant VL. De la même manière, après introduction de la mutation F447L dans le variant VL, ils observaient un phénotype similaire à celui de VA.

Ces résultats suggèrent que les résidus en positions 447, 458, et 478 sont responsables à la fois d'une entrée virale accrue, et de l'échappement à la neutralisation par les anticorps.

L'équipe du Professeur Baumert nous a alors proposé de faire des analyses bioinformatiques sur les séquences de la glycoprotéine E2, et plus précisément sur ces 3 positions.

Nous avons donc extrait 2074 séquences de E2 de la base de données euHCVdb (Combet et al., 2007), dont un sous-ensemble de séquences de génotype 1b (correspondant au génotype des variants du patient P01). A partir de là, nous avons produit 2 alignements multiples qui nous ont permis de calculer les répertoires de résidus de E2, à l'aide de l'outil *ComputeRepertoire*.

Les fréquences des résidus pour les positions 447, 458, et 478 sont présentées sur la Figure 37 sous forme d'histogrammes. Il y a donc un répertoire correspondant aux séquences E2 de tous les génotypes (en noir), et un répertoire des séquences E2 de génotype 1b (en blanc).

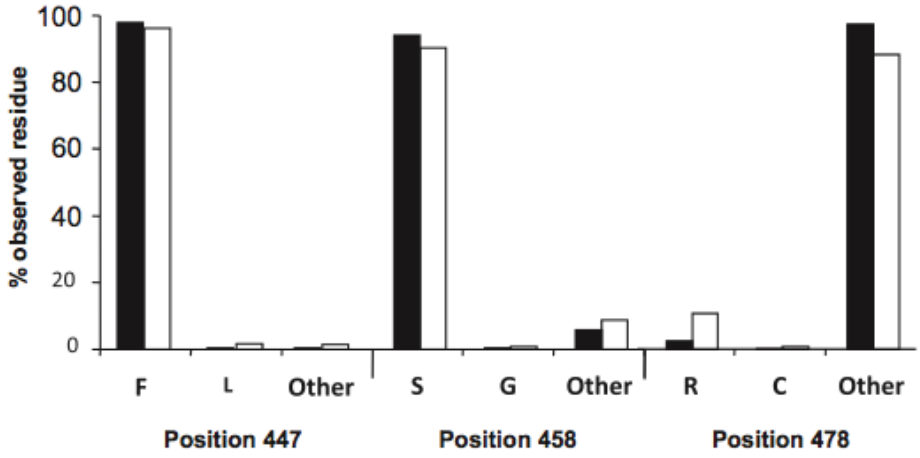


Figure 37 : Pourcentage des résidus observés aux positions 447, 458 et 478.
A partir des séquences de tous génotypes en noir, et des séquences de génotype 1b en blanc. (Fofana et al., 2012)

Les résidus F, S et R sont observés beaucoup plus fréquemment aux positions 447, 458, et 478 que L, G et C. F et S sont les résidus prédominants dans les positions 447 et 458 dans la grande majorité des souches 1b, respectivement (F447 tous génotypes : 98,4% ; 1b : 96,2% ; S458 tous génotypes : 94% ; 1b : 90,3%). La position 478 est variable, mais R (tous génotypes : 2,4% ; 1b : 10,8%) est plus fréquent que C (tous génotypes : 0,2% ; 1b : 0,9%). La prévalence élevée de résidus identifiés soutient leur pertinence fonctionnelle pour la survie et la sélection du virus, car des résidus plus pertinents structuralement et fonctionnellement seront observés plus fréquemment. Ces données suggèrent que l'épitope contenant les résidus identifiés aux positions 447, 458, et 478 est responsable non seulement de l'échappement viral aux anticorps antiviraux au cours de la transplantation hépatique, mais peut aussi contribuer à l'évasion virale dans l'infection chronique par le VHC en général.

D'après un modèle de l'organisation structurale de la protéine E2 (Krey et al., 2010), les positions 447, 458, 478 sont à proximité du domaine de liaison au récepteur CD81. Or, cette étude a montré que les mutations F447L, S458G, et R478C modulaient la dépendance à CD81 de l'entrée du VHC, altéraient l'interaction avec CD81. On peut donc penser que ces positions pourraient se trouver dans les 2 boucles appartenant à un groupe de boucles exposées à la surface.

Ces boucles sont probablement impliquées dans l'interaction E2-CD81 soit directement, soit indirectement, comme un point clé de la réorganisation structurale au cours de l'entrée virale.

Les résidus polaires S et R présents dans le variant d'échappement peuvent former des interactions avec d'autres résidus par des liaisons hydrogène et des ponts salins, respectivement. Ces interactions pourraient augmenter la stabilité de l'interface d'interaction CD81-E2, permettant l'entrée efficace du variant d'échappement VL à travers les complexes de corécepteurs E2-CD81-CLDN1, qui sont des facteurs déterminants pour l'entrée du virus.

Par ailleurs, le groupe de boucles E2 contenant les mutations portent des épitopes linéaires mais définit aussi au moins un épitope conformationnel qui est une cible pour les anticorps neutralisants. Selon les propriétés physico-chimiques des résidus, les résidus S458 et R478 du variant VL améliorent l'hydrophilie des boucles et peuvent favoriser leur exposition à la surface. Ce changement pourrait moduler les interactions

E2-CD81 plus loin et affecter la liaison des anticorps neutralisants en bloquant l'accès à leurs épitopes cibles.

En conclusion, ces données ont permis d'identifier les principaux déterminants de l'échappement à la réponse immunitaire *in vivo*. Des mutations conférant un échappement à la neutralisation altèrent l'utilisation du récepteur CD81 et augmentent l'entrée cellulaire.

3.3.2 Vers une modélisation du domaine RNase H du VHB

La polymérase du VHB est composée de 4 domaines : le domaine TP, le domaine charnière (ou spacer), le domaine transcriptase inverse (RT) et le domaine RNase H. Ce dernier domaine porte l'activité enzymatique ribonucléase. Il permet de dégrader le brin d'ARN, correspondant à l'ARNpg servant de matrice à la réplication, qui forme un hétéro-duplex avec le brin moins d'ADN en cours de polymérisation. Une fois le brin d'ARNpg dégradé, le brin plus d'ADN pourra être synthétisé pour former le nouveau génome viral sous forme d'ADN-RC.

Le domaine RNase H de la Pol du VHB n'est pas très bien connu, et notamment, on ne connaît pas sa structure tridimensionnelle, tout comme on ne connait pas celle du reste de la Pol.

En revanche, certaines structures de RNase H virales et non virales on été résolues. Parmi les RNase H virales dont la structure tridimensionnelle est connue, il y a celle de Moloney Murine Leukemia Virus (MoMLV) (Lim et al., 2006) et celle du VIH-1 (virus de l'immunodéficience humaine de type 1). La structure de la RNase H du VIH-1 correspond en réalité à la structure de la sous-unité p66, qui comprend le domaine RT et le domaine RNase H, liés par un domaine de connexion (Sarafianos et al., 2001).

D'autres structures de RNase H sont disponibles pour d'autres organismes procaryotes et eucaryotes. Il existe deux types de RNase H : type 1 et type 2 (Tadokoro and Kanaya, 2009; Champoux and Schultz, 2009). Les RNase H de type 1, auxquelles appartiennent celles du VIH-1 et du MoMLV, ont toutes une organisation structurale similaire. Certains organismes, eucaryotes et procaryotes, possèdent des RNase H des deux types (Tadokoro and Kanaya, 2009; Cerritelli and Crouch, 2009).

Le site catalytique des RNase H, quel que soit le type, est formé d'une tétrade catalytique. Les RNase H de type 1 assurent leur activité catalytique grâce une tétrade conservée de résidus D-E-D-D, avec un mécanisme impliquant deux cations divalents

(Mg^{2+}) (Cerritelli and Crouch, 2009; Champoux and Schultz, 2009; Tadokoro and Kanaya, 2009). La polymérase du VHB possédant des domaines RT et RNase H, tout comme les rétrovirus, on peut penser que leur organisation est similaire. Ceci suggère que la RNase H du VHB est une RNase H de type 1, comme celles des rétrovirus, et qu'elle possède une tétrade catalytique de type D-E-D-D.

Les RNase H de type 1 ont également une autre caractéristique : une protrusion basique, qui peut comporter ou non une hélice α, appelée hélice C, suivie d'une boucle basique. D'après Cerritelli et Crouch, chez les eucaryotes la protrusion basique comprend deux hélices α : l'hélice C et l'hélice D (Cerritelli and Crouch, 2009).

Dans les RNase H de type 1, cette protrusion basique est située, en séquence, entre les deux derniers acides aspartiques catalytiques de la tétrade (Tadokoro and Kanaya, 2009). Cette protrusion basique est impliquée dans l'interaction avec l'acide nucléique. Un exemple d'organisation de RNase H de type 1, celle de *E. coli*, est présenté sur la Figure 38.

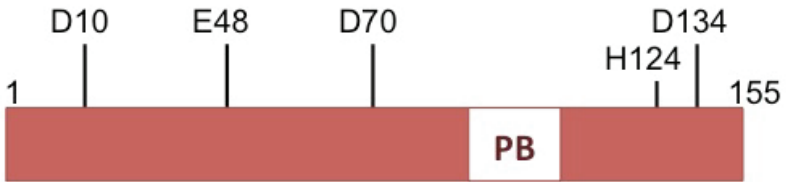


Figure 38 : Organisation de la RNase H de type 1 de *E. coli* présentant la tétrade catalytique.
La protrusion basique est représentée par la portion blanche notée PB. L'histidine conservée précédant le dernier D catalytique est représentée (H124).

Concernant les virus, MoMLV possède cette protrusion basique, qui comprend l'hélice α C d'après Lim *et al*., alors que chez le VIH-1, l'hélice C est absente. Visiblement chez le VIH-1, d'un point de vue structural, l'hélice C est remplacée par un domaine de connexion plus important, dont une région serait impliquée dans l'interaction avec l'acide nucléique, et donc remplacerait la fonction de l'hélice C (Lim et al., 2006). Chez les procaryotes, il y a également l'exemple de *E. coli* qui possède une hélice C, alors que *B. halodurans* n'en possède pas (Champoux and Schultz, 2009).

La Figure 39 présente une superposition des structures de RNase H de *E. coli* (cyan, PDB : 2RN2), MoMLV (vert, PDB : 2HB5) et VIH-1 (rose, 1HRH). L'hélice C de la RNase H du MoMLV est délétée dans la protéine cristallisée donc elle n'apparaît pas, mais l'hélice C de *E. coli* est représentée en bleu foncé.

Figure 39 : Superposition de 3 structures expérimentales de RNases H.
E. coli (cyan), MoMLV (vert) et VIH-1 (rose). Représentation en mode cartoon.

Comme la structure tridimensionnelle de la RNase H du VHB n'est pas connue et que les structures des RNase H connues sont similaires, nous avons choisi de modéliser la structure de la RNase H du VHB en utilisant une approche de modélisation par homologie, dont la première étape est l'identification d'une empreinte.

3.3.2.1 Recherche d'homologues pour la sélection d'une empreinte

Il s'agit de trouver une séquence homologue à la séquence à modéliser, et dont la structure tridimensionnelle est connue.

Après plusieurs BLAST et PSI-BLAST avec la séquence de RNase H du VHB comme requête, et des séquences de UniProtKB et de la PDB comme base, aucune séquence homologue de structure connue n'est ressortie. J'ai alors fait l'hypothèse que la RNase H du VIH-1 devait avoir une structure proche de celle du VHB, tout comme cette hypothèse à été faite pour les modèles de RT publiés. J'ai donc utilisé Dali server (Holm and Rosenström, 2010) pour récupérer des structures proches de celle de la RNase H du VIH-1 (PDB : 1HRH). Parmi les résultats, il y avait les structures de RNase H de d'autres

organismes : MoMLV (Moloney Murine Leukemia Virus, PDB : 2HB5) (Lim et al., 2006), un virus xénotropique relié à MoMLV (XMRV, PDB : 3P1G) (Kirby et al., 2012), l'humain (PDB : 2QKK) (Nowotny et al., 2007), *E. coli* (PDB : 1GOA) (Ishikawa et al., 1993), et une Archae hyperthermophile (*Sulfolobus tokodaii str. 7*, PDB : 2EHG). J'ai donc sélectionné ces structures et calculé un alignement structural à l'aide de Dali. J'ai utilisé cet alignement structural (Figure 40) pour calculer un profil HMM (avec *hmmbuild*, du package HMMER 3).

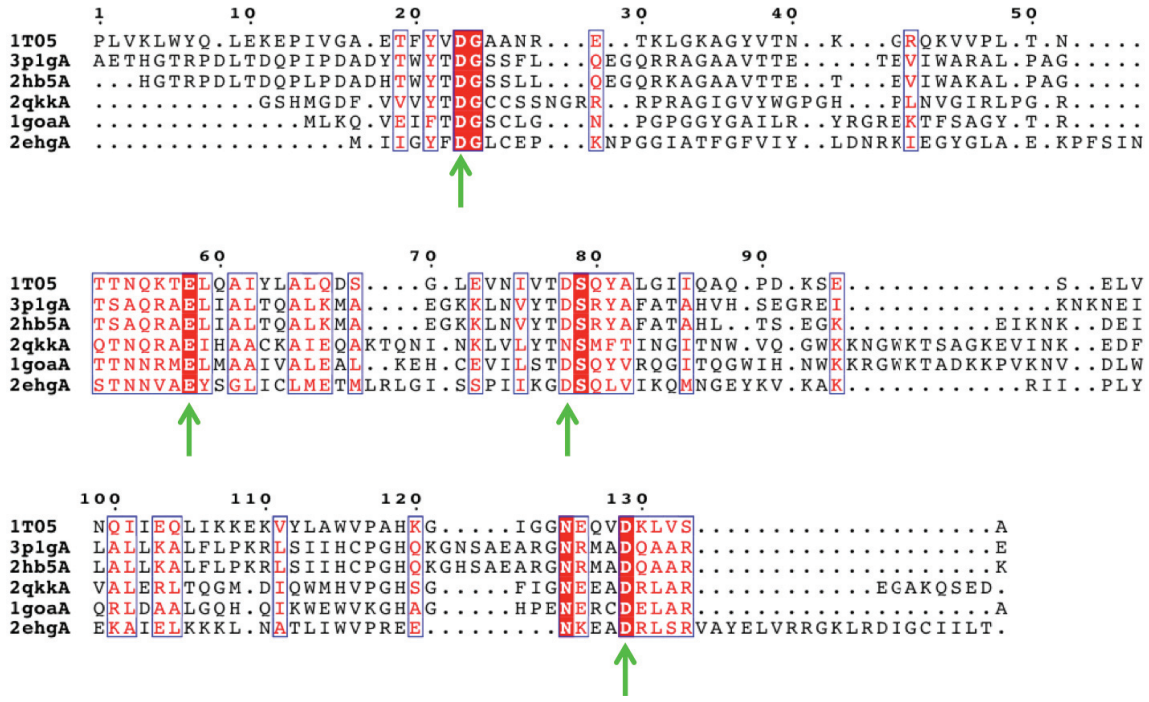


Figure 40 : Alignement structural de RNases H virales, procaryote, eucaryote, et d'Archae.
1T05 : VIH-1, 3P1G : XMRV, 2HB5 : MoMLV, 2QKK : humaine (le 3ème D catalytique est muté en N), 1GOA : E. coli, 2EHG : Archae hperthermophile. Les flèches vertes pointent sur les résidus de la tétrade catalytique. Figure réalisée avec ESPript (Gouet et al., 2003).

Ensuite, j'ai cherché dans UniProtKB, les séquences qui s'alignaient avec ce profil, à l'aide de l'outil *hmmsearch*. Après plusieurs enrichissements du profil avec des séquences qui s'alignait dessus et renouvellement de la recherche, aucune séquence du VHB n'est apparue dans les résultats.

J'ai ensuite calculé un alignement structural à partir des RNase H d'origine virale uniquement. Cet alignement comprenait les structures suivantes : la partie RNase H de la structure PDB : 1T05 (sous-unité p66 du VIH-1) (Tuske et al., 2004), PDB : 2HB5

(MoMLV) et PDB : 3P1G (XMRV). Après la construction d'un profil HMM et la recherche de séquences avec ce profil, puis plusieurs itérations et enrichissements du profil, aucune séquence du VHB n'a été détectée.

J'ai effectué le même type de recherches plusieurs fois en utilisant des alignements structuraux de départ différents à chaque fois, calculés par Dali ou par Matt (Menke et al., 2008) et en combinant plusieurs structures de RNase H. Ces analyses n'ont pas abouti à détecter une séquence de RNase H du VHB.

J'ai ensuite réalisé une recherche similaire en utilisant l'alignement structural des RNase H de MoMLV, VIH-1, *E. coli* et *B. halodurans* (PDB : 1ZBF) (Nowotny et al., 2005) publié par Lim *et al.* (Lim et al., 2006). Même après plusieurs itérations de *hmmsearch* avec enrichissement du profil, la séquence de RNase H du VHB n'est pas sortie dans les résultats.

Malgré la sensibilité des méthodes de profils, l'homologie ces RNase H est donc difficile à prouver à partir des séquences. Cependant, les données structurales et fonctionnelles sur les RNase H virales, procaryotes et eucaryotes laissent penser qu'il s'agit d'homologues distants. En effet, les structures des RNase H de type 1 étant assez conservées (Tadokoro and Kanaya, 2009), même si les séquences sont très divergentes, on pourrait utiliser comme empreinte, l'une des deux RNase H virales dont on dispose de la structure, VIH-1 ou MoMLV.

Comme il a été précisé précédemment, certaines RNase H possèdent une hélice α nommée hélice C. Cette hélice, impliquée dans les interactions avec l'hétéro-duplex, est présente chez MoMLV et chez *E. coli* mais elle est absente de la RNase H du VIH-1. D'après la longueur de la séquence du domaine RNase H du VHB, et d'après des prédictions de structures secondaires, on peut émettre l'hypothèse que la RNase H du VHB possède l'hélice C.

J'ai réalisé un alignement structural des RNase H de MoMLV (PDB : 2HB5) et *E. coli* (PDB : 2RN2), qui possèdent toutes les deux une hélice C. Cependant, dans le fichier PDB de la RNase H de MoMLV, l'hélice C n'est pas présente car elle a été délétée dans la protéine cristallisée. J'ai donc inséré manuellement dans mon alignement la séquence correspondant à l'hélice C de MoMLV, selon l'alignement publié par Lim *et al*. (Lim et al., 2006). Puis, à l'aide de Clustal W, j'ai aligné la séquence de RNase H du VHB à cet alignement structural (Figure 41). Dans l'alignement résultant, la région prédite en

hélice α à partir des méthodes de prédiction de structures secondaires s'aligne avec les hélices C de MoMLV et de *E. coli*.

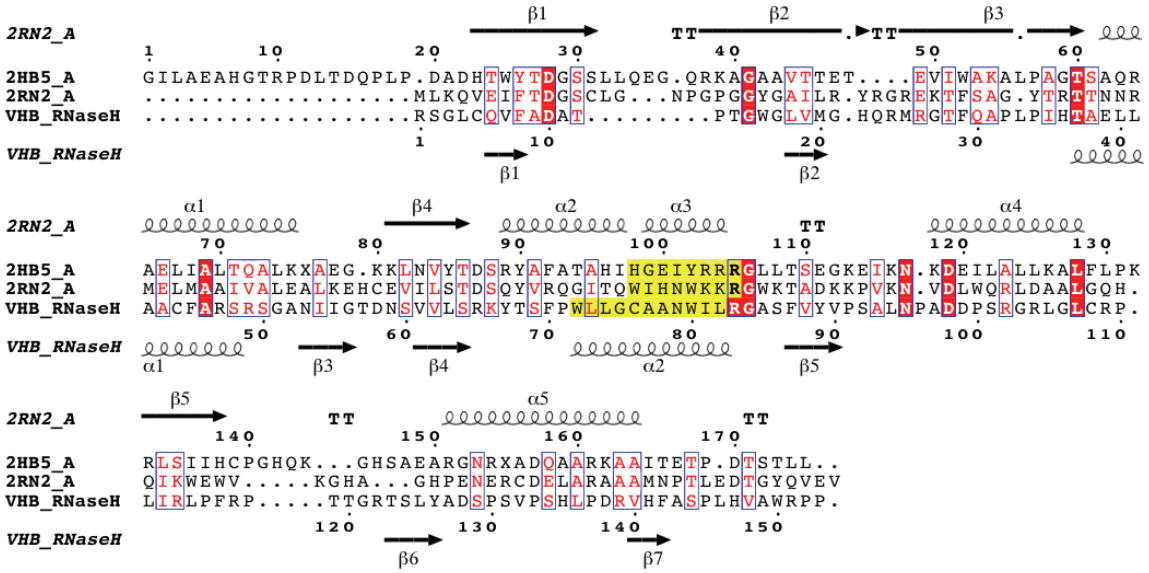


Figure 41 : RNase H du VHB alignée avec l'alignement structural de celles de MoMLV et *E. coli*. 2HB5 : MoMLV, 2RN2 : *E. coli*. Les résidus des hélices C sont surlignés en jaune. Les structures secondaires de E. coli et prédites pour le VHB sont présentées. Figure réalisée avec ESPript (Gouet et al., 2003).

La séquence du VHB présente une insertion par rapport au VIH, qui pourrait donc être assimilée aux hélices C de MoMLV et de *E. coli*. La meilleure empreinte pour la modélisation semble alors être la structure de la RNase H du MoMLV. Or, la structure de RNase H de MoMLV disponible dans la PDB (2HB5) présente la délétion au niveau de l'hélice C (Lim et al., 2006).

Pour cette raison, nous avons sélectionné la structure de la RNase H de *E. coli* (PDB : 1RDD) (Katayanagi et al., 1993) comme empreinte, comme l'avait fait Potenza *et al*. qui avait publié un modèle de la RNase H du VHB (Potenza et al., 2007). L'étape d'alignement entre la séquence du VHB à modéliser et la séquence de l'empreinte, sera présentée dans le chapitre de modélisation moléculaire (partie 4.3.1.1).

Nous avons fait d'autres analyses au niveau des séquences de la RNase H du VHB, notamment pour étudier la variabilité des séquences et aussi pour déterminer les positions des 4 résidus acides formant la tétrade catalytique.

3.3.2.2 Analyse de la variabilité de la RNase H

3.3.2.2.1 Positions présentant des variations

3.3.2.2.1.1 Etude des séquences de tous les génotypes à partir de HBVdb

Dans la rubrique des « dataset » protéiques de HBVdb (Release 1.0), nous avons récupéré les alignements multiples, les fichiers de répertoires de résidus, et les fichiers de fréquence des résidus de la protéine Pol pour chacun des 8 génotypes. A partir de ces données, nous avons analysé la variabilité des 153 dernières positions de la Pol, correspondant au domaine RNase H. A partir de chaque répertoire de résidus et du fichier de fréquences associé, j'ai relevé, pour chacun des génotypes, les positions présentant au moins deux résidus présents dans plus de 1% des séquences. Le Tableau 9 expose les positions qui présentent des variations. Le domaine RNase H débute par le motif RSGLCQV à la position 680 (numérotation UniProtKB : C7DSI5).

Génotype	Nombre de séquences récupérées	Nombre de positions présentant des variations	Positions qui varient (numérotation RNase H)
A	676	9	681, 732, 756, 787, 795, 796, 807, 817, 830
B	121	22	681, 702, 704, 707, 709, 713, 732, 747, 756, 764, 772, 790, 793, 794, 795, 796, 798, 807, 813, 814, 817, 830
C	963	16	681, 683, 699, 702, 704, 732, 733, 734, 786, 792, 796, 807, 814, 815, 817, 830
D	193	33	680, 681, 682, 698, 699, 702, 709, 710, 712, 720, 721, 733, 734, 749, 756, 764, 767, 786, 787, 791, 792, 793, 795, 796, 798, 807, 817, 826, 827, 828, 829, 830, 831
E	223	23	681, 699, 701, 704, 712, 732, 752, 760, 762, 786, 787, 792, 796, 798, 806, 812, 813, 814, 815, 817, 820, 830, 832
F	97	20	681, 704, 709, 732, 733, 734, 765, 778, 786, 787, 792, 793, 794, 796, 807, 814, 817, 822, 828, 830
G	29	14	720, 732, 752, 758, 759, 760, 762, 771, 781, 786, 787, 794, 798, 809
H	24	12	704, 709, 733, 753, 773, 787, 794, 798, 813, 815, 822, 828

Tableau 9 : Positions de la RNase H présentant des variations, pour chaque génotype.
D'après des données de HBVdb (Release 1.0).

Au total, si l'on prend en compte tous les génotypes, il y a 63 positions qui présentent des variations dans le domaine RNase H. Parmi ces 63 positions, 31 sont retrouvées comme variant dans au moins 2 génotypes dont 15 dans au moins la moitié des

génotypes. Il a également 32 positions qui varient spécifiquement dans un génotype. Elles sont présentées dans le Tableau 10.

Génotype	Positions présentant des variations spécifiques	Nombre de positions variant spécifiquement
A	-	0
B	707, 713, 747, 772, 790	5
C	683	1
D	680, 682, 698, 710, 721, 749, 767, 791, 826, 827, 829, 831	12
E	701, 806, 812, 820, 832	5
F	765, 778	2
G	758, 759, 771, 781, 809	5
H	753, 773	2

Tableau 10 : Positions présentant des variations spécifiquement dans un génotype.

On peut ainsi remarquer que, même si le nombre de séquences de génotype D n'est pas le plus important, ces séquences sont celles qui présentent le plus de variations d'après ces analyses.

Afin d'approfondir l'analyse au niveau du génotype D, nous avons procédé à une analyse similaire à partir de données issues d'une étude ultra-sensible de pyroséquençage (UDPS) sur des sérums de patients infectés par des virus de génotype D. Nous avons réalisé cette analyse en collaboration avec le Docteur Christophe Rodriguez et le Professeur Jean-Michel Pawlotsky, du département de Virologie de l'hôpital Henri Mondor, à Créteil.

3.3.2.2.1.2 Etude des séquences de génotype D issues de séquençage à haut débit

La première étape de cette étude consistait à séquencer toutes les quasi-espèces du VHB provenant de 73 patients naïfs de tout traitement, chroniquement infectés par des virus de génotype D, et ayant un antigène HBe négatif.

La réaction de pyroséquençage a été réalisée avec le kit de séquençage GS FLX Titanium, sur un FLX Genome Sequencer (454 Life Sciences).

Les premiers tris et filtrations de séquences ont été réalisés par Christophe Rodriguez, à l'aide d'un pipeline de traitement des données d'UDPS qu'il a développé sous forme d'un package d'outils appelé Pyropackage ® (Rodriguez *et al*., 2012, soumis).

Près de 960000 séquences ont été générées à partir de ces échantillons, soit en moyenne 13130 ± 3220 séquences par patient. Leur longueur moyenne est de 302 pb (entre 221 et 369 pb) et 63,8% de l'ensemble des séquences a été jugé d'excellente

qualité. Après avoir éliminé les séquences non fiables ou trop courtes par l'outil PyroMute ® (Rodriguez *et al.*, 2012, soumis), plus de 900 000 séquences étaient disponibles pour l'analyse ultérieure.

Nous avons récupéré le fichier des fréquences des résidus calculées à partir des séquences filtrées et alignées, issues des quasi-espèces des 73 patients. Nous avons généré le répertoire de résidus à partir de ces fréquences.

A partir des fréquences, j'ai repéré les positions qui présentaient au moins deux résidus différents, présents dans au moins 1% des séquences. Au total, il y a 65 positions présentant des variations de résidus, dont 21 qui avaient été relevées dans les séquences de génotype D de HBVdb. Parmi ces 21 positions communes, 15 sont également retrouvées dans les autres génotypes et 6 sont visiblement spécifiques au génotype D. Dans les 65 positions déterminées comme présentant des variations dans les séquences issues de l'UDPS, 12 sont retrouvées dans les séquences HBVdb de génotypes autres que D (Figure 42). Il reste donc 32 positions présentant des variations, qui sont retrouvées exclusivement dans les séquences de génotype D issues de l'analyse UDPS.

On remarque que l'UDPS permet de détecter deux régions qui semblent être variables, la première allant des résidus 714 à 722, et la deuxième, plus longue, des résidus 735 à 753 (sauf le résidu 752).

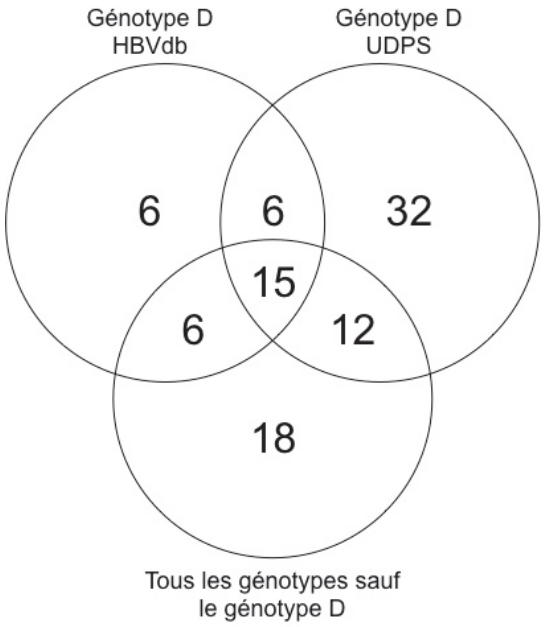


Figure 42 : Diagramme de Venn des positions de la RNase H présentant des variations pour le génotype D (données HBVdb et UDPS) et les autres génotypes.

Si on fait le bilan de cette analyse, on retrouve 44 positions spécifiques du génotype D qui présentent des variations : 32 trouvées uniquement dans les séquences issues de l'analyse UDPS, 6 trouvées dans les séquence de génotype D de HBVdb, et 6 communes aux deux (Figure 42).

Pour avoir une idée plus quantitative de la variabilité de la RNase H, nous avons procédé à une analyse similaire, mais en utilisant l'entropie de Shannon normalisée

3.3.2.2.2 Analyse quantitative de la variabilité de la RNase H par l'entropie de Shannon

A partir des alignements calculés pour les données de HBVdb et à partir des fréquences des résidus pour les données d'UDPS, nous avons calculé l'entropie de Shannon normalisée à chaque position des alignements.

Globalement, les séquences de RNase H du VHB ne sont pas variables, du fait de la pression exercée par le chevauchement de cadres de lecture. Dans cette analyse, on considère que les positions ayant une entropie supérieure ou égale à 0,1 sont variables. Partant de cette hypothèse, il y a 4 positions variables pour les séquences de génotype A, 8 pour celles de génotype B, 6 pour les C, E et F, 7 pour les D, 3 pour les H et aucune pour les séquences de génotype G.

Les valeurs d'entropie normalisée pour les 8 génotypes le long de la séquence de RNase H sont représentées sur la Figure 43.

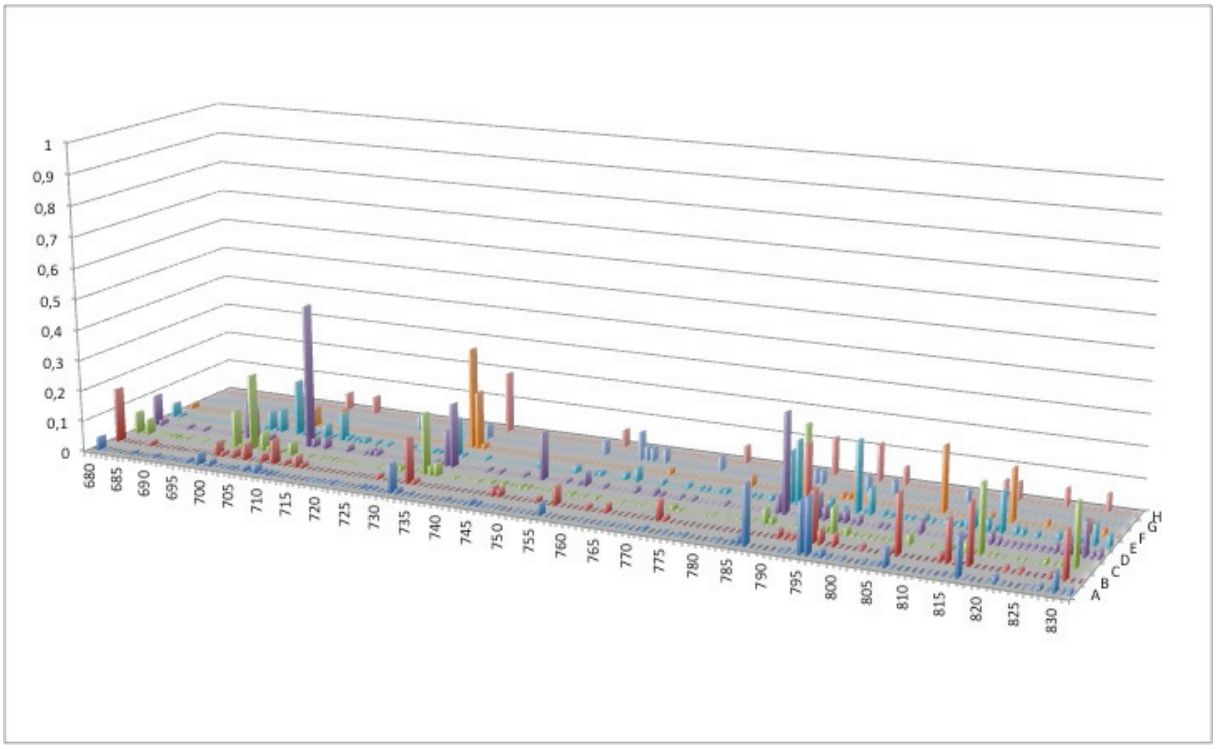


Figure 43 : Entropie de Shannon normalisée à chaque position de la RNase H et pour chaque génotype du VHB.
L'axe X correspond aux positions de la séquence, l'axe Y à l'entropie de Shannon normalisée, et l'axe Z au génotype.

Puisque nous nous sommes intéressés au génotype D, nous avons voulu déterminer les positions qui sont variables spécifiquement dans ce génotype. Pour ceci, nous avons réparti les données en 4 jeux de données : tous les génotypes (*all*), tous les génotypes sauf D (*notD*), données du génotype D à partir de HBVdb (*Dpop*), et les données du génotype D extraites de l'analyse UDPS (*Dudps*). L'analyse de la variabilité des séquences de ces 4 jeux de données est représentée sur l'histogramme 3D de la Figure 44.

Les résultats indiquent que 81,7% des positions (125/153) avaient une entropie de Shannon normalisée inférieure à 0,1. Les 28 positions restantes présentent une entropie supérieure à 0,1, dont 10 ont une entropie comprise entre 0,2 et 0,463 et sont énumérées dans l'Annexe 9. Selon ces données, la séquence de RNase H est conservée. Le jeu de données *Dudps* a permis d'identifier une région présentant des positions légèrement variables (741 à 753) avec une entropie entre 0,101 et 0,205, sauf la position 749 (détectée aussi dans le jeu de données *Dpop* avec une entropie de 0,151) et la position 752 (entropie inférieure à 0,1).

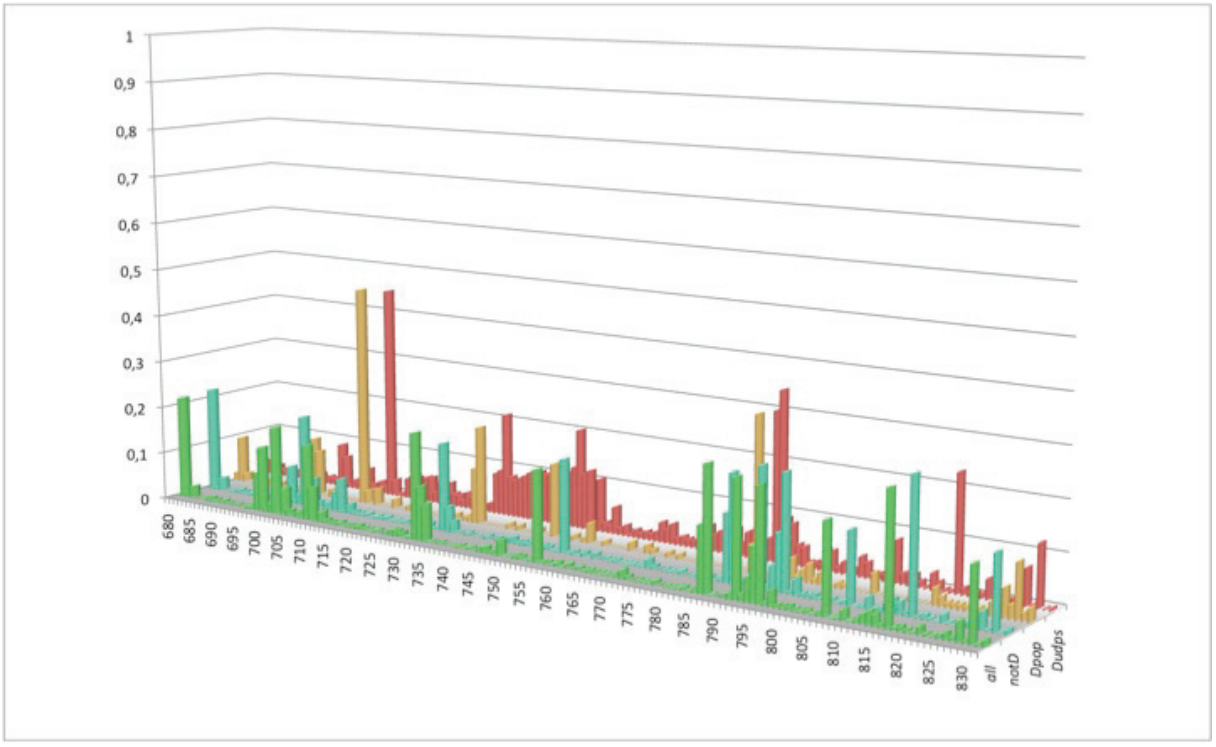


Figure 44 : Entropie de Shannon normalisée à chaque position de la RNase H et pour les 4 jeux de données.
all (vert), *notD* (cyan), *Dpop* (orange), *Dudps* (rouge).

3.3.2.2.3 La tétrade catalytique

Comme indiqué précédemment, les RNase H de type 1 ont un site actif composé d'une tétrade catalytique de type D-E-D-D.

D'après les multiples alignements de RNase H que nous avons réalisés, les 3 premiers résidus catalytiques sont D689, E718 et D737 (séquence UniProtKB : C7DSI5). Ces 3 positions n'ont pas été détectées comme étant variables par les analyses précédentes. Elles sont très conservées à travers tous les génotypes, avec une fréquence de 0,999 pour D à ces 3 positions (données HBVdb). Les fréquences sont 0,993, 0,962 et 0,928 pour les données issues de l'UDPS, même si chez 12 patients les séquences présentent la substitution D737A pour 15 à 50% des quasi-espèces virales.

En revanche, pour le quatrième D catalytique, la question se pose. Sur la séquence UniProtKB : C7DSI5, il y a quatre résidus D après le 3ème en position 737 : D777, D778, D807 et D817. Nous avons analysé ces résidus en terme de conservation. Les deux premiers, D777 et D778 sont très conservés (fréquence de 0,999 et 0,997 respectivement, pour tous les génotypes à partir de HBVdb ; 0,983 et 0,962 pour les données UDPS). Il y a donc visiblement une pression de sélection sur ces résidus. En

effet, des études ont montré que la mutation du résidu D777 affectait de manière drastique l'élongation du brin moins d'ADN (Chen and Marion, 1996).

Quant aux deux autres résidus, ils sont moins conservés, surtout le D807 qui a une fréquence de 0,172 (contre 0,796 pour V et 0,031 pour A). Le D817 a une fréquence de 0,621 (0,33 pour V, 0,034 pour A et 0,011 pour G). Ces fréquences sont calculées sur les séquences de tous les génotypes confondus (HBVdb).

Le D807 avait été désigné comme le 4ème D catalytique par Potenza *et al*. (Potenza et al., 2007), mais visiblement il est peu probable que ce soit le cas, compte tenu de la conservation de l'acide aspartique à cette position.

La Figure 45 montre les fréquences des résidus pour les positions 807 et 817, en fonction du génotype. A la position 807, pour la moitié des génotypes, le résidu le plus fréquent est un V, pour les autres génotypes c'est un D, sauf pour le génotype F pour lequel on observe la moitié de A et la moitié de D.

Concernant la position 817, on peut remarquer que pour les génotypes A et H, le résidu le plus fréquent n'est pas un D (V et A respectivement). D'après les données d'UDPS, la fréquence du D807 est de 0,938 et celle de D817 est de 0,681.

Figure 45 : Histogrammes des fréquences des résidus aux positions 807 et 817 (pour les 8 génotypes)

Nous pensons que le 4ème résidu catalytique ne peut pas être le D777, ni le D778, pour des raisons structurales et d'alignement, et ce, même s'ils sont nécessaires à la fonction. En effet, comme expliqué précédemment, chez les RNases H de type 1, le 4ème D catalytique se situe après la protrusion basique, à la suite d'une histidine conservée (Tadokoro and Kanaya, 2009). Or, nos alignements avec d'autres RNases H et nos prédictions de structures secondaires ont suggéré la présence d'une hélice C chez le VHB. L'hélice est prédite entre les résidus 751-762 (UniProtKB : C7DSI5). On remarque que quelques résidus après l'hélice C putative, la région allant de 779 à 801 est riche en arginines et peut être assimilée à une partie de la protrusion basique. Ainsi, la

protrusion basique du VHB pourrait s'étendre du résidu 751 (début de l'hélice C putative) au résidu 801 (ou avant), et comprendrait donc les D777 et D778 conservés. Ceci indiquerait qu'il est peu probable que l'un de ces deux résidus soit le 4ème catalytique.

De plus, il y a une histidine conservée en position 814 (fréquence 0,984), précédant le D817. Le D817, malgré le fait qu'il ne soit pas parfaitement conservé, semble être le meilleur candidat pour être le 4ème D catalytique de la tétrade.

Nous avons voulu approfondir l'étude et considérer la RNase H dans son contexte, la polymérase. En effet, le domaine RNase H est situé à la suite du domaine de transcriptase inverse (RT), et n'ayant pas les structures tridimensionnelles de ces domaines, on ne connaît pas vraiment leur organisation spatiale.

De plus, nous avons fait l'hypothèse de la présence d'une hélice C dans la RNase H du VHB, et il serait intéressant de voir son positionnement par rapport au reste de la protéine, et vis à vis de l'hétéro-duplex. Nous avons donc voulu construire un modèle de la polymérase le plus complet à ce jour, comprenant le domaine RT et le domaine RNase H.

De plus, il est connu que chez les rétrovirus possédant ces deux domaines, par exemple MoMLV et VIH-1, RT et RNase H sont liés par un domaine de connexion. Or, chez le VHB, ce domaine de connexion n'a pas été mentionné. La construction d'un tel modèle pour le VHB pourrait permettre de répondre à la question sur l'existence d'un domaine de connexion entre RT et RNase H.

3.3.3 Vers une modélisation moléculaire des domaines RT et RNase H du VHB

Dans cette partie, je présenterai les analyses de séquences préliminaires que j'ai effectuées pour la modélisation par homologie des domaines RT et RNase H de la polymérase du VHB, avec notamment la recherche d'homologues de structures connues pour la sélection de l'empreinte.

3.3.3.1 Alignement multiple de polymérases du VHB

Nous avons réalisé une étude préliminaire sur les polymérases des différents virus de l'hépatite B, notamment les virus infectant le canard, la marmotte, l'écureuil et

certains primates. Pour cela, nous avons extrait 66 séquences de polymérases de la bases de données UniProtKB : SwissProt (realease 2011_04). Nous avons ensuite effectué un alignement multiple de ces séquences à l'aide de Clustal W (version 1.8). Cet alignement est présenté dans l'Annexe 10.

Cet alignement révèle des insertions et des délétions dans les séquences des virus aviaires (*avihepadnavirus*), dans la partie TP (terminal protein) de la polymérase, par rapport aux séquences de virus de mammifères (*orthohepadnavirus*). La partie charnière (ou spacer), n'est pas très conservée au sein des *hepadnaviridae*. Le domaine RT, quant à lui, est plutôt bien conservé, mais on peut observer une longue insertion (48 aa) chez les *orthohepadnavirus* par rapport aux séquences de virus aviaires. Cette insertion se situe entre les résidus rt113 et rt160, selon la numérotation standard du domaine RT (Stuyver et al., 2001). Enfin, les virus aviaires présentent une insertion de 3 résidus, juste avant le début du domaine RNase H, ainsi qu'une délétion d'une douzaine de résidus au niveau de la séquence qui suit l'hélice C putative, et d'une dizaine à l'extrémité C-terminale du domaine. La Figure 46 représente un extrait tronqué de l'alignement des 66 séquences de polymérases, sur lequel on peut visualiser l'insertion dans les séquences de virus de mammifères, et la délétion correspondante dans les séquences des virus aviaires.

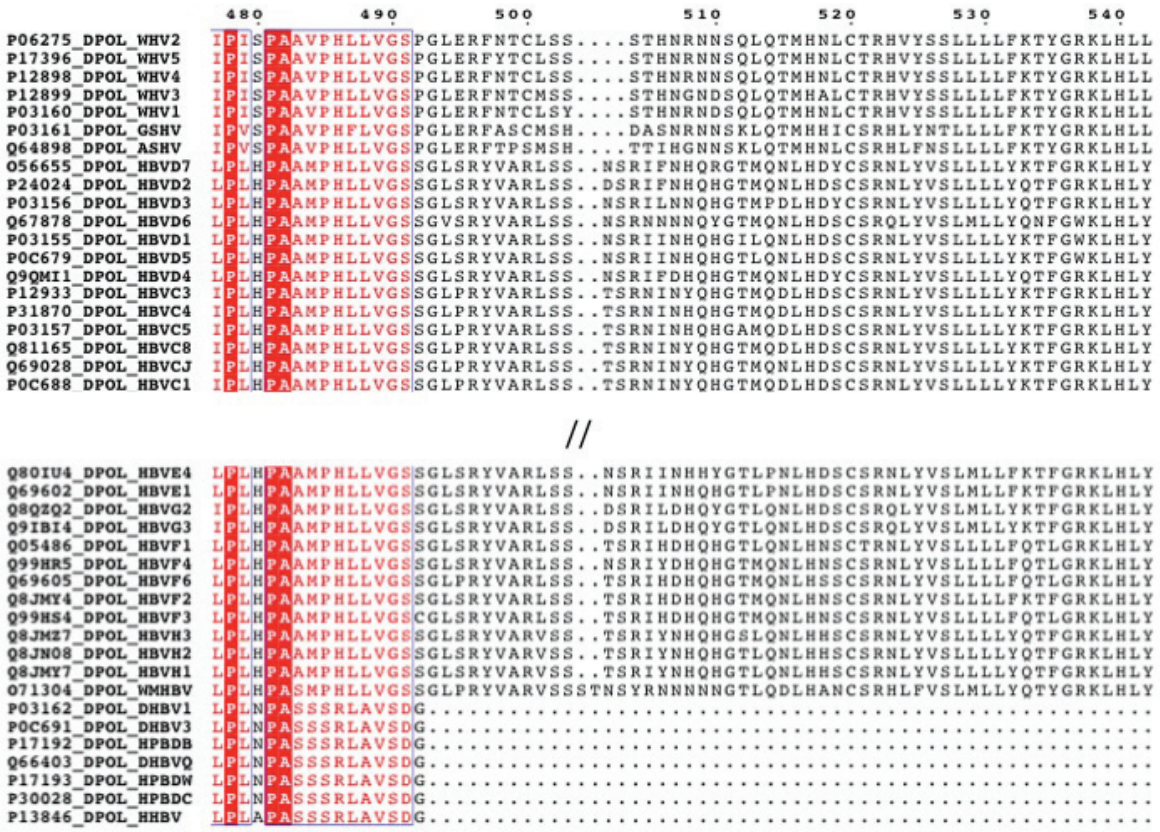


Figure 46 : Alignement montrant l'insertion dans les séquences des orthohepadnavirus.

3.3.3.2 Recherche d'homologues pour la sélection d'une empreinte

Les deux domaines enzymatiques de la polymérase du VHB partagent des similitudes fonctionnelles avec des domaines RT et RNases H d'organismes de règnes différents. Afin de confirmer ces similitudes au niveau de la séquence, nous avons effectué des recherches par profils (PSI-BLAST et HMMER).

Dans un premier temps, nous avons fait une recherche par PSI-BLAST contre la base de données *nr* (non-redundant protein sequences) qui contient des entrées de UniProtKB et de la PDB, avec une séquence complète de polymérase du VHB, pour trouver des séquences homologues de structures connues. Cette recherche n'ayant pas aboutie, nous avons effectué la même recherche, mais en utilisant qu'une partie de la polymérase comprenant le domaine RT, allant des résidus 300 à 700. Après 2 itérations de PSI-BLAST, des séquences de la PDB sont apparues dans les résultats. En effet, les résultats comptaient des RT du MoMLV, du VIH-1 et du VIH-2.

A l'inverse, nous avons essayé de détecter des séquences de Pol du VHB en utilisant des profils HMM construits à partir d'alignements structuraux. Les recherches ont été initiées par des alignements de structures de transcriptases inverses résolues expérimentalement, stockées dans la PDB. Nous avons effectué plusieurs alignements structuraux en combinant les structures du VIH-1 (PDB: 1T05) (Tuske et al., 2004) et du MoMLV (PDB : 1RW3) (Das and Georgiadis, 2004). Les domaines RT ont une organisation structurale en « main droite » avec des sous-domaines correspondant au pouce, à la paume et aux doigts.

Après avoir isolé pour les deux structures, la partie correspondant aux sous-domaines de la paume et des doigts, nous avons produit un alignement structural de ces régions. Par ailleurs, nous avons aligné les deux pouces (MoMLV et VIH-1). Enfin, nous avons combiné ces deux alignements pour créer un alignement structural optimisé des domaines RT du VIH-1 et du MoMLV, que nous avons utilisé pour construire un profil HMM. Nous avons cherché dans UniProtKB des séquences s'alignant avec ce profil. Après 5 itérations avec enrichissement du profil, nous avons détecté des séquences de Pol du VHB dans les résultats. Ces analyses montrent que les séquences de RT du VHB, VIH-1 et MoMLV sont des homologues distants.

Les recherches d'homologues pour une empreinte du domaine RNase H ont été décrites dans la partie précédente, et elles n'ont pas permis d'identifier de RNase H de structure connue dont la séquence est homologue à la RNase H du VHB. C'est la structure de la RNase H de *E. coli* qui avait été utilisée, car elle possède une hélice C.

Nous avons donc décidé de choisir comme empreinte, une structure de l'alignement structural optimisé de VIH-1 et MoMLV utilisé pour la recherche par profil, présentant les deux domaines, RT et RNase H. Les structures disponibles pour le MoMLV ne présentant pas les deux domaines, mais l'un ou l'autre des domaines, notre choix s'est porté sur une structure de RT-RNase H du VIH-1 (PDB : 1T05). Le domaine RT, organisé en main droite, est relié au domaine RNase H par un domaine de connexion.

Plusieurs modèles moléculaires du domaine RT du VHB ont déjà été publiés, et les auteurs ont utilisé la structure du domaine RT du VIH-1 comme empreinte pour la modélisation par homologie. Cependant, la différence entre les modèles réside dans l'alignement entre la séquence de l'empreinte et la séquence à modéliser. Parmi les

modèles publiés, on peut distinguer 3 alignements différents entre la RT du VIH-1 et celle du VHB (Das et al., 2001; Daga et al., 2010; Bartholomeusz et al., 2004).

L'alignement que nous avons produit est encore différent de ces 3 alignements, et surtout, il comprend les deux domaines RT et RNase H. Il sera présenté dans le prochain chapitre (4.3.2.1).

Chapitre 4 :

Modélisation moléculaire des domaines RT et RNase H de la polymérase du VHB

Ce chapitre, structuré sous forme d'article, présente les travaux de modélisation moléculaire des domaines RT et RNase H du VHB. Ces modélisations par homologie s'appuient sur les analyses de séquences du VHB décrites au chapitre 3.

4.1 Introduction

La cible principale des traitements anti-VHB est la polymérase virale, et plus précisément, l'activité enzymatique de transcriptase inverse (RT) qu'elle porte. Les inhibiteurs de cette activité, sont les analogues de nucléos(t)ides (NA). Certains de ces NA ont été testés dans le traitement contre le VHB parce qu'ils avaient montré une efficacité contre le VIH (Virus de l'Immunodéficience Humaine), qui utilise également une RT pour se répliquer, puisque c'est un rétrovirus.

Le traitement de l'hépatite B chronique avec des inhibiteurs de RT induit une diminution rapide de la virémie. Toutefois, des traitements à long terme avec des médicaments de première génération tels que la lamivudine ou l'adéfovir, avec une faible barrière génétique à la résistance, ont été associés à des rebonds de charge virale du fait de la sélection de variants du VHB résistants aux médicaments (Zoulim and Locarnini, 2009; Locarnini, 2008). Les médicaments anti VHB de seconde génération aujourd'hui utilisés comme première ligne, l'entécavir et le ténofovir, sont puissants et ont une plus grande barrière à la résistance. Ainsi, la sélection de souches résistantes à VHB est rare au cours des 5 premières années de traitement. Cependant, la résistance peut se produire plus tard dans le cours de la thérapie chez les patients susceptibles d'utiliser ces médicaments à vie.

Cependant, les mécanismes moléculaires qui interviennent dans ces résistances sont mal connus, du fait qu'on ne dispose pas d'une structure tridimensionnelle de la polymérase du VHB résolue expérimentalement.

Pour pallier ce manque, plusieurs auteurs ont tenté de modéliser le domaine RT de la polymérase, par modélisation par homologie, en utilisant la structure de la RT du VIH-1 comme empreinte.

En 2001, Das *et al.* (Das et al., 2001) ont publié le premier modèle 3D de la RT du VHB, avec l'alignement des séquences de RT du VIH-1 et du VHB, utilisé pour la modélisation. Par la suite, d'autres auteurs ont utilisé ce modèle, ou l'ont reconstruit à partir de l'alignement, pour effectuer des dynamiques moléculaires et des analyses de

docking des interactions entre les NA et le domaine RT du VHB (Chong and Chu, 2002; Yadav and Chu, 2004; Langley et al., 2007; Sharon and Chu, 2008).

Bartholomeusz *et al.* (Bartholomeusz et al., 2004) ont publié un modèle de RT en 2004, avec un alignement différent de celui de Das *et al.*, mais la publication ne présente qu'une partie de l'alignement.

En 2010, Daga *et al.* (Daga et al., 2010) ont publié un nouveau modèle avec un nouvel alignement des séquences de RT du VHB et du VIH-1. Un autre modèle de la RT du VHB a été publié, mais il ne présente pas l'alignement utilisé pour la modélisation (Mukaide et al., 2010).

Dans le contexte de l'émergence de résistances aux inhibiteurs de RT, il est plus sûr de mettre au point de nouvelles générations de médicaments anti-VHB efficaces, et ayant d'autres cibles que la RT.

Le domaine RNase H de la polymérase du VHB représente une cible thérapeutique potentielle, car il a une activité enzymatique qui est essentielle pour la réplication du VHB. Cependant, la découverte de médicaments ciblés nécessite une connaissance approfondie de la structure de la RNase H. Or, tout comme pour les autres domaines de la polymérase, les tentatives visant à déterminer expérimentalement la structure 3D de la RNase H ont été vaines jusqu'à présent. Cependant, plusieurs structures de RNase H de divers organismes ont été résolues expérimentalement (Katayanagi et al., 1993; Lim et al., 2006; Tadokoro and Kanaya, 2009).

Une caractéristique structurale importante des RNases H, est l'existence d'une hélice α, appelé hélice C, impliquée dans la reconnaissance du substrat (Champoux and Schultz, 2009). Elle est présente chez certains organismes (par exemple *E. coli*, Moloney Murine Leukemia Virus), mais pas chez d'autres (par exemple le VIH-1).

La disponibilité de ces structures 3D et le fait que les structures de RNase H sont conservées à travers les espèces (Tadokoro and Kanaya, 2009), ouvre la voie de la modélisation moléculaire de ce domaine, basée sur l'alignement de la séquence du VHB avec la séquence de l'empreinte sélectionnée. D'ailleurs, un modèle moléculaire de la RNase H du VHB utilisant la RNase H de *E. coli*, a été publié en 2007 (Potenza et al., 2007).

Nous avons construit un nouveau modèle par homologie de la RNase H du VHB, différent de celui déjà publié, et qui prend en compte l'existence éventuelle d'une hélice C dans la RNase H du VHB.

Dans ce chapitre, les méthodes et outils utilisés pour la modélisation et les analyses sur les modèles seront décrits dans un premier temps.

Ensuite, je développerai la modélisation de la RNase H ainsi que l'étude réalisée sur le modèle, qui vise à cartographier les positions variables décrites lors des analyses de séquences, dans le chapitre précédant.

Enfin, dans la dernière partie, je décrirai le travail de modélisation effectué pour construire le modèle de la polymérase le plus complet à ce jour, comprenant les deux domaines RT et RNase H. Je détaillerai ensuite les analyses faites sur le site catalytique de la RT, et les comparaisons avec les autres modèles de RT publiés.

4.2 Matériel et méthodes

Les analyses préliminaires à la modélisation par homologie, ainsi que la recherche d'empreinte ont été décrites dans le chapitre précédent. Ce chapitre présente donc les étapes suivantes : l'alignement entre la séquence à modéliser et la séquence de l'empreinte, la modélisation, et les analyses réalisées à partir des modèles.

4.2.1 Alignements et optimisations

Les alignements de séquences ont été réalisés avec Clustal W (Thompson et al., 1994), et ils ont été optimisés manuellement.

Ces optimisations ont été notamment guidées par des prédictions de structures secondaires, réalisées sur le serveur NPS@ (Combet et al., 2000), en combinant les méthodes PHD (Rost and Sander, 1993), DSC (King and Sternberg, 1996) et SOPMA (Geourjon and Deléage, 1995). Elles ont également été guidées par l'observation des structures tridimensionnelles ainsi que par les structures secondaires déduites de la structure 3D par l'algorithme DSSP (Kabsch and Sander, 1983).

4.2.2 Modélisation moléculaire par homologie

Les modèles moléculaires ont été calculés avec le logiciel Geno3D (Combet et al., 2002), qui peut prendre en entrée l'alignement de la séquence à modéliser avec la séquence de l'empreinte, et le fichier PDB de la structure de l'empreinte. Geno3D calcule les modèles selon un protocole similaire à celui utilisé pour la résolution de structure par résonance magnétique nucléaire (RMN). Il utilise des contraintes géométriques

d'angles dièdres et de distances, la géométrie des distances, ainsi que la dynamique moléculaire et des algorithmes de minimisation de l'énergie, comme implémenté dans le logiciel CNS (version 1.2). Geno3D mesure les contraintes spatiales (angles dièdres et distances interatomiques) à partir de la structure de l'empreinte pour les résidus similaires en fonction de l'alignement.

L'avantage de ce protocole est qu'il fournit un faisceau de modèles, qui échantillonne l'espace conformationnel, en particulier au niveau des boucles. La valeur du RMSD (écart quadratique moyen) est faible au niveau des éléments de structure secondaire et augmente au niveau des boucles en présence d'insertions ou de délétions. Ces dernières sont traitées d'une manière automatique dans ce protocole par l'utilisation de contraintes géométriques, mais la qualité du modèle obtenu dépend grandement de l'alignement et surtout de la position des gaps. Ainsi, si les gaps sont localisés au niveau des boucles, le modèle sera plus fiable que si les gaps sont localisés dans les structures secondaires. Pour chaque modélisation, 10 modèles sont calculés, et pour chacun d'eux, la stéréochimie est évaluée par le logiciel PROCHECK (Laskowski et al., 1996), et les rapports et graphiques sont fournis par Geno3D. Ensuite, le modèle le plus correcte est sélectionné en fonction de la valeur d'énergie, de la stéréochimie, du nombre de violations de contraintes (non respect d'une contrainte d'intervalle de distance ou d'angle), et du RMSD (déviation standard, root mean square deviation) avec l'empreinte, pour des analyses plus approfondies.

4.2.3 Superpositions structurales

La plupart des superpositions entre les modèles et les structures sont effectuées par Geno3D, les autres ont été réalisées par DeepView (Guex and Peitsch, 1997).

4.2.4 Cartographie des positions variables sur le modèle

Les positions variables déterminées par les analyses des jeux de données issus de HBVdb et de l'étude de pyroséquençage (UDPS) ont été cartographiées sur le modèle moléculaire RNase H dans le logiciel Pymol, grâce à un script permettant de colorer les positons en fonction de la valeur d'entropie de Shannon normalisée. J'ai développé une méthode Java qui prend en entrée le fichier contenant les valeurs d'entropie de Shannon

normalisé par position et le fichier PDB du modèle, et qui crée le script Pymol permettant de colorer les positions variables selon un gradient.

Les couleurs utilisées selon les valeurs d'entropie sont :

- gris clair : 0,0 ≤ ESN < 0,1
- jaune : 0,1 ≤ ESN < 0,2
- orange clair : 0,2 ≤ ESN < 0,3
- orange : 0,3 ≤ ESN < 0,4
- rouge : 0,4 ≤ ESN < 0,5
- framboise : 0,5 ≤ ESN < 0,6
- brique : 0,6 ≤ ESN < 0,7
- brun : 0,7 ≤ ESN < 0,8
- chocolat : 0,8 ≤ ESN < 0,9
- noir : 0,9 ≤ ESN ≤ 1,0

4.2.5 Analyses d'interactions avec un ligand

Les analyses d'interactions entre les ligands et les structures ou modèles ont été effectuées par le logiciel LIGPLOT, qui est disponible sur le site web de PDBsum.

Toutes les images de structures sont réalisées avec Pymol.

4.3 Résultats

4.3.1 La modélisation du domaine RNase H

Les recherches d'empreintes pour modéliser le domaine RNase H ont été présentées dans le chapitre précédent. La RNase H de *E. coli* qui a été sélectionnée, et l'étape suivante correspond à l'alignement de la séquence de l'empreinte avec la séquence à modéliser.

4.3.1.1 L'alignement des RNases H du VHB et d'E. coli

La séquence du VHB utilisée pour la modélisation est la séquence UniProtKB : C7DSI5 (génotype D) à partir du résidu 680 (motif RSGL), et l'empreinte correspondant à la structure de RNase H de *E. coli* est l'entrée PDB : 1RDD (Katayanagi et al., 1993).

Un alignement de la séquence de RNase H du VHB avec celle de *E. coli* est le résultat d'une combinaison de deux alignements (avec Clustal W, version 1.8):

- l'alignement publié par Lim *et al.* (Lim et al., 2006) des RNase H du VIH-1, du MoMLV, de *E. coli*, et de *B. halodurans*
- l'alignement par paire des RNases H de *E. coli* et du VHB publié par Potenza *et al.* (Potenza et al., 2007).

Dans l'alignement résultant de cette combinaison, seulement les séquences du VHB et de *E. coli* ont été conservées pour garder un alignement par paire. Puis, cet alignement a été optimisé manuellement en observant la structure 3D de *E. coli*, les structures secondaires déduites de la structure 3D (algorithme DSSP) ainsi que les prédictions de structures secondaires pour la séquence du VHB. L'alignement est présenté sur la Figure 47.

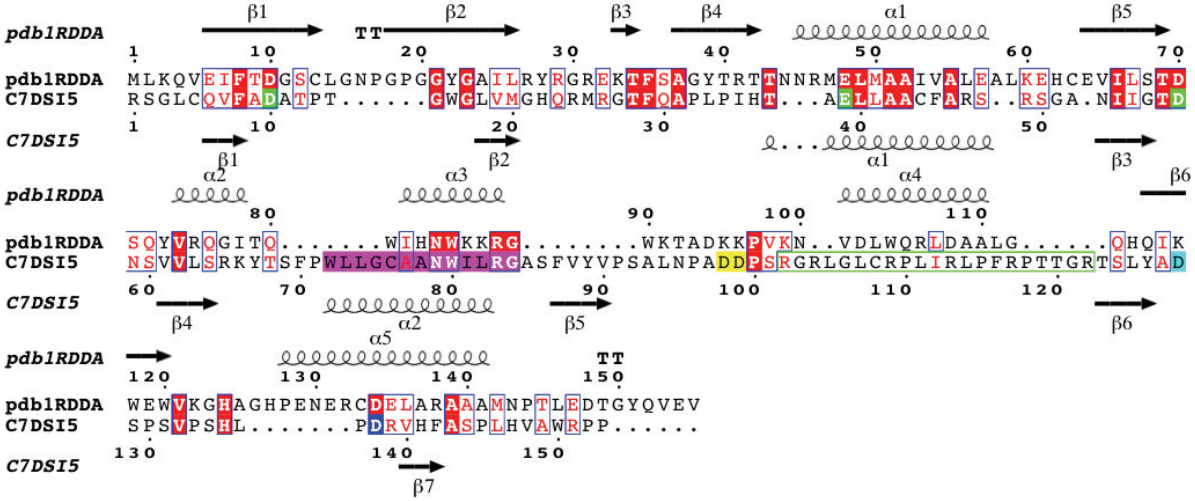


Figure 47 : Alignement des séquences de RNases H de *E. coli* et du VHB.
Les structures secondaires déduites et prédites sont représentées. L'hélice C est surlignée en rose, les 3 premiers résidus catalytiques en vert : D10, E39, D58 (689, 718, 737 dans la séquence UniProtKB : C7DSI5), D98 et D99 en jaune (D777 et D778), D128 (D807) en cyan, et D138 en bleu (D717). La région riche en arginines est encadrée en vert. Figure réalisée avec ESPript (Gouet et al., 2003).

Lorsque l'on analyse l'alignement des séquences de RNase H du VHB et de *E. coli,* on ne trouve que 14,6% d'identité, 39,9% de similarité et 26,9% de gaps.

On suggère que la région de 24 résidus F748-D771 (UniProtKB : C7DSI5) comprend l'hélice C présente dans les RNases H de MoMLV et de *E. coli*. D'après l'alignement et les prédictions de structures secondaires, cette hélice s'étend du résidu W751 au résidu

R762. Au niveau des séquences, les hélices C de *E. coli* et du VHB partagent un motif conservé (758-NWXXRG-763).

Il est connu que l'hélice C, faisant partie de la protrusion basique, comprend habituellement un cluster de résidus chargés positivement, qui permettent notamment d'interagir avec l'acide nucléique. Dans le cas du VHB, l'hélice C putative ne comporte qu'une arginine à la position R762. Cependant, une région située quelques résidus après l'hélice C putative (UniProtKB:C7DSI5:779-801) est riche en arginines, ce qui suggère que la RNase H du VHB possède bien une protrusion basique.

Les RNases H de type 1 possèdent un site catalytique composé de 4 résidus : D-E-D-D. Dans l'alignement, les 3 premiers résidus catalytiques de *E. coli* sont alignés avec les positions D689, E718 et D737. Il a été montré dans le chapitre précédant que ces 3 résidus de la RNase H sont très conservés, pour tous les génotypes du VHB. En revanche, la question se pose pour le quatrième D du site catalytique. Potenza *et al*. (Potenza et al., 2007) avait désigné le résidu D807 comme étant le 4e résidu catalytique. Or, d'après les analyses de séquences décrites dans le chapitre précédent, cette position est variable et on ne retrouve pas toujours un D, en fonction du génotype.

Les hypothèses du chapitre précédent (paragraphe 3.3.2.2.3) quant à la position du 4ème D, et l'optimisation manuelle de l'alignement, guidée par les structures secondaires de l'empreinte et par celles prédites pour la RNase H du VHB, nous ont amené à aligner le 4ème D catalytique de *E. coli* avec le résidu D817. En effet, les analyses de séquences ont montré que ce résidu est globalement mieux conservé que le résidu D807, même si pour certains génotypes, le résidu le plus fréquent à cette position est un V (génotype A) ou un A (génotype H). Il reste donc une incertitude quant à la position et à l'importance de ce quatrième résidu catalytique chez le VHB.

4.3.1.2 La modélisation moléculaire du domaine RNase H du VHB

Le calcul de modélisation a été lancé en fournissant à Geno3D l'alignement entre les RNases H de *E. coli* et du VHB, et le fichier PDB de 1RDD, contenant toutes les coordonnées atomiques de la structure de la RNase H de *E. coli*.

Geno3D a produit 10 modèles, dont la qualité a été vérifiée par PROCHECK (Laskowski et al., 1996). L'accord entre les modèles était bon, avec une RMSD de 1,50 Å sur 81 des atomes de Cα.

Le modèle le plus correct a ensuite été sélectionné parmi les 10, selon plusieurs critères :

- une valeur énergétique la plus basse possible
- un nombre de violations de contraintes le plus bas possible
- une RMSD faible avec l'empreinte
- un nombre élevé de résidus dans des régions favorables (stéréochimie)

Le Tableau 11 présente les valeurs des critères pour les 10 modèles générés.

N° du modèle	Valeur énergie (kcal/mol)	RMSD avec l'empreinte (Å)	Nombre de violations de contraintes	Résidus dans la région favorable (%)
1	-4 329,40	1,66	7	81,5
2	-4 508,93	1,63	7	85,5
3	-4 509,07	1,61	7	86,2
4	-4 456,77	1,62	7	83,9
5	-4 647,11	1,6	6	79,8
6	-4 470,33	1,73	6	83,1
7	-4 507,87	1,55	6	**86,3**
8*	**-4 730,94**	**1,29**	**5**	**86,3**
9	-4 601,71	1,61	7	79,8
10	-4 484,14	1,5	6	78,3

Tableau 11 : Valeurs des critères de sélection pour les 10 modèles de RNase H générés.
(Gras : meilleure valeur, vert* : modèle sélectionné.)

Le modèle 8, qui a été sélectionné selon ces critères, est celui utilisé pour les analyses, telles que la cartographie des positions variables déterminées par l'étude de pyroséquençage et à partir des données de HBVdb. La valeur de RMSD du modèle 8 avec la structure PDB : 1RDD est de 1,29Å (calculée sur 81 atomes Cα).

Bien que cette valeur indique un bon accord entre le modèle et la structure de l'empreinte, il y a des différences locales dans les environs proches de l'hélice C putative, du fait que l'insertion qui comprend l'hélice C putative est plus longue de l 'hélice C de *E. coli*.

4.3.1.3 Superposition du modèle avec l'empreinte

La superposition de la structure de l'empreinte et du modèle sélectionné est présentée dans la Figure 48.

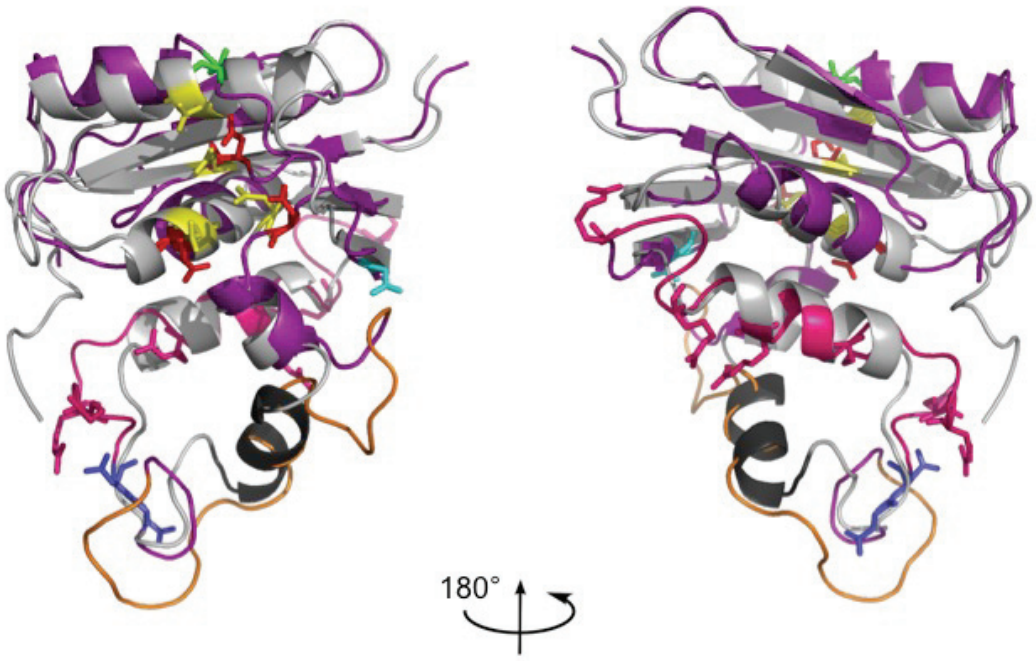


Figure 48 : Superposition de la structure de la RNase H de *E. coli* et du modèle VHB sélectionné
E. coli : gris-noir, VHB : violet-rose-orange.

La structure de *E. coli* est présentée en gris et le modèle de la RNase H du VHB en violet. On peut voir que dans la partie supérieure du modèle, des brins β et des hélices α ont été modélisés et ils superposent bien à ceux de la structure de *E. coli*, même si les brins modélisés sont plus courts que les brins de la structure. Dans la partie basse de la figure, on peut observer l'hélice C de *E. coli*, qui est représentée en noir. La partie contenant l'hélice C putative du VHB est représentée en orange et est évidemment bien plus longue que l'hélice C de *E. coli*. Malgré cela, la partie du modèle correspondant aux résidus alignés avec l'hélice C de *E. coli* suit parfaitement les tours de l'hélice C de *E. coli*. Les autres résidus de cette région sont repliés en boucle, puisqu'il n'y a pas de résidus leur correspondant dans l'empreinte.

La partie en rose représente toute la portion du modèle qui correspond à la région riche en arginines située directement en C-terminal de la région comprenant l'hélice C putative. On peut remarquer qu'une partie de ces résidus sont repliés en hélice α, puisqu'ils sont alignés avec l'hélice D de *E. coli*. Les arginines de cette région sont représentées en sticks pour que l'on puisse visualiser l'orientation des chaînes latérales. On peut voir qu'elles sont quasiment toutes dirigées vers l'extérieur du modèle, suggérant une interaction avec l'acide nucléique.

D'autres résidus sont représentés sous forme de sticks, pour la visualisation de leurs chaînes latérales, ce sont les résidus catalytiques sûrs et potentiels. La tétrade catalytique de *E. coli* est colorée en jaune, quant aux 3 premiers résidus catalytiques du VHB, ils sont présentés en rouge. Les résidus D777 et D778, très bien conservés d'après les analyses de séquences, sont situés en C-terminal de l'hélice C, et sont présentés en bleu. Ils sont à l'opposé du site catalytique. Le résidu D807, suggéré comme le 4ème catalytique par Potenza *et al*., est représenté en cyan, et est assez éloigné du site catalytique, du fait de sa position dans notre alignement.

Le résidu D817, que nous suggérons comme étant le 4ème résidu catalytique est coloré en vert et se trouve dans l'hélice de la partie supérieure.

4.3.1.4 Cartographie des positions variables de la RNase H sur le modèle

La cartographie des positions ayant une entropie de Shannon normalisée supérieure ou égale à 0,1 (jaune et orange) sur le modèle moléculaire (Figure 49) indique qu'elles sont pour la plupart situées sur la face de la molécule qui n'est pas en contact avec le substrat acide nucléique. La position 817 qui pourrait porter le quatrième résidu catalytique (D) est une exception à cette règle (D817 en bleu). Les positions identifiées uniquement par le jeu de données *Dudps* (jaune) semble se situer entre la protrusion basique (cyan avec les résidus arginine en violet clair) et le site catalytique (les trois premiers résidus en vert, le D807 en magenta et le D817 en bleu), mais sont pour la plupart sur le côté (image en haut à gauche) et l'arrière (image en bas à gauche) de la molécule vis à vis des résidus catalytiques. Cependant, d'après la faible résolution de notre modèle (en raison de la faible similarité de séquence entre la requête et l'empreinte), surtout dans la région qui contient des insertions (peu de contraintes pour modéliser cette région), il est difficile de déterminer les résidus qui pourraient jouer un rôle dans la fixation du substrat.

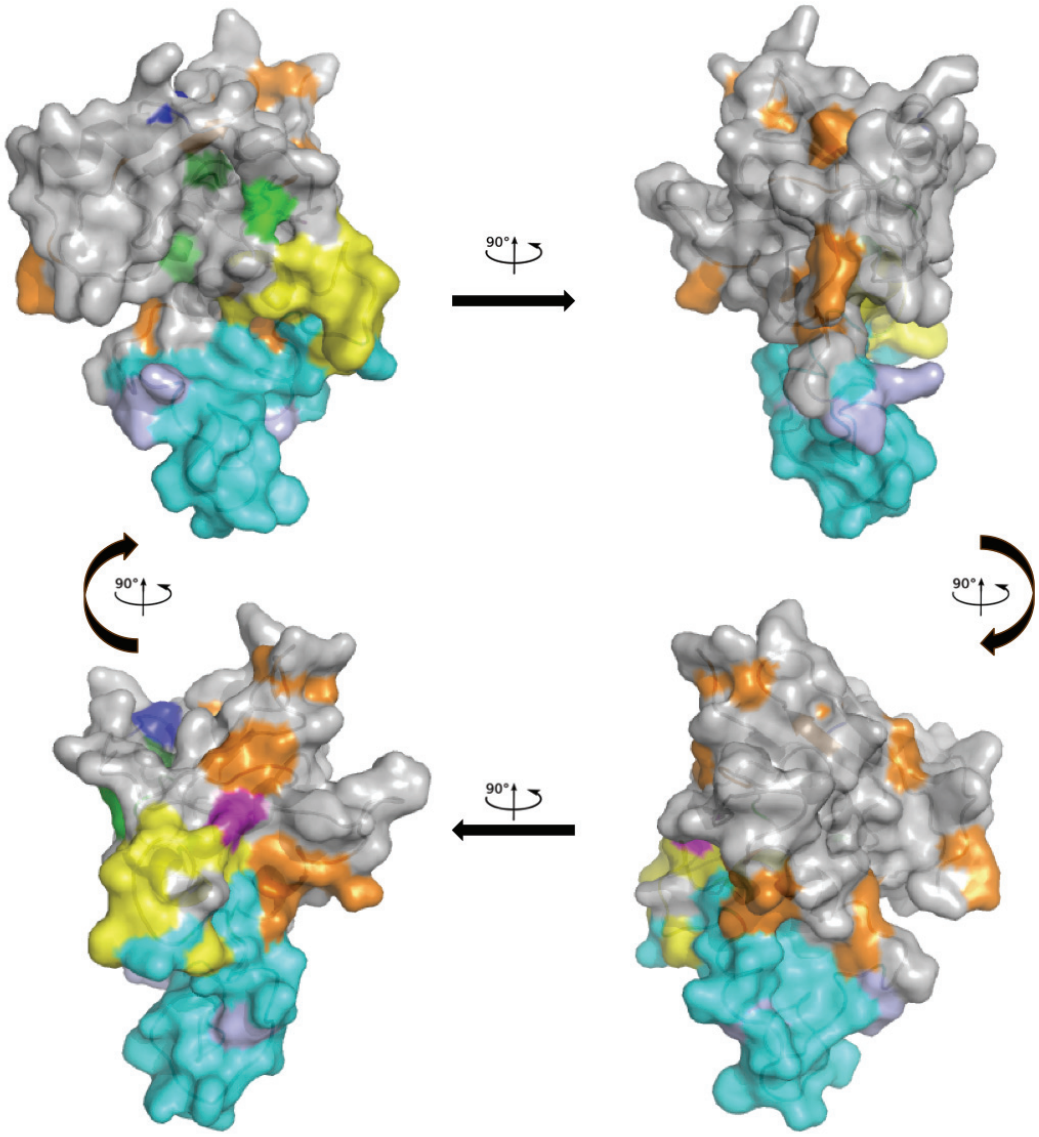


Figure 49 : Cartographie des positions variables sur le modèle de RNase H du VHB.

Les figures où sont cartographiées les positions variables (entropie ≥ 0,1) pour chacun des 4 jeux de données : *all*, *notD*, *Dpop* et *Dudps*, sont présentées dans l'Annexe 11, avec le gradient de couleurs décrit dans la partie matériel et méthodes.

4.3.2 La modélisation des domaines RT et RNase H

Afin de positionner notre modèle du domaine RNase H dans son contexte, et de vérifier le positionnement de l'hélice C putative, nous avons décidé de créer un modèle plus complet que tous ceux publiés jusqu'à présent, qui comprendrait les deux domaines enzymatiques principaux de la polymérase : le domaine RT (transcriptase inverse) et le

domaine RNase H. Un tel modèle peu permettre d'avoir une idée sur la structure globale de la polymérase, et sur l'organisation de ces deux domaines dans l'espace.

Les analyses de séquences préliminaires et la recherche d'empreinte pour la modélisation ont été présentées au chapitre précédent (3.3.3). Cette partie débute donc par l'étape d'alignement des séquences de l'empreinte et du VHB, pour se poursuivre par la modélisation et les analyses faites sur le modèle.

Elle présentera également une comparaison entre le domaine RT de notre modèle et des modèles de RT déjà publiés, ainsi qu'une analyse plus approfondie du site actif du domaine RT de notre modèle.

4.3.2.1 L'alignement des domaines RT et RNase H du VHB et du VIH-1

La séquence de polymérase utilisée pour la modélisation est la séquence UniProtKB:Swiss-Prot:Q05486. C'est une séquence issue d'un génome de génotype F. Pour modéliser les domaines RT et RNase H, seulement la séquence de 497 résidus allant des positions 347 à 843 a été utilisée.

Nous avons d'abord calculé un alignement entre la séquence du VHB décrite ci-dessus et la séquence de la sous-unité p66 du VIH-1 (PDB : 1T05) à l'aide du programme Clustal W. Dans le domaine RT, cet alignement a ensuite été optimisé manuellement en fonction d'un alignement profil-profil qui a été calculé par programme MUSCLE (version 3.8.31) à partir de deux alignements :

- l'alignement multiple des 66 polymérases de virus de l'hépatite B infectant différents organismes aviaires et mammifères, contenant la séquence complète de UniProtKB:Swiss-Prot:Q05486
- l'alignement structural des domaines RT de MoMLV et VIH-1

Pour le domaine de la RNase H, il s'agissait maintenant d'aligner la séquence du VHB avec celle du VIH-1, et plus avec celle de *E. coli*.

Nous avons donc utilisé la combinaison de l'alignement de structures de Lim *et al*. (Lim et al., 2006) et l'alignement de Potenza *et al.* (Potenza et al., 2007), à laquelle nous avons ajouté l'alignement multiple des 66 polymérases de divers virus d'hépatite B. L'alignement par paire VHB - VIH-1 dans le domaine de la RNase H a été optimisé selon l'alignement résultant de cette combinaison d'alignements.

Une optimisation manuelle finale de l'alignement par paire a été réalisée en particulier dans le sous-domaine correspondant au pouce de la RT ainsi que dans le domaine de connexion entre RT et RNase H (à partir du motif conservé 596-598 MGY au motif 691-694 RPGL) pour fournir l'alignement final utilisé pour construire le modèle. Les optimisations manuelles ont été guidées par les structures secondaires observées, déduites de la structure tridimensionnelle par l'algorithme DSSP (Kabsch and Sander, 1983), et prédites par une combinaison de méthodes (PHD, DSC, SOPMA).

L'alignement utilisé pour construire le modèle est présenté à la Figure 50. Le pourcentage d'identité est de 12,7, le pourcentage de similarité est de 34,6 et le pourcentage de gaps est de 38,4.

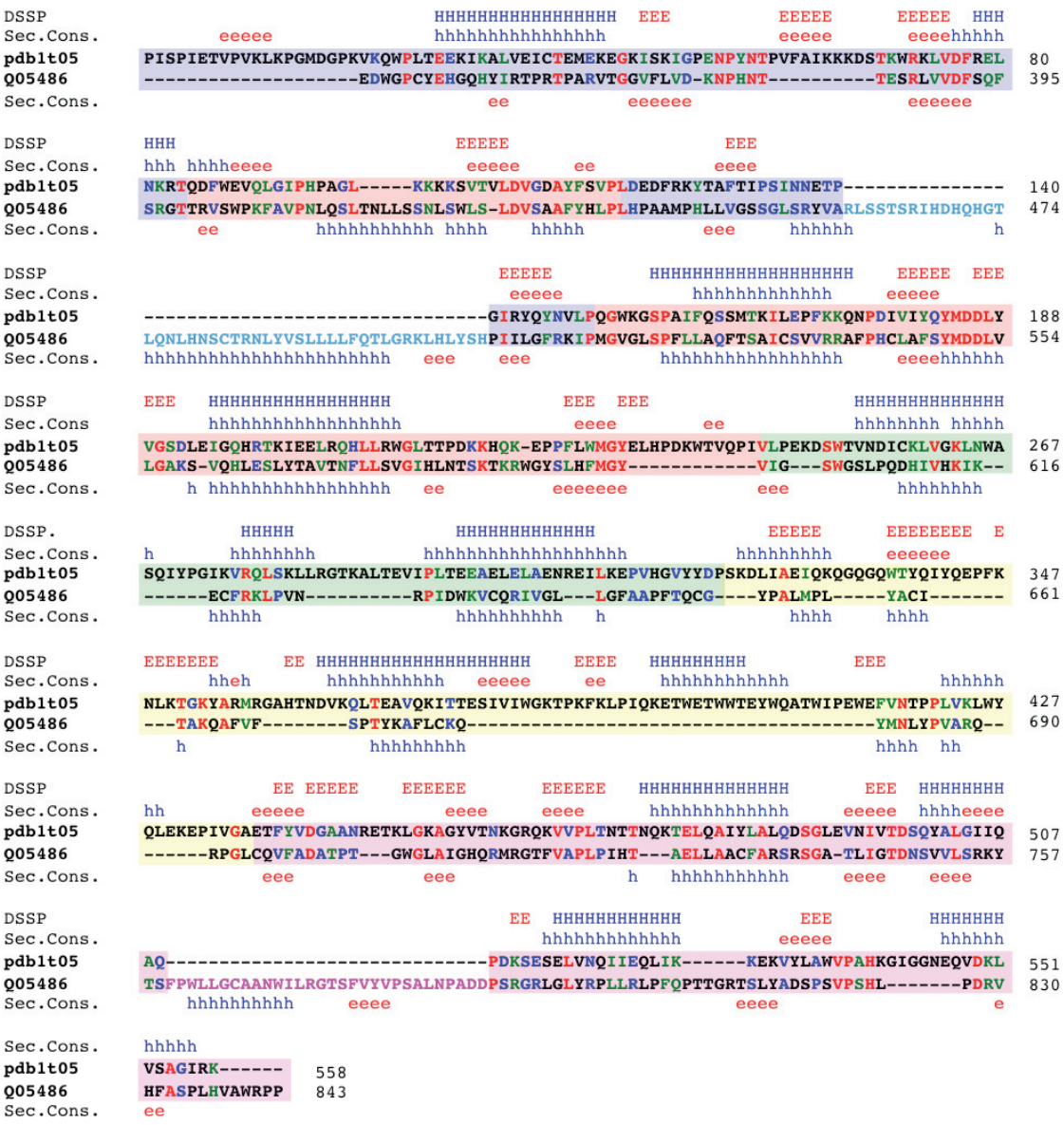


Figure 50 : Alignement de la séquence du VHB à modéliser avec la séquence du VIH-1 (PDB : 1T05), prédictions de structures secondaires (lignes Sec. Cons.), et structures secondaires déduites de la structure 3D (H et h=hélices, E et e=brins).

L'alignement montre la longue insertion de 47 résidus (en bleu clair) entre les sous-domaines A et B de la RT, au niveau des doigts (bleu). Cette insertion de 47 résidus est en accord avec l'alignement précédemment publié par Bartholomeusz *et al.* (Bartholomeusz et al., 2004). Parmi tous les alignements précédemment publiés et les différents alignements que nous avons calculé, il y a seulement une région stable de 83 résidus dans l'alignement, autour du motif YMDD, des résidus rtP170 à rtD252 (516-598 dans la séquence UniProtKB : Swiss-Prot : Q05486), surlignés en rouge (paume). Ces résidus correspondent au fond de la poche catalytique.

La combinaison des données à partir des alignements de RT et de RNase H et l'optimisation manuelle guidée par les structures secondaires prédites et observées a permis la définition d'un domaine de liaison entre les domaines RT et RNase H. L'alignement présente de nombreux blocs de délétions dans la séquence du VHB, au niveau du pouce (vert) et du domaine de connexion (jaune), par rapport à la séquence du VIH. Ceci indique un pouce et un domaine de connexion plus court dans la polymérase du VHB.

En effet, il manque 98 acides aminés dans la séquence du VHB par rapport à la séquence du VIH, entre le motif MGY situé aux positions rt250 à rt252 (UniProtKB : Swiss-Prot : Q05486 : 596-598), à l'extrémité C-terminale de la paume, et le résidu N-terminal du domaine RNase H (UniProtKB : Swiss-Prot : Q05486 : 691). Selon notre alignement, le pouce pourrait être compris entre les positions rt253 et rt304 (UniProtKB : Swiss-Prot : Q05486 : 599-650), indiqué en vert, et le domaine de connexion entre les positions rt305 et rt344 (UniProtKB : Swiss-Prot : Q05486 : 651-690), indiqué en jaune.

L'alignement des domaines RNases H du VHB et du VIH-1 (surlignage rose), est plutôt aisé dans la région N-terminale, avec des courts motifs conservés, et notamment les 3 premiers résidus catalytiques (D-E-D), qui sont situés avant l'hélice C putative. Ici, la grande différence avec l'alignement des RNases H du VHB et de *E. coli* réside dans le fait que le VIH-1 ne possède pas d'hélice C. De ce fait, dans cet alignement les résidus 760 à 789 du VHB représentent une grande insertion, avec une série de gaps correspondants dans la séquence du VIH-1 (en rose dans l'alignement).

La discussion à propos du 4ème D catalytique a été abordée dans la partie précédente, et en accord avec cette discussion, dans cet alignement nous avons aligné le 4ème D catalytique de la RNase H du VIH-1 avec le résidu D828 de la séquence UniProtKB : Swiss-Prot : Q05486 (qui correspond au résidu D817 dans la séquence UniProtKB : C7DSI5 de génotype D utilisée pour le modèle de RNase H dans la partie précédente).

Pour la comparaison avec les modèles de RT déjà publiés, il y a en réalité trois alignements à reproduire, dont un qui n'est pas complet dans la publication, pour reconstruire les modèles publiés. Nous avons donc aligné les séquences UniProtKB : Swiss-Prot : Q05486 et PDB : 1T05 en fonction des alignements de Das *et al*. (Das et al., 2001), de Bartholomeusz *et al*. (Bartholomeusz et al., 2004), et de Daga *et al*. (Daga et al., 2010) (Annexe 12).

4.3.2.2 La modélisation moléculaire des domaines RT et RNase H du VHB

De la même manière que précédemment, nous avons utilisé Geno3D pour générer les modèles à partir de l'empreinte du VIH-1 (PDB : 1T05).

Nous avons généré les modèles de Das *et al*. (Das et al., 2001) et de Daga *et al*. (Daga et al., 2010) à partir des alignements VHB-1T05 que nous avons reproduits, et du fichier de structure PDB : 1T05. Le modèle de Bartholomeusz *et al*. (Bartholomeusz et al., 2004) n'a pas été reproduit, puisque l'alignement du domaine RT n'est pas complet dans la publication.

Nous avons ensuite généré notre modèle à partir de l'alignement VHB-1T05 présenté sur la Figure 50, et de la structure PDB de 1T05, correspondant aux domaines RT et RNase H du VIH-1, reliés par un domaine de connexion.

Nous avons récupéré des contraintes utilisées pour générer le modèle de la RNase H à partir de *E.coli*. Ces contraintes sont celles qui avaient été mesurées sur la structure de *E. coli* et utilisées pour modéliser l'hélice C putative du VHB. Nous avons ajouté manuellement ces contraintes d'angles et de distances au fichier de contraintes pour pouvoir modéliser le repliement de l'hélice C dans ce modèle complet, alors qu'elle n'existe pas chez VIH-1, et n'est donc pas présente dans le fichier de structure de l'empreinte.

Pour chaque modélisation, comme précédemment, 10 modèles ont été générés puis vérifiés et validés. Parmi les 10 modèles, le modèle le plus exact a été sélectionné selon les critères d'énergie, de stéréochimie (PROCHECK), la distance avec l'empreinte (RMSD), et les violations de contraintes.

Pour les modèles reproduits de Das *et al*. et de Daga *et al*., à l'issue du processus de génération des modèles, les modèles 10 et 9 ont été sélectionnés respectivement, pour les analyses comparatives des poches catalytiques et pour les superpositions avec notre modèle.

Concernant notre modèle, l'ensemble des dix modèles a été calculé avec 14293 contraintes de distances et 1118 contraintes d'angles dièdres, impliquant 296 résidus. Ces chiffres incluent 70 contraintes de distances et 20 contraintes de dièdres pour l'hélice C putative. Parmi les dix modèles générés, l'un des modèles présentait une valeur trop élevée de RMSD avec l'empreinte et a été supprimé pour les analyses suivantes. L'accord entre les neuf modèles restants est bon, avec une RMSD de 1,85 Å, calculé sur 291 atomes Cα. La Figure 51A présente la superposition des 9 modèles.

L'accord est meilleur dans le domaine de RT dans lequel la séquence VHB partage plus de similitudes avec la séquence de l'empreinte que dans le domaine RNase H. Le modèle numéro 5 a été sélectionné pour les analyses supplémentaires, en fonction des critères cités précédemment (Annexe 13). Il est présenté sur la Figure 51B, en mode surface transparente, avec les structures secondaires apparentes (cartoon). Les sous domaines de la RT sont colorés en rouge pour la paume, bleu pour les doigts, et vert pour le pouce. Le domaine de connexion est présenté en orange et la RNase H en violet avec l'hélice C putative en cyan.

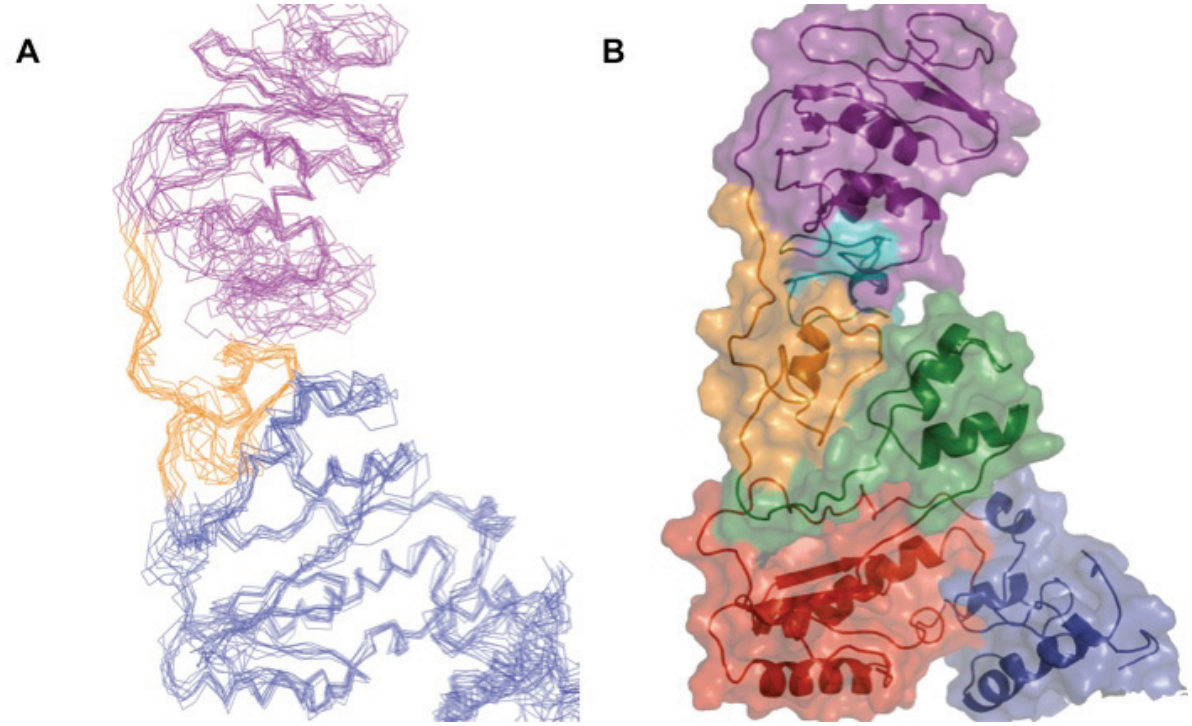


Figure 51 : Modèles des domaines RT-RNase H de la polymérase du VHB.
(A) Superposition des 9 modèles générés (RT en bleu, connecteur en orange, RNase H en violet). **(B)** Modèle 5 en représentation cartoon et surface (RT : paume en rouge, doigts en bleu, pouce en vert ; connecteur en orange ; RNase H en violet et hélice C cyan).

4.3.2.3 Comparaisons structurales - Superpositions

La superposition de la structure de l'empreinte (PDB : 1T05) et du modèle 5 réalisée par Geno3D est présentée sur la Figure 52 A et C. La valeur de RMSD correspondante est de 1,41 Å sur 291 atomes Cα. La valeur de RMSD entre le domaine RT du modèle et le domaine RT de l'empreinte est de 1,04 Å sur 200 atomes Cα. En dépit d'une bonne similitude structurale globale, il y a quelques différences locales, en particulier dans le

sous-domaine des doigts, où il y a la longue insertion spécifique des polymérases des virus de l'hépatite B infectant les mammifères, et dans le pouce, qui est plus court que celui de la RT du VIH-1.

La valeur de RMSD entre le modèle et l'empreinte pour le domaine RNase H est de 1,14 Å (calculée en utilisant 80 atomes Cα). Bien que cette valeur indique un bon accord entre le modèle et la structure de l'empreinte, il existe d'importantes différences locales situées autour de l'hélice C putative, puisque l'hélice C est absente chez VIH-1.

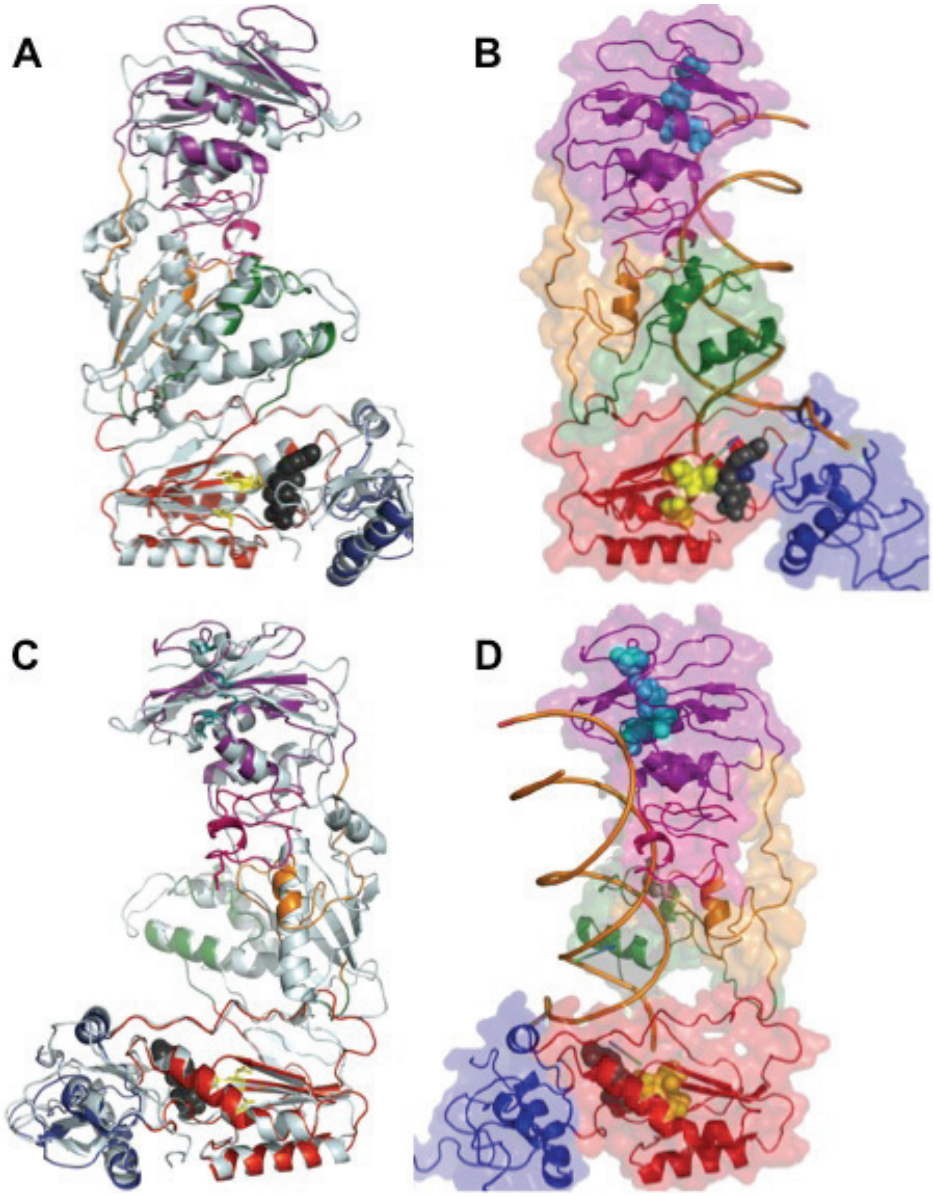


Figure 52 : Superposition du modèle avec l'empreinte 1T05.
(A) La structure de la RT du VIH est colorée en gris clair et le modèle du VHB est coloré en fonction des domaines et sous-domaines. **(B)** Modèle du VHB avec le TFV (noir) dans la poche catalytique et en interaction avec un acide nucléique. La triade catalytique de la RT est présentée en jaune, la tétrade catalytique de la RNase H en cyan, et l'hélice C en rose. **(C)** et **(D)** images A et B à 180°.

La Figure 52 B et D présente le modèle RT-RNase H du VHB, en interaction avec une double hélice d'ADN et le TFV (noir). Ces éléments sont présents dans la fichier PDB : 1T05 de la structure du VIH et ont été placés dans notre modèle lors de la superposition entre le modèle et l'empreinte. Les résidus de la triade catalytique du domaine RT sont représentés par des sphères jaunes, et ceux de la tétrade de la RNase H par des sphères cyan. La protrusion basique contenant l'hélice C est colorée en rose dans le domaine RNase H..

Afin de comparer les modèles de RT publiés de Das *et al*. et de Daga *et al*. avec le domaine RT de notre modèle, les 3 modèles sélectionnés ont été superposés à l'aide de Geno3D. Des valeurs de RMSD ont été calculées à partir de ces superpositions. Les valeurs de RMSD correspondant aux superpositions par paires des modèles de Das *et al*., Daga *et al*., et du domaine RT de notre modèle sont présentées dans le Tableau 12. Ces valeurs montrent que les 3 modèles sont structuralement assez différents les uns des autres.

Ces différences varient selon les sous-domaines RT. Les modèles sont plutôt similaires au niveau de la paume avec une RMSD moyenne pour les 3 modèles de 1,29 Å sur 85 atomes Cα. En effet, la paume est la région la plus conservée entre les séquences du VHB et du VIH-1, et donc la plus comparable dans les différents alignements. Dans les doigts, où la valeur moyenne de RMSD entre les 3 modèles est de 2,71 Å (sur 45 atomes Cα), les résidus supplémentaires du VHB (longue insertion), sont insérés de manière différente pour les 3 modèles. De ces différences dans les alignements résultent les différences structurales observées. Pour le pouce (2,80 Å RMSD plus de 27 atomes Cα), tous les modèles sont structuralement différents. Ceci s'explique par le fait que les 3 alignements sont très différents au niveau de cette région.

	Daga *et al.*	Notre modèle
Das *et al.*	2,85	2,33
Daga *et al.*	/	2,25

Tableau 12 : RMSD (en Å) entre les différents modèles de RT Das *et al*., Daga *et al*. et le nôtre.

4.3.2.4 Analyses de la poche catalytique

Comme pour beaucoup d'enzymes, la poche catalytique de la transcriptase inverse comprend une triade catalytique composée de 3 acides aspartiques (D). La triade catalytique du domaine RT du VIH-1 est composée des résidus D110, D185 et D186 (PDB : 1T05), les deux derniers résidus appartenant au motif YMDD. Pour le VHB, le premier D catalytique est le rtD83 et les deux autres, appartenant au motif YMDD sont les résidus rtD205 et rtD206 (UniProtKB : Swiss-Prot : Q05486 : D429, D551, D552).

Les analyses suivantes visent à explorer plus en détails les résidus de la poche catalytique, et notamment les positions spatiales des mutations connues de résistance aux traitements, ainsi que les résidus des modèles et de l'empreinte qui interagissent avec un analogue de nucléotide : le ténofovir.

4.3.2.4.1 Cartographie des mutations de résistance

Nous avons cartographié les mutations connues de résistance aux analogues de nucléos(t)ides (Zoulim and Locarnini, 2009) sur notre modèle, en les sélectionnant et leur attribuant une couleur, à l'aide du logiciel DeepView.

Après la cartographie des mutations de résistance sur notre modèle, nous avons défini des sphères autour du motif YMDD afin d'avoir la liste des résidus de la RT du VHB qui sont dans la poche catalytique. Ces analyses ont révélé que les dix positions connues pour être impliquées dans un profil de résistance à l'adéfovir (ADV), et/ou à l'entécavir (ETV) et/ou à la lamivudine (LMV), sont toutes situées à une distance inférieure ou égale à 10 Å autour du motif YMDD.

En effet, nous avons trouvé 6 positions : rtL180, rtA181, rtT184, rtM204, rtS202 et rtM250, à des distances ne dépassant pas 5 Å du motif YMDD. Les 4 positions restantes : rtL80, rtV173, rtI169 et rtN236, sont situées à une distance comprise entre de 5 à 10 Å du motif YMDD. La cartographie des mutations de résistance est présentée sur la Figure 53, avec les drogues concernées représentées par des pastilles de couleur.

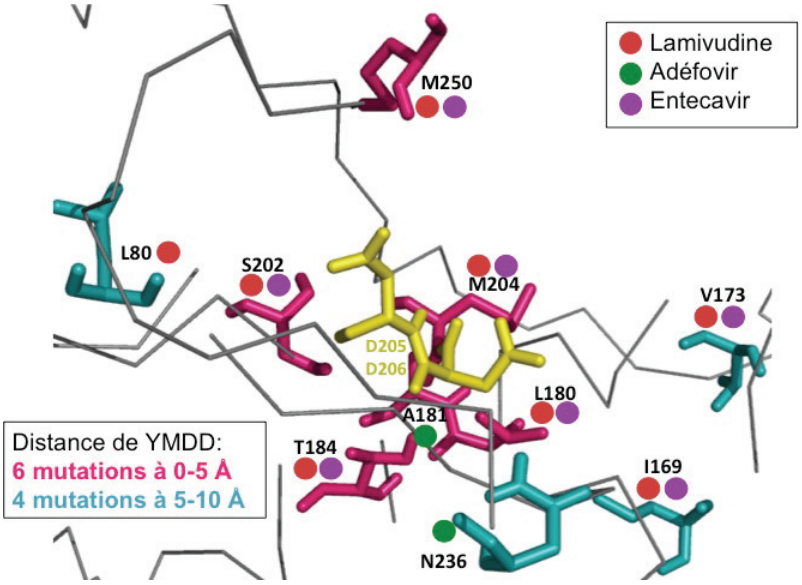


Figure 53 : Cartographie des positions de mutations de résistance connues dans le domaine RT du modèle.

4.3.2.4.2 Interactions avec le ténofovir

La structure de l'empreinte, PDB : 1T05, correspond en réalité à la structure de la sous-unité p66 du VIH-1 en complexe avec un analogue de nucléotide : le ténofovir (TFV). Or, le ténofovir est également utilisé dans le traitement contre le VHB.

Le TFV a donc été inséré dans la poche catalytique, pour chaque modèle, par une superposition avec 1T05, grâce au logiciel Geno3D. Les résidus de la sous-unité p66 ont ensuite été éliminés pour ne conserver que le TFV. Puis, les modèles avec TFV ont été régularisés par une minimisation d'énergie (100 étapes de l'algorithme « steepest descent ») en utilisant DeepView (Guex and Peitsch, 1997).

Les interactions entre TFV et les résidus de la poche catalytique ont ensuite été analysées pour les trois modèles, et pour la structure du VIH-1, avec le logiciel LIGPLOT (Wallace et al., 1995), disponible sur le site web PDBsum. LIGPLOT effectue des analyses détaillées des interactions protéine-ligand et produit des tracés précis de ces interactions.

Ces tracés présentent 2 types d'interactions entre les résidus et le ligand : les interactions hydrophobes et les liaisons hydrogènes. Le Tableau 13 résume les interactions entre le ténofovir et les résidus du domaine RT de la structure 1T05, du modèle reproduit de Das *et al.*, du modèle reproduit de Daga *et al.*, et du domaine RT de notre modèle.

1T05		**Das *et al.***		**Daga *et al.***		**Notre modèle**	
Résidu	**Interaction**	**Résidu**	**Interaction**	**Résidu**	**Interaction**	**Résidu**	**Interaction**
K65	H	rtD31	H				
				rtN33	B		
				rtP34	B		
				rtN36	H		
R72	H	rtR41*	B+H	rtR41*	B+H	rtR41*	H
D110	B	rtD83	B				
V111	H	rtV84	B	rtV84	B		
G112	B	rtS85	H			rtS85	B
D113	H	rtA86*	B	rtA86*	B	rtA86*	H
A114	B	rtA87	H				
		rtF88*	B	rtF88*	B	rtF88*	B
						rtY89	B
Q151	B	rtM171*	B	rtM171*	B	rtM171*	B
		rtM204	B				
D185	H	rtD205*	H	rtD205*	B	rtD205*	H

Tableau 13 : Liste des résidus de 1T05 et des 3 modèles interagissant avec le ténofovir.
Les colonnes « résidu » listent pour chaque structure et modèle les résidus en interaction avec le TFV. Les lignes indiquent les résidus structuralement équivalents entre eux. Les colonnes « interaction » indiquent le type d'interaction détectée par le LIGPLOT : H=liaison hydrogène, B=contact hydrophobe. Les résidus marqués d'une * sont ceux communs aux trois modèles.

Les tracés de LIGPLOT présentant les interactions listées dans le Tableau 13 sont présentés sur la Figure 54.

Selon l'analyse LIGPLOT, le TFV interagit avec 9 résidus du domaine RT du VIH-1. Pour les trois domaines RT des modèles VHB, le nombre de résidus en interaction avec le TFV est de 11 pour le modèle de Das *et al.*, 9 pour celui de Daga *et al.*, et 7 pour le nôtre. Cinq résidus : rtR41, rtA86, rtF88, rtM171 et rtD205 (UniProtKB : Swiss-Prot : Q05486 : R387, A432, F434, M517, D551) se trouvent être en interaction avec le TFV dans les trois modèles. Parmi les cinq résidus communs, tous les résidus, à l'exception rtF88, ont leurs résidus équivalents dans la structure du VIH-1 qui sont en interaction

avec le TFV. Le résidu rtF88 du VHB correspond au résidu PDB : 1T05 : Y115, qui n'interagit pas avec le TFV.

En dépit de quelques résidus communs, situés au fond de la poche catalytique, l'environnement du TFV est différent dans les 3 modèles. La diminution du nombre de résidus qui interagissent avec le TFV dans notre modèle pourrait être expliquée par la longue insertion au niveau les doigts.

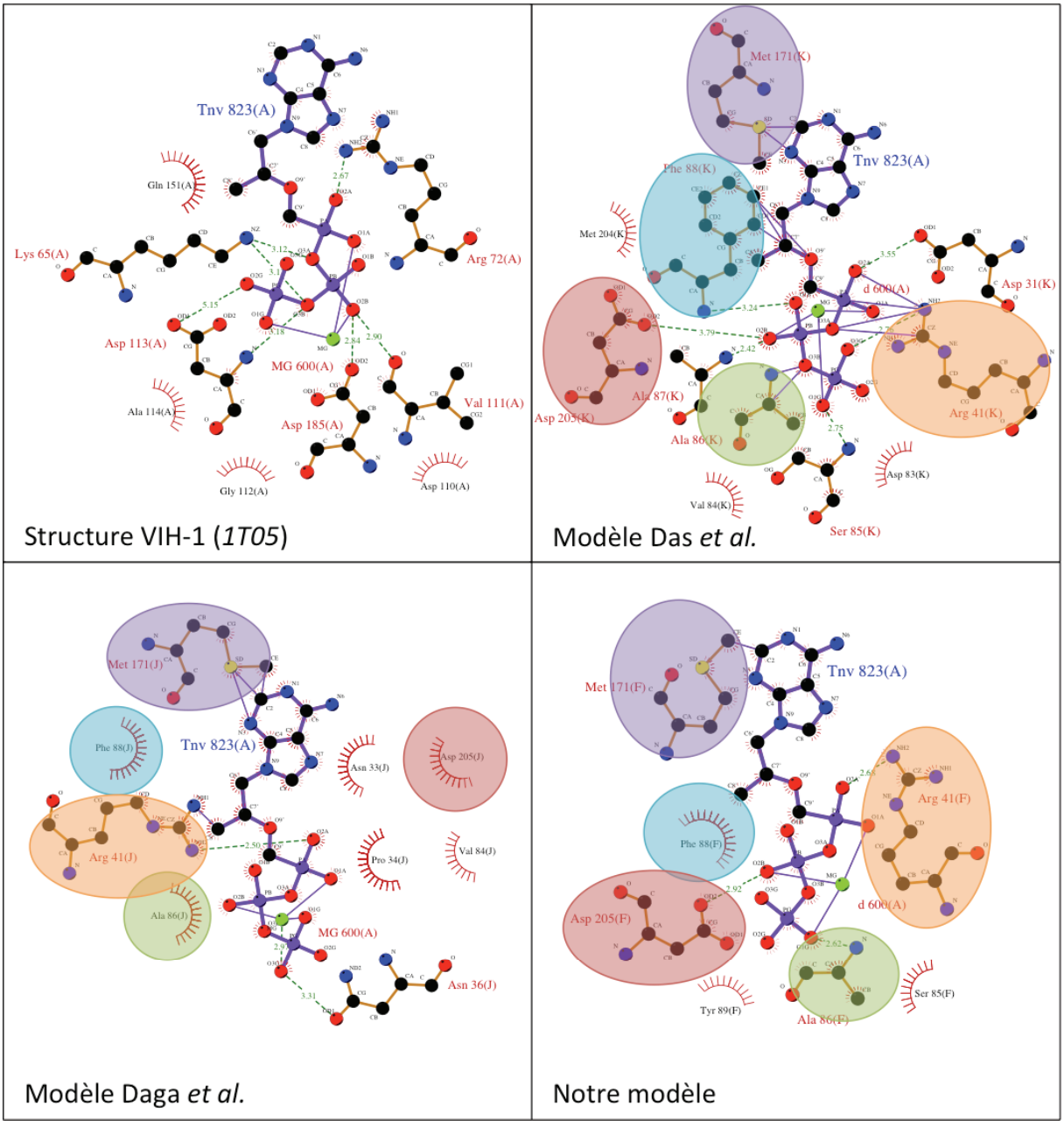


Figure 54 : Interactions entre le ténofovir et les résidus de la structure du VIH-1 et des 3 modèles du VHB.
Les 5 résidus communs aux trois modèles sont présentés dans les cercles colorés (R41 orange, A86 vert, F88 cyan, M171 violet, et D205 rouge).

4.3.2.5 Conception de mutants à partir des analyses sur le modèle

A partir des analyses réalisées sur le domaine RT des différents modèles, nous avons déterminé 5 positions qui semblent être en interaction avec le ténofovir : rtR41, rtA86, rtF88, rtM171 et rtD205. On peut alors supposer que des mutations sur ces positions ont un effet sur la sensibilité du virus au TFV. Afin de vérifier cette hypothèse, nous avons alors conçu plusieurs mutations pour chacune de ces positions afin de tester leur effet expérimentalement. Comme le cadre de lecture de la Pol est chevauchant avec le cadre de lecture des protéines de surface, pour chaque mutation envisagée, nous avons regardé l'impact sur les codons correspondants dans HBs.

C'est Judith Fresquet, de l'équipe du Pr. Fabien Zoulim à Lyon, qui a réalisé les constructions de mutants, et qui a effectué les tests *in vitro* sur ces mutants. Pour chaque mutant, le niveau de réplication a été mesuré, ainsi que le niveau de sécrétion de HBs. Pour les mutants présentant un niveau de réplication le permettant, le TFV a été ajouté afin de mesurer l'IC50 et l'IC90, qui représentent les concentrations de drogue (TFV) nécessaires pour une inhibition de 50% et de 90%, respectivement. Ces valeurs témoignent de la sensibilité du virus à la drogue. Si l'IC50 mesurée pour un mutant est beaucoup plus élevée que l'IC50 mesurée pour une souche sauvage, cela signifie que le mutant résiste à la drogue.

Des résultats préliminaires de cette étude sont présentés dans le Tableau 14. On y retrouve : les substitutions d'acides aminés pour les 5 positions de la RT, ainsi que les modifications de codons et le nombre de changements dans le codon, les informations correspondantes pour HBs, le pourcentage de réplication, l'IC50 et le pourcentage de sécrétion d'HBs. Les changements de codons entrainant des substitutions silencieuses dans HBs ne sont pas indiqués.

Mutant (rt)	Codons	Nb nt changés	Réplication (%)	IC50 (µM)	Mutant HBs	Codons	Nb nt changés	Sécrétion HBs (%)
pCH9-3091-GFP			100 +/- 32,3	6,8 +/- 0,9				100 +/- 0,0
R41K	aga -> aAG	2	37,1 +/- 16,6	> 100	D33S	gac -> AGc	2	0,0 +/- 0,0
R41Q	aga -> CAa	2	3,8 +/- 1,9	> 100	D33N	gac -> Aac	1	0,0 +/- 0,0
R41E	aga -> GAa	2	0,1 +/- 0,01	ND	D33N	gac -> Aac	1	ND
R41A	aga -> GCa	2	13,3 +/- 6,5	> 100	D33H	gac -> Cac	1	0,0 +/- 0,0
A86S	**gcg -> Tcg**	**1**	**47,6 +/- 24,8**	**> 100**		**cgg -> cgg**	**0**	**28,5 +/- 12,3**
A86N	gcg -> AAT	3	40,8 +/- 21,3	> 70	R78M	cgg -> ATg	2	0,0 +/- 0,0
A86D	gcg -> gAC	2	4,7 +/- 0,5	69,0 +/- 19,4	R78T	cgg -> ACg	2	0,0 +/- 0,0
A86K	gcg -> AAg	2	1,6 +/- 0,5	ND		cgg -> Agg	1	ND
F88Y	ttt -> tAt	1	7,8 +/- 0,5	en cours	F80I	ttt -> Att	1	en cours
F88L	ttt -> Ctt	1	8,2 +/- 3,9	> 100		ttt -> ttt	0	0,0 +/- 0,0
F88A	ttt -> GCt	2	9,6 +/- 4,6	> 60	F80L	ttt -> Ctt	1	0,0 +/- 0,0
M171N	atg -> aAT	2	4,3 +/- 0,9	> 60	W163M	tgg -> ATg	2	0,0 +/- 0,0
M171K	atg -> aAg	1	4,6 +/- 2,1	> 60	W163R	tgg -> Agg	1	29,8 +/- 11,1
M171A	atg -> GCg	2	21,0 +/- 9,6	> 100	W163R	tgg -> Cgg	1	46,7 +/- 7,4

Tableau 14 : Résultats préliminaires pour les mutants des 5 positions interagissant avec le TFV

Pour induire une résistance au ténofovir, il faut que le mutant soit capable de se répliquer, que l'impact sur la sécrétion de HBs soit le plus faible possible, et bien sûr qu'il présente une IC50 assez élevée. Evidemment, plus le nombre de changements dans le codon est faible, moins le risque d'apparition spontanée du mutant est élevé. D'après ces résultats, le mutant le plus probable est le mutant rtA86S car il n'implique qu'un seul changement dans le codon de RT et aucun dans le codon de HBs, il a un niveau de réplication de quasiment 50%, et l'IC50 est plus de 10 fois supérieure à celle d'une souche sauvage. La mutation rtA86S est donc une mutation de résistance potentielle au ténofovir.

4.4 Discussion et perspectives

4.4.1 Le modèle de RNase H

La combinaison de la modélisation moléculaires et des analyses de séquences sur le domaine RNase H du VHB nous a permis de déterminer les caractéristiques importantes de ce domaine enzymatique. Il s'agit d'une RNase H de type 1 avec un domaine de protrusion basique contenant une hélice α C. Les RNases H de type 1 fonctionnent avec un site actif en tétrade catalytique D-E-D-D. Pour la RNase H du VHB, les 3 premiers résidus catalytiques apparaissent de manière évidente (UniProtKB : C7DSI5 : D689, E718 et D737, ou UniProtKB : SwissProt : Q05486 : D700, E729 et D748).

La question se pose pour le 4ème D catalytique, pour lequel nous proposons le résidu UniProtKB : C7DSI5 : D817.

Les diverses analyses de séquences et prédiction de structures secondaires nous ont conduit à éliminer les positions D777 et D778 comme candidats pour le 4ème D catalytique, bien qu'ils soient très conservés.

Les D807 et D817 sont plus susceptibles d'être alignés avec le 4ème D catalytique de la RNase H de *E. coli*. Cependant, ces deux positions présentent de la variabilité, ce qui est inattendu pour un résidu catalytique. Si on regarde le D807, après exclusion des génotypes F et G du jeu de données *Dpop* (car il n'y a pas suffisamment de séquences pour ces génotypes), on remarque que les génotypes A, C et E présentent principalement une valine à cette position. Un intérêt majeur de l'ensemble de données *Dudps* est de fournir la variabilité observée au niveau du patient. Pour la position 807, le jeu de données *Dudps* indiquait que trois patients avaient principalement les substitutions D807A et D807V. Pour la position 817, la conservation est meilleure (sauf pour le génotype A où V est prédominant, après exclusion des génotypes F, G et H) d'après les jeux de données *all*, *notD* et *Dpop*. En revanche, le jeu de données *Dudps* a révélé que 20 patients avaient une substitution D817V (avec un patient présentant une mutation du D807) et que 2 patients avaient une substitution D817A. Le patient montrant la double mutation nous a permis d'écarter un mécanisme de complémentation possible entre la D807 et D817.

Ces données nous conduisent à deux questions :

- la RNase H peut elle être active avec seulement trois résidus catalytiques et ainsi chélater un seul cation divalent ?
- Pourquoi dans le sérum de certains patients, les mutants ayant une RNase H potentiellement défective peuvent représenter la population dominante, sans pour autant réduire la charge virale?

Nous avons essayé de répondre à ces questions et émis quelques hypothèses.

La première est qu'il est peu probable que la RNase H du VHB puisse tolérer de travailler avec un seul cation magnésium selon les données sur le RNases H de type 1 virales et non virales, et selon des études mécanistiques (Katayanagi et al., 1993; Lim et al., 2006; Nowotny et al., 2005; Nowotny et al., 2007; Kim et al., 2012; Ho et al., 2010) publiées à ce jour.

Ainsi, le D817 reste le meilleur candidat pour être le 4ème résidu catalytique. Cette hypothèse est renforcée par la présence d'une histidine conservée (H814) quelques résidus en amont de cette position (Tadokoro and Kanaya, 2009).

La deuxième hypothèse concerne l'explication possible de l'existence de ces mutants défectifs assez répandus. L'hypothèse est qu'une autre pression de sélection est exercée, et s'oppose à la sélection du D à la position du 4ème catalytique. Sur ce point, l'organisation génomique du VHB joue un rôle important, et le chevauchement du gène codant pour la protéine HBx peut être responsable de ce signal brouillé de conservation du 4ème résidu catalytique. Il faut noter que le 3ème nucléotide d'un codon de HBx correspond au 2ème nucléotide d'un codon de la RNase H. Ainsi, le changement de la 3ème base dans HBx se traduira par une substitution D vers A, V ou G dans la RNase H. Ce sont les substitutions effectivement observées.

Si l'on s'intéresse au fait que la charge virale ne réduit pas chez des patients présentant des mutants potentiellement défectifs pour la RNase H, on peut proposer plusieurs hypothèses.

La première hypothèse pouvant expliquer la présence de ces mutants défectifs est celle d'une trans-complémentation par des RNases H virales actives. En effet, au sein de la quasi-espèce, on peut trouver des virus présentant un domaine RNase H avec une tétrade catalytique complète et fonctionnelle. Les polymérases de ces virus pourraient alors complémenter la RNase H défective pour dégrader l'ARN de l'hétéro-duplex. Cette hypothèse suggère un fonctionnement en dimère. Or, il n'y a pas de données quant à

l'existence de dimère pour la polymérase du VHB. Il semblerait qu'elle soit plutôt à l'état de monomère, comme c'est le cas pour MoMLV.

La dernière hypothèse est que, compte tenu de la conservation de la fonction et de la structure des RNases H à travers les espèces, une RNase H cellulaire de l'hôte pourrait trans-complémenter la RNase H virale défective.

4.4.2 Le modèle RT-RNase H

Nous avons construit un model étendu de la polymérase du VHB, incluant les deux domaines enzymatiques RT et RNase H. Ceci dans le but de placer la RNase H dans son contexte, notamment pour vérifier le positionnement de la protrusion basique contenant l'hélice C, mais aussi pour répondre à la question sur l'existence d'un domaine de connexion entre les domaines RT et RNase H.

Ces travaux de modélisation ont souligné que, bien que les séquences de polymérase du VHB et celles du VIH-1 et du MoMLV soient des homologues distants, la séquence du VHB reste compatible avec le repliement observé chez ces deux virus, avec un domaine de connexion permettant de relier les deux domaines catalytiques.

Alors que l'homologie entre le VIH-1 et le VHB polymérase est fortement soutenue par de nombreuses études fonctionnelles publiées à ce jour, le calcul d'un alignement de séquences fiable et d'un modèle a été difficile en raison de la faible identité de séquence et du pourcentage de gaps élevé (61 résidus de différence entre les 2 protéines). En effet, il y a trois régions dans l'alignement avec des *indels* (insertions-délétions) importantes entre les séquences.

La première zone est située dans le sous-domaine des doigts du domaine RT où les séquences de VHB de mammifères ont une longue insertion de 47 résidus. En effet, c'est une insertion que l'on n'observe pas chez les virus de l'hépatite B infectant les oiseaux, comme le DHBV. Cette région représente également une insertion vis à vis de la séquence du VIH. Dans les modèles de RT précédemment publiés, les auteurs avaient distribué cette insertion en deux (Das et al., 2001) ou trois (Daga et al., 2010) insertions dans leurs alignements. Nous avons choisi de ne mettre qu'une longue insertion (entre rtA113 et rtH160, *i.e.* UniProtKB : Swiss-Prot : Q05486 : A459–H506), puisque c'est ce que l'on observe entre les *orthohepadnavirus* et les *avihepadnavirus*. De ce point de vue,

notre alignement est en accord avec celui de Bartholomeusz *et al.* (Bartholomeusz et al., 2004).

Le rôle structural et fonctionnel de cette insertion reste inconnu. Cette insertion semble être structurée d'après les prédictions de structures secondaires. Selon ces données, nous pouvons faire l'hypothèse que l'insertion pourrait être impliquée dans l'interaction avec des protéines cellulaires spécifiques de cellules de mammifères. Toutefois, la liaison à l'acide nucléique ou à d'autres protéines virales, y compris la polymérase, ne peut être exclue.

La seconde région qui a été difficile à aligner correspond au pouce et au domaine de connexion. Dans cette région, la séquence du VHB présente une délétion de 98 résidus par rapport à la séquence du VIH-1. Sur la base de quelques petits motifs conservés et des structures secondaires, il est possible de définir un domaine de connexion avec un pouce plus court en contrepartie. Le domaine de connexion montre de nombreuses délétions par comparaison avec le VIH-1 et le MoMLV. Une des délétions est située au niveau des résidus 358-365 du VIH-1 (PDB : 1T05). Selon Lim *et al*., chez le VIH-1, ces résidus du connecteur peuvent compenser structuralement l'absence de l'hélice C (Lim et al., 2006; Schultz and Champoux, 2008).

La troisième région est la partie C-terminale de la RNase H à partir du résidu S759 (UniProtKB : Swiss-Prot : Q05486). Sans tenir compte de l'existence éventuelle d'une hélice C, on ne peut pas produire d'alignement satisfaisant. Selon notre alignement, la RNase H du VHB présente une longue insertion de 30 résidus par rapport au VIH-1. Cette insertion contient l'hélice C putative, que l'on a pu modéliser grâce au modèle de RNase H basé sur la structure de *E. coli*.

Cette insertion correspondant très certainement à une partie de la protrusion basique, et se situant structuralement dans la partie inférieure de la RNase H (partie la plus proche de la RT), elle pourrait compenser structuralement le court domaine de connexion (à l'inverse du VIH-1).

Enfin, les analyses sur le site catalytique de notre modèle ont montré que les positions portant des mutations connues de résistance aux drogues se situaient à des distances du motif YMDD n'excédant pas les 10 Å. Cette information valide la fiabilité du modèle, au moins pour la poche catalytique.

Les analyses d'interactions avec le ténofovir ont révélé 5 résidus qui étaient en interaction avec le TFV dans les 3 modèles (le notre, celui de Das *et al*., et celui de Daga *et al*.). A partir de ces résultats, nous avons conçu des mutants pour chacune de ces positions. Ces mutants ont été construits, et leur réplication en présence et en absence de ténofovir a été testée, ainsi que l'impact sur la protéine HBs (codée sur le cadre de lecture chevauchant). Cette étude a fait ressortir le mutant rtA86S comme un mutant potentiellement résistant au TFV. Or, aucune étude ne décrit ce mutant et les données issues de HBVdb montrent une conservation complète du A à cette position. D'ailleurs, on n'a pas encore observé l'émergence de mutants résistants au TFV. D'après ces résultats préliminaires, on a pu voir qu'il est possible de trouver des mutants qui résistent au TFV. Alors pourquoi ces mutants ne sont pas observés dans la nature ? Peut être parce que la barrière génétique du TFV est trop élevée pour observer leur émergence. La question reste ouverte.

L'ensemble de ces données nous amène à penser que notre modèle est plausible. Cependant, le faible pourcentage d'identité entre la polymérase du VHB et les empreintes potentielles pour la modélisation moléculaire nous invitent à prendre tous les modèles avec précaution, en particulier dans l'analyse des mutations de résistance ou dans la prédiction de résistances.

Enfin, la partie N-terminale (TP) de la polymérase du VHB impliquée dans l'amorçage de la transcription inverse et la région charnière reliant le domaine TP et le domaine RT, essentielle pour la flexibilité de la protéine, restent à modéliser, en l'absence d'une structure expérimentale de l'enzyme.

Des études expérimentales peuvent être envisagées, sur la base des hypothèses formulées à partir de ces modèles. Par exemple, des mutants peuvent être construits afin de vérifier les hypothèses sur le 4ème résidu catalytique de la RNase H.

De plus, ces travaux de modélisation ont permis de décrire des caractéristiques de la protéine, qui pourraient être utiles pour résoudre la structure de manière expérimentale.

Conclusion

La plupart des objectifs fixés au début du projet ont été atteints. La base de connaissances HBVdb et son processus de génération et de mise à jour automatique sont fonctionnels. Elle est disponible sur le web depuis fin juin 2012 et déjà utilisée par des chercheurs à travers le monde.

Elle permet à chaque utilisateur de récupérer et analyser des données de la base, mais aussi d'analyser ses propres séquences grâce aux outils spécialisés d'annotation, de génotypage et de détection de profils de résistance aux traitements.

A partir des données d'HBVdb, nous avons pu réaliser plusieurs analyses de séquences, qui nous ont aidé, notamment pour construire un modèle moléculaire du domaine RNase H de la polymérase du virus de l'hépatite B.

Grâce à ces analyses et à la modélisation par homologie à partir de la RNase H de *E. coli*, nous avons pu montrer que la RNase H du VHB est une RNase H de type 1, qui possède un domaine de protrusion basique contenant une hélice C, impliqué dans l'interaction avec l'hétéro-duplex ARN-ADN.

Les RNases H de type 1 assurent leur activité catalytique avec une tétrade catalytique. Nous avons déterminé les 3 premiers résidus formant cette tétrade chez le VHB, puis nous avons fait une proposition quant à la position du 4ème D catalytique parmi 4 positions possibles. Or, d'après les analyses de séquences, il existe une proportion importante de virus mutés au niveau de ce 4ème résidu catalytique, donc une proportion assez important de virus potentiellement défectifs au niveau de la RNase H. Nous avons proposé des hypothèses pour expliquer ce phénomène, qui pourraient être testées expérimentalement.

Enfin, dans le but de placer le modèle de la RNase H dans un contexte plus global, et pour vérifier le positionnement de l'hélice C, nous avons construit un modèle plus étendu, comprenant le domaine RT et le domaine RNase H. Ce modèle nous a permis d'émettre une hypothèse quant à l'existence d'un domaine de connexion entre les deux domaines.

Grâce à une analyse approfondie de la poche catalytique de la RT, en y plaçant un analogue de nucléotide, le ténofovir, nous avons conçu des mutants potentiellement résistants au ténofovir. Ces mutants ont été construits puis testés, et cette étude est toujours en cours.

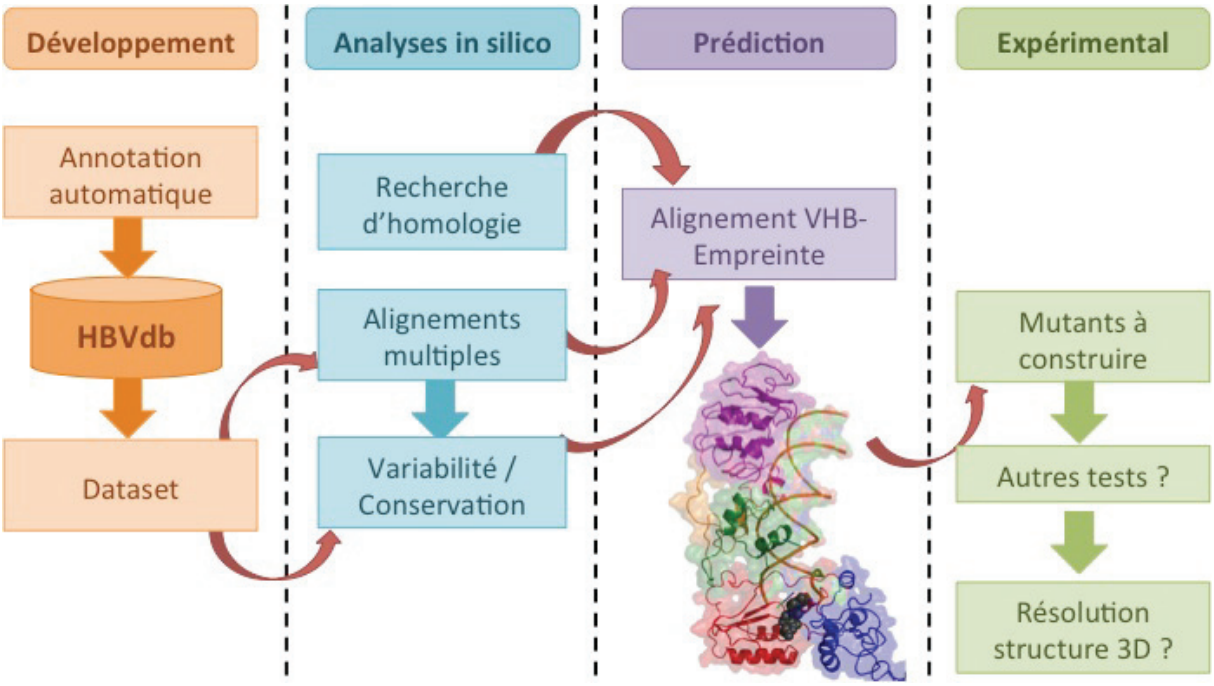


Figure 55 : Mes travaux de thèse intégrés à ma vision simplifiée de la bioinformatique.

Références bibliographiques

1. Akuta, N., and Kumada, H. (2005). Influence of hepatitis B virus genotypes on the response to antiviral therapies. J Antimicrob Chemother *55*, 139-142.

2. Alcantara, L.C., Cassol, S., Libin, P., Deforche, K., Pybus, O.G., Van Ranst, M., Galvão-Castro, B., Vandamme, A.M., and de Oliveira, T. (2009). A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res *37*, W634-642.

3. Alter, H.J., and Blumberg, B.S. (1966). Further studies on a "new" human isoprecipitin system (Australia antigen). Blood *27*, 297-309.

4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

5. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

6. Amid, C., Birney, E., Bower, L., Cerdeño-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Gibson, R., Goodgame, N., Hunter, C., Jang, M., Leinonen, R., Liu, X., Oisel, A., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Rivière, S., Rossello, M., Senf, A., Smirnov, D., Ten Hoopen, P., Vaughan, D., Vaughan, R., Zalunin, V., and Cochrane, G. (2012). Major submissions tool developments at the European Nucleotide Archive. Nucleic Acids Res *40*, D43-47.

7. Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res *36*, D419-425.

8. Angus, P., Vaughan, R., Xiong, S., Yang, H., Delaney, W., Gibbs, C., Brosgart, C., Colledge, D., Edwards, R., Ayres, A., Bartholomeusz, A., and Locarnini, S. (2003). Resistance to adefovir dipivoxil therapy associated with the selection of a novel mutation in the HBV polymerase. Gastroenterology *125*, 292 - 297.

9. Arauz-Ruiz, P., Norder, H., Robertson, B.H., and Magnius, L.O. (2002). Genotype H: a new Amerindian genotype of hepatitis B virus revealed in Central America. J Gen Virol *83*, 2059-073.

10. Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res *28*, 45-48.

11. Bartenschlager, R., and Schaller, H. (1992). Hepadnaviral assembly is initiated by polymerase binding to the encapsidation signal in the viral RNA genome. EMBO J *11*, 3413-420.

12. Bartholomeusz, A., Tehan, B.G., and Chalmers, D.K. (2004). Comparisons of the HBV and HIV polymerase, and antiviral resistance mutations. Antivir Ther *9*, 149-160.

13. Bayard, F., Godon, O., Nalpas, B., Costentin, C., Zhu, R., Soussan, P., Vallet-Pichard, A., Fontaine, H., Mallet, V., Pol, S., and Michel, -L. (2012). T-cell responses to hepatitis B splice-generated protein of hepatitis B virus and inflammatory cytokines/chemokines in chronic hepatitis B patients. ANRS study: HB EP 02 HBSP-FIBRO. J Viral Hepat *19*, 872-880.

14. Beck, J., and Nassal, M. (2007). Hepatitis B virus replication. World J Gastroenterol *13*, 48-64.

15. Benhenda, S., Cougot, D., Buendia, M.-A., and Neuveut, C. (2009). Chapter 4 Hepatitis B Virus X Protein: Molecular Functions and Its Role in Virus Life Cycle and Pathogenesis. In Advances in Cancer Research, G.F.V. Woude, and G. Klein, eds. (Academic Press).

16. Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., and Sayers, E.W. (2012). GenBank. Nucleic Acids Res *40*, D48-D53.

17. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem *80*, 319-324.

18. Blumberg, B.S., Alter, H.J., and Visnich, S. (1965). A "new" antigen in leukemia sera. JAMA *191*, 541-46.

19. Blumberg, B.S., Gerstley, B.J., Hungerford, D.A., London, W.T., and Sutnick, A.I. (1967). A serum antigen (Australia antigen) in Down's syndrome, leukemia, and hepatitis. Ann Intern Med *66*, 924-931.

20. Bock, C.T., Schwinn, S., Locarnini, S., Fyfe, J., Manns, M.P., Trautwein, C., and Zentgraf, H. (2001). Structural organization of the hepatitis B virus minichromosome. J Mol Biol *307*, 183-196.

21. Bock, C.T., Tillmann, H.L., Maschek, H.J., Manns, M.P., and Trautwein, C. (1997). A preS mutation isolated from a patient with chronic hepatitis B infection leads to virus retention and misassembly. Gastroenterology *113*, 1976-982.

22. Bond W. W., Favero M. S., Petersen N. J., Gravelle C. R., Ebert J. W., & Maynard J. E. (1981). Lancet. In *Survival of hepatitis B virus after drying and storage for one week.*. ENGLAND.

23. Bouchard, M.J., and Schneider, R.J. (2004). The enigmatic X gene of hepatitis B virus. J Virol *78*, 12725-734.

24. Bourne, C.R., Finn, M.G., and Zlotnick, A. (2006). Global structural changes in hepatitis B virus capsids induced by the assembly effector HAP1. J Virol *80*, 11055-061.

25. Brunelle, M.-N., Jacquard, A.-C., Pichoud, C., Durantel, D., Carrouée-Durantel, S., Villeneuve, J.-P., Trépo, C., and Zoulim, F. (2005). Susceptibility to antivirals of a human HBV strain with mutations conferring resistance to both lamivudine and adefovir. Hepatology *41*, 1391-98.

26. Bruss, V. (2004). Envelopment of the hepatitis B virus nucleocapsid. Virus Res *106*, 199-209.

27. Bruss, V. (2007). Hepatitis B virus morphogenesis. World J Gastroenterol *13*, 65-73.

28. Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S., and Venter, J.C. (1996). Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science *273*, 1058-073.

29. Carman, W.F., Jacyna, M.R., Hadziyannis, S., Karayiannis, P., McGarvey, M.J., Makris, A., and Thomas, H.C. (1989). Mutation preventing formation of hepatitis B e antigen in patients with chronic hepatitis B infection. Lancet *2*, 588-591.

30. Carman, W.F., Zanetti, A.R., Karayiannis, P., Waters, J., Manzillo, G., Tanzi, E., Zuckerman, A.J., and Thomas, H.C. (1990). Vaccine-induced escape mutant of hepatitis B virus. Lancet *336*, 325-29.

31. Ceres, P., and Zlotnick, A. (2002). Weak protein-protein interactions are sufficient to drive assembly of hepatitis B virus capsids. Biochemistry *41*, 11525-531.

32. Cerritelli, S.M., and Crouch, R.J. (2009). Ribonuclease H: the enzymes in eukaryotes. FEBS Journal *276*, 1494-1505.

33. Champoux, J.J., and Schultz, S.J. (2009). Ribonuclease H: properties, substrate specificity and roles in retroviral reverse transcription. FEBS Journal *276*, 1506-516.

34. Chen, W.-N., Chen, J.-Y., Jiao, B.-Y., Lin, W.-S., Wu, Y.-L., Liu, L.-L., and Lin, X. (2012). Interaction of the Hepatitis B Spliced Protein with Cathepsin B Promotes Hepatoma Cell Migration and Invasion. J Virol *86*, 13533-541.

35. Chen, Y., and Marion, P.L. (1996). Amino acids essential for RNase H activity of hepadnaviruses are also required for efficient elongation of minus-strand viral DNA. J Virol *70*, 6151-56.

36. Chong, Y., and Chu, C.K. (2002). Understanding the unique mechanism of L-FMAU (clevudine) against hepatitis B virus: molecular dynamics studies. Bioorg Med Chem Lett *12*, 3459-462.

37. Chou, P.Y., and Fasman, G.D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol *47*, 45-148.

38. Clercq, E.D. (2007). The acyclic nucleoside phosphonates from inception to clinical use: Historical perspective. Antiviral Res *75*, 1 - 13.

39. Combet, C., Blanchet, C., Geourjon, C., and Deleage, G. (2000). NPS@: network protein sequence analysis. Trends in biochemical sciences *25*, 147.

40. Combet, C., Garnier, N., Charavay, C., Grando, D., Crisan, D., Lopez, J., Dehne-Garcia, A., Geourjon, C., Bettler, E., Hulo, C., Le Mercier, P., Bartenschlager, R., Diepolder, H., Moradpour, D., Pawlotsky, J.M., Rice, C.M., Trépo, C., Penin, F., and Deléage, G. (2007). euHCVdb: the European hepatitis C virus database. Nucleic Acids Res *35*, D363-66.

41. Combet, C., Jambon, M., Deléage, G., and Geourjon, C. (2002). Geno3D: automatic comparative molecular modelling of protein. Bioinformatics *18*, 213-14.

42. Cooreman, M.P., Leroux-Roels, G., and Paulij, W.P. (2001). Vaccine- and hepatitis B immune globulin-induced escape mutations of hepatitis B virus surface antigen. J Biomed Sci *8*, 237-247.

43. Couroucé, A.M., Lee, H., Drouet, J., Canavaggio, M., and Soulier, J.P. (1983). Monoclonal antibodies to HBsAg: a study of their specificities for eight different HBsAg subtypes. Dev Biol Stand *54*, 527-534.

44. Craxi, A., and Cooksley, W.G. (2003). Pegylated interferons for chronic hepatitis B. Antiviral Res *60*, 87-89.

45. Crowther, R.A., Kiselev, N.A., Böttcher, B., Berriman, J.A., Borisova, G.P., Ose, V., and Pumpens, P. (1994). Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy. Cell *77*, 943-950.

46. Cuff, A.L., Sillitoe, I., Lewis, T., Redfern, O.C., Garratt, R., Thornton, J., and Orengo, C.A. (2009). The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res *37*, D310-14.

47. Daga, P.R., Duan, J., and Doerksen, R.J. (2010). Computational model of hepatitis B virus DNA polymerase: molecular dynamics and docking to understand resistant mutations. Protein Sci *19*, 796-807.

48. Dane, D.S., Cameron, C.H., and Briggs, M. (1970). Virus-like particles in serum of patients with Australia-antigen-associated hepatitis. Lancet *1*, 695-98.

49. Das, D., and Georgiadis, M.M. (2004). The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. Structure *12*, 819-829.

50. Das, K., Xiong, X., Yang, H., Westland, C.E., Gibbs, C.S., Sarafianos, S.G., and Arnold, E. (2001). Molecular modeling and biochemical characterization reveal the mechanism of hepatitis B virus polymerase resistance to lamivudine (3TC) and emtricitabine (FTC). J Virol *75*, 4771-79.

51. Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978). In Atlas of protein sequence and structure .

52. Delaney, W.E., Ray, A.S., Yang, H., Qi, X., Xiong, S., Zhu, Y., and Miller, M.D. (2006). Intracellular metabolism and in vitro activity of tenofovir against hepatitis B virus. Antimicrob Agents Chemother *50*, 2471-77.

53. Delaney, W.E., Yang, H., Westland, C.E., Das, K., Arnold, E., Gibbs, C.S., Miller, M.D., and Xiong, S. (2003). The hepatitis B virus polymerase mutation rtV173L is selected during lamivudine therapy and enhances viral replication in vitro. J Virol *77*, 11833-841.

54. Dény, P., and Zoulim, F. (2010). Hepatitis B virus: from diagnosis to treatment. Pathol Biol (Paris) *58*, 245-253.

55. Dienstag, J.L. (2008). Hepatitis B Virus Infection. N Engl J Med *359*, 1486-1500.

56. Dutta, S., Burkhardt, K., Young, J., Swaminathan, G.J., Matsuura, T., Henrick, K., Nakamura, H., and Berman, H.M. (2009). Data Deposition and Annotation at the Worldwide Protein Data Bank. Molecular Biotechnology *42*, 1-13.

57. Eble, B.E., MacRae, D.R., Lingappa, V.R., and Ganem, D. (1987). Multiple topogenic sequences determine the transmembrane orientation of the hepatitis B surface antigen. Mol Cell Biol *7*, 3591-3601.

58. Eddy, S.R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol *4*, e1000069.

59. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, e1002195.

60. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-97.

61. European Association For The Study Of The Liver (2012). EASL Clinical Practice Guidelines: Management of chronic hepatitis B virus infection. J Hepatol *57*, 167-185.

62. Fafi-Kremer, S., Fofana, I., Soulier, E., Carolla, P., Meuleman, P., Leroux-Roels, G., Patel, A.H., Cosset, F.L., Pessaux, P., Doffoël, M., Wolf, P., Stoll-Keller, F., and Baumert, T.F. (2010). Viral entry and escape from antibody-mediated neutralization influence hepatitis C virus reinfection in liver transplantation. J Exp Med *207*, 2019-031.

63. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., and Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature *260*, 500-07.

64. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995).

Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

65. Fofana, I., Fafi-Kremer, S., Carolla, P., Fauvelle, C., Zahid, M.N., Turek, M., Heydmann, L., Cury, K., Hayer, J., Combet, C., Cosset, F.L., Pietschmann, T., Hiet, M.S., Bartenschlager, R., Habersetzer, F., Doffoël, M., Keck, Z.Y., Foung, S.K., Zeisel, M.B., Stoll-Keller, F., and Baumert, T.F. (2012). Mutations that alter use of hepatitis C virus cell entry factors mediate escape from neutralizing antibodies. Gastroenterology *143*, 223-233.e9.

66. Galibert, F., Chen, T.N., and Mandart, E. (1982). Nucleotide sequence of a cloned woodchuck hepatitis virus genome: comparison with the hepatitis B virus sequence. J Virol *41*, 51-65.

67. Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P., and Charnay, P. (1979). Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in E. coli. Nature *281*, 646-650.

68. Geourjon, C., and Deléage, G. (1994). SOPM: a self-optimized method for protein secondary structure prediction. Protein Eng *7*, 157-164.

69. Geourjon, C., and Deléage, G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci *11*, 681-84.

70. Geourjon, C., Combet, C., Blanchet, C., and Deléage, G. (2001). Identification of related proteins with weak sequence identity using secondary structure information. Protein Sci *10*, 788-797.

71. Ghany, M., and Liang, T.J. (2007). Drug targets and molecular mechanisms of drug resistance in chronic hepatitis B. Gastroenterology *132*, 1574-585.

72. Glebe, D., and Urban, S. (2007). Viral and cellular determinants involved in hepadnaviral entry. World J Gastroenterol *13*, 22-38.

73. Gnaneshan, S., Ijaz, S., Moran, J., Ramsay, M., and Green, J. (2007). HepSEQ: International Public Health Repository for Hepatitis B. Nucleic Acids Res *35*, D367-370.

74. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. (1996). Life with 6000 genes. Science *274*, 546, 563-67.

75. Gouet, P., Robert, X., and Courcelle, E. (2003). ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. Nucleic Acids Res *31*, 3320-23.

76. Greene, L.H., Lewis, T.E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J.M., and Orengo, C.A. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res *35*, D291-97.

77. Guex, N., and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis *18*, 2714-723.

78. Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A *89*, 10915-19.

79. Ho, M.H., De Vivo, M., Dal Peraro, M., and Klein, M.L. (2010). Understanding the effect of magnesium ion concentration on the catalytic activity of ribonuclease H through computation: does a third metal binding site modulate endonuclease catalysis? J Am Chem Soc *132*, 13702-712.

80. Hollinger, .B. (2007). Hepatitis B virus genetic diversity and its impact on diagnostic assays. J Viral Hepat *14*, 11-15.

81. Holm, L., and Rosenström, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res *38*, W545-49.

82. Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., and Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res *39*, D576-582.

83. Innaimo, S.F., Seifer, M., Bisacchi, G.S., Standring, D.N., Zahler, R., and Colonno, R.J. (1997). Identification of BMS-200475 as a potent and selective inhibitor of hepatitis B virus. Antimicrob Agents Chemother *41*, 1444-48.

84. Ishikawa, K., Nakamura, H., Morikawa, K., Kimura, S., and Kanaya, S. (1993). Cooperative stabilization of Escherichia coli ribonuclease HI by insertion of Gly-80b and Gly-77-->Ala substitution. Biochemistry *32*, 7136-142.

85. Jadeau, F., Grangeasse, C., Shi, L., Mijakovic, I., Deléage, G., and Combet, C. (2012). BYKdb: the Bacterial protein tYrosine Kinase database. Nucleic Acids Res *40*, D321-24.

86. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-2637.

87. Kaminuma, E., Kosuge, T., Kodama, Y., Aono, H., Mashima, J., Gojobori, T., Sugawara, H., Ogasawara, O., Takagi, T., Okubo, K., and Nakamura, Y. (2011). DDBJ progress report. Nucleic Acids Res *39*, D22-27.

88. Kane, A., Lloyd, J., Zaffran, M., Simonsen, L., and Kane, M. (1999). Transmission of hepatitis B, hepatitis C and human immunodeficiency viruses through unsafe injections in the developing world: model-based regional estimates. Bull World Health Organ *77*, 801-07.

89. Kann, M., Sodeik, B., Vlachou, A., Gerlich, W.H., and Helenius, A. (1999). Phosphorylation-dependent binding of hepatitis B virus core particles to the nuclear pore complex. J Cell Biol *145*, 45-55.

90. Karlin, S., and Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A *87*, 2264-68.

91. Karsch-Mizrachi, I., Nakamura, Y., Cochrane, G., and International Nucleotide Sequence Database Collaboration (2012). The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res *40*, D33-37.

92. Katayanagi, K., Okumura, M., and Morikawa, K. (1993). Crystal structure of Escherichia coli RNase HI in complex with Mg2+ at 2.8 A resolution: proof for a single Mg(2+)-binding site. Proteins *17*, 337-346.

93. Kay, A., and Zoulim, F. (2007). Hepatitis B virus genetic variability and evolution. Virus Res *127*, 164-176.

94. Kenney, J.M., von Bonsdorff, C.H., Nassal, M., and Fuller, S.D. (1995). Evolutionary conservation in the hepatitis B virus core structure: comparison of human and duck cores. Structure *3*, 1009-019.

95. Kidd-Ljunggren, K., Miyakawa, Y., and Kidd, A.H. (2002). Genetic variability in hepatitis B viruses. J Gen Virol *83*, 1267-280.

96. Kim, J.H., Kang, S., Jung, S.K., Yu, K.R., Chung, S.J., Chung, B.H., Erikson, R.L., Kim, B.Y., and Kim, S.J. (2012). Crystal structure of xenotropic murine leukaemia virus-related virus (XMRV) ribonuclease H. Biosci Rep *32*, 455-463.

97. King, R.D., and Sternberg, M.J. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci *5*, 2298-2310.

98. Kirby, K.A., Marchand, B., Ong, Y.T., Ndongwe, T.P., Hachiya, A., Michailidis, E., Leslie, M.D., Sietsema, D.V., Fetterly, T.L., Dorst, C.A., Singh, K., Wang, Z., Parniak, M.A., and Sarafianos, S.G. (2012). Structural and inhibition studies of the RNase H function of xenotropic murine leukemia virus-related virus reverse transcriptase. Antimicrob Agents Chemother *56*, 2048-061.

99. Koike, K. (2009). Hepatitis B virus X gene is implicated in liver carcinogenesis. Cancer Lett *286*, 60-68.

100. Kramvis, A., and Kew, M.C. (2005). Relationship of genotypes of hepatitis B virus to mutations, disease progression and response to antiviral therapy. J Viral Hepat *12*, 456-464.

101. Kramvis, A., Arakawa, K., Yu, M.C., Nogueira, R., Stram, D.O., and Kew, M.C. (2008). Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. J Med Virol *80*, 27-46.

102. Krey, T., d'Alayer, J., Kikuti, C.M., Saulnier, A., Damier-Piolle, L., Petitpas, I., Johansson, D.X., Tawar, R.G., Baron, B., Robert, B., England, P., Persson, M.A., Martin, A., and Rey, F.A. (2010). The disulfide bonds in glycoprotein E2 of hepatitis C virus reveal the tertiary organization of the molecule. PLoS Pathog *6*, e1000762.

103. Lai, C.-L., Chien, R.-N., Leung, N.W., Chang, T.-T., Guan, R., Tai, D.-I., Ng, K.-Y., Wu, P.-C., Dent, J.C., Barber, J., Stephenson, S.L., and Gray, D.F. (1998). A One-Year Trial of Lamivudine for Chronic Hepatitis B. N Engl J Med *339*, 61-68.

104. Lai, C.-L., Leung, N., Teo, E.-K., Tong, M., Wong, F., Hann, H.-W., Han, S., Poynard, T., Myers, M., Chao, G., Lloyd, D., and Brown, N.A. (2005). A 1-Year Trial of Telbivudine, Lamivudine, and the Combination in Patients With Hepatitis B e Antigen—Positive Chronic Hepatitis B. Gastroenterology *129*, 528 - 536.

105. Langley, D.R., Walsh, A.W., Baldick, C.J., Eggers, B.J., Rose, R.E., Levine, S.M., Kapur, A.J., Colonno, R.J., and Tenney, D.J. (2007). Inhibition of hepatitis B virus polymerase by entecavir. J Virol *81*, 3992-4001.

106. Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR *8*, 477-486.

107. Lavanchy, D. (2004). Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures. J Viral Hepat *11*, 97-107.

108. Lee, J., Shin, M.K., Lee, H.J., Yoon, G., and Ryu, W.S. (2004). Three novel cis-acting elements required for efficient plus-strand DNA synthesis of the hepatitis B virus genome. J Virol *78*, 7455-464.

109. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V., and Cochrane, G. (2011). The European Nucleotide Archive. Nucleic Acids Res *39*, D28-D31.

110. Levin, J.M., and Garnier, J. (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. Biochim Biophys Acta *955*, 283-295.

111. Lim, D., Gregorio, G.G., Bingman, C., Martinez-Hackert, E., Hendrickson, W.A., and Goff, S.P. (2006). Crystal structure of the moloney murine leukemia virus RNase H domain. J Virol *80*, 8379-389.

112. Liu, N., Ji, L., Maguire, M.L., and Loeb, D.D. (2004). cis-Acting sequences that contribute to the synthesis of relaxed-circular DNA of human hepatitis B virus. J Virol *78*, 642-49.

113. Locarnini, S. (2008). Primary resistance, multidrug resistance, and cross-resistance pathways in HBV as a consequence of treatment failure. Hepatol Int *2*, 147.

114. Locarnini, S.A., and Yuen, L. (2010). Molecular genesis of drug-resistant and vaccine-escape HBV mutants. Antivir Ther *15*, 451-461.

115. Lok, A.S., Lai, C.-L., Leung, N., Yao, G.-B., Cui, Z.-Y., Schiff, E.R., Dienstag, J.L., Heathcote, E.J., Little, N.R., Griffiths, D.A., Gardner, S.D., and Castiglia, M. (2003). Long-term safety of lamivudine treatment in patients with chronic hepatitis B. Gastroenterology *125*, 1714 - 1722.

116. Lucifora, J., Arzberger, S., Durantel, D., Belloni, L., Strubin, M., Levrero, M., Zoulim, F., Hantz, O., and Protzer, U. (2011). Hepatitis B Virus X protein is

essential to initiate and maintain virus replication after infection. J Hepatol *55*, 996-1003.

117. Mandart, E., Kay, A., and Galibert, F. (1984). Nucleotide sequence of a cloned duck hepatitis B virus genome: comparison with woodchuck and human hepatitis B virus sequences. J Virol *49*, 782-792.

118. Marcellin, P., Asselah, T., and Boyer, N. (2005). Treatment of chronic hepatitis B. J Viral Hepat *12*, 333-345.

119. Marcellin, P., Chang, T.-T., Lim, S.G., Tong, M.J., Sievert, W., Shiffman, M.L., Jeffers, L., Goodman, Z., Wulfsohn, M.S., Xiong, S., Fry, J., and Brosgart, C.L. (2003). Adefovir Dipivoxil for the Treatment of Hepatitis B e Antigen–Positive Chronic Hepatitis B. N Engl J Med *348*, 808-816.

120. Marcellin, P., Gane, E., Buti, M., Afdhal, N., Sievert, W., Jacobson, I.M., Washington, M.K., Germanidis, G., Flaherty, J.F., Schall, R.A., Bornstein, J.D., Kitrinos, K.M., Subramanian, G.M., McHutchison, J.G., and Heathcote, E.J. (2012). Regression of cirrhosis during treatment with tenofovir disoproxil fumarate for chronic hepatitis B: a 5-year open-label follow-up study. Lancet *381*, 468-475.

121. Margeridon-Thermet, S., and Shafer, R.W. (2010). Comparison of the Mechanisms of Drug Resistance among HIV, Hepatitis B, and Hepatitis C. Viruses *2*, 2696-2739.

122. Melegari, M., Bruno, S., and Wands, J.R. (1994). Properties of hepatitis B virus pre-S1 deletion mutants. Virology *199*, 292-300.

123. Melegari, M., Wolf, S.K., and Schneider, R.J. (2005). Hepatitis B virus DNA replication is coordinated by core protein serine phosphorylation and HBx expression. J Virol *79*, 9810-820.

124. Menke, M., Berger, B., and Cowen, L. (2008). Matt: local flexibility aids protein multiple structure alignment. PLoS Comput Biol *4*, e10.

125. Moolla, N., Kew, M., and Arbuthnot, P. (2002). Regulatory elements of hepatitis B virus transcription. J Viral Hepat *9*, 323-331.

126. Mukaide, M., Tanaka, Y., Shin-I, T., Yuen, M.F., Kurbanov, F., Yokosuka, O., Sata, M., Karino, Y., Yamada, G., Sakaguchi, K., Orito, E., Inoue, M., Baqai, S., Lai, C.L., and Mizokami, M. (2010). Mechanism of entecavir resistance of hepatitis B virus with viral breakthrough as determined by long-term clinical assessment and molecular docking simulation. Antimicrob Agents Chemother *54*, 882-89.

127. Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540.

128. Mutimer, D.J., and Oo, Y.H. (2011). Hepatitis B. Medicine *39*, 545 - 549.

129. Myers, R., Clark, C., Khan, A., Kellam, P., and Tedder, R. (2006). Genotyping Hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR. J Gen Virol *87*, 1459-464.

130. Myers, R., Gnaneshan, S., Ijaz, S., Tedder, R., Ramsay, M., Green, J., and HepSEQ Steering Committee (2008). HepSEQ--an integrated hepatitis B epidemiology and sequence analysis platform. Euro Surveill *13*

131. Nassal, M. (1992). The arginine-rich domain of the hepatitis B virus core protein is required for pregenome encapsidation and productive viral positive-strand DNA synthesis but not for virus assembly. J Virol *66*, 4107-116.

132. Nassal, M. (2008). Hepatitis B viruses: reverse transcription a different way. Virus Res *134*, 235-249.

133. NCBI Resource Coordinators (2013). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *41*, D8-D20.

134. Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443 - 453.

135. Nguyen, D.H., Ludgate, L., and Hu, J. (2008). Hepatitis B virus-cell interactions and pathogenesis. J Cell Physiol *216*, 289-294.

136. Norder, H., Couroucé, A.M., and Magnius, L.O. (1992). Molecular basis of hepatitis B virus serotype variations within the four major subtypes. J Gen Virol *73 (Pt 12)*, 3141-45.

137. Norder, H., Couroucé, A.-M., and Magnius, L.O. (1994). Complete Genomes, Phylogenetic Relatedness, and Structural Proteins of Six Strains of the Hepatitis B Virus, Four of Which Represent Two New Genotypes. Virology *198*, 489 - 503.

138. Nowotny, M., Gaidamakov, S.A., Crouch, R.J., and Yang, W. (2005). Crystal Structures of RNase H Bound to an RNA/DNA Hybrid: Substrate Specificity and Metal-Dependent Catalysis. Cell *121*, 1005-016.

139. Nowotny, M., Gaidamakov, S.A., Ghirlando, R., Cerritelli, S.M., Crouch, R.J., and Yang, W. (2007). Structure of Human RNase H1 Complexed with an RNA/DNA Hybrid: Insight into HIV Reverse Transcription. Molecular Cell *28*, 264-276.

140. Okamoto, H., Imai, M., Tsuda, F., Tanaka, T., Miyakawa, Y., and Mayumi, M. (1987). Point mutation in the S gene of hepatitis B virus for a d/y or w/r subtypic change in two blood donors carrying a surface antigen of compound subtype adyr or adwr. J Virol *61*, 3030-34.

141. Okamoto, H., Tsuda, F., Sakugawa, H., Sastrosoewignjo, R.I., Imai, M., Miyakawa, Y., and Mayumi, M. (1988). Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. J Gen Virol *69 (Pt 10)*, 2575-583.

142. de Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E.J., Wensing, A.M., van de Vijver, D.A., Boucher, C.A., Camacho, R., and Vandamme, A.M. (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics *21*, 3797-3800.

143. Omata, M., Hirota, K., and Yokosuka, O. (1986). In vivo study of the mechanism of action of antiviral agents against hepadna virus replication in the liver. Resistance of supercoiled viral DNA. J Hepatol *3 Suppl 2*, S49-S55.

144. Ono, Y., Onda, H., Sasada, R., Igarashi, K., Sugino, Y., and Nishioka, K. (1983). The complete nucleotide sequences of the cloned hepatitis B virus DNA; subtype adr and adw. Nucleic Acids Res *11*, 1747-757.

145. Packianathan, C., Katen, S.P., Dann, C.E., and Zlotnick, A. (2010). Conformational changes in the hepatitis B virus core protein are consistent with a role for allostery in virus assembly. J Virol *84*, 1607-615.

146. Pagani, I., Liolios, K., Jansson, J., Chen, I.M., Smirnova, T., Nosrat, B., Markowitz, V.M., and Kyrpides, N.C. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res *40*, D571-79.

147. Panjaworayan, N., Roessner, S.K., Firth, A.E., and Brown, C.M. (2007). HBVRegDB: annotation, comparison, detection and visualization of regulatory elements in hepatitis B virus sequences. Virol J *4*, 136.

148. Pawlotsky, J.-M. (2005). The concept of hepatitis B virus mutant escape. Journal of Clinical Virology *34, Supplement 1*, S125 - S129.

149. Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J., and Orengo, C. (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res *33*, D247-251.

150. Pearson, W.R. (1998). Empirical statistical estimates for sequence similarity searches. J Mol Biol *276*, 71-84.

151. Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A *85*, 2444-48.

152. Pollack, J.R., and Ganem, D. (1994). Site-specific RNA binding by a hepatitis B virus reverse transcriptase initiates two distinct reactions: RNA packaging and DNA synthesis. J Virol *68*, 5579-587.

153. Potenza, N., Salvatore, V., Raimondo, D., Falanga, D., Nobile, V., Peterson, D.L., and Russo, A. (2007). Optimized expression from a synthetic gene of an untagged RNase H domain of human hepatitis B virus polymerase which is enzymatically active. Protein Expression and Purification *55*, 93 - 99.

154. Protein Data Bank (1971). Protein Data Bank. Nature New Biol *233*, 223.

155. Protein Data Bank (1973). Protein Data Bank. Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry *29*, 1746.

156. Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *35*, D61-65.

157. Rhee, S.Y., Margeridon-Thermet, S., Nguyen, M.H., Liu, T.F., Kagan, R.M., Beggel, B., Verheyen, J., Kaiser, R., and Shafer, R.W. (2010). Hepatitis B virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery. Antiviral Res *88*, 269-275.

158. Rost, B., and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol *232*, 584-599.

159. Rozanov, M., Plikat, U., Chappey, C., Kochergin, A., and Tatusova, T. (2004). A web-based genotyping resource for viral sequences. Nucleic Acids Res *32*, W654-59.

160. Ruiz-Opazo, N., Chakraborty, P.R., and Shafritz, D.A. (1982). Evidence for supercoiled hepatitis B virus DNA in chimpanzee liver and serum Dane particles: possible implications in persistent HBV infection. Cell *29*, 129-136.

161. Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., and Belshaw, R. (2010). Viral mutation rates. J Virol *84*, 9733-748.

162. Sarafianos, S.G., Das, K., Tantillo, C., Clark, A.D., Ding, J., Whitcomb, J.M., Boyer, P.L., Hughes, S.H., and Arnold, E. (2001). Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. EMBO J *20*, 1449-461.

163. Schaefer, S. (2007). Hepatitis B virus taxonomy and hepatitis B virus genotypes. World J Gastroenterol *13*, 14-21.

164. Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res *18*, 6097-6100.

165. Schultz, A.K., Bulla, I., Abdou-Chekaraou, M., Gordien, E., Morgenstern, B., Zoaulim, F., Dény, P., and Stanke, M. (2012). jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. Nucleic Acids Res *40*, W193-98.

166. Schultz, A.K., Zhang, M., Leitner, T., Kuiken, C., Korber, B., Morgenstern, B., and Stanke, M. (2006). A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. BMC Bioinformatics *7*, 265.

167. Schultz, S.J., and Champoux, J.J. (2008). RNase H activity: structure, specificity, and function in reverse transcription. Virus Res *134*, 86-103.

168. Seeger, C., and Mason, W.S. (2000). Hepatitis B virus biology. Microbiol Mol Biol Rev *64*, 51-68.

169. Shannon, C.E., Weaver, W., Blahut, R.E., and Hajek, B. (1948). The mathematical theory of communication (University of Illinois press Urbana).

170. Sharon, A., and Chu, C.K. (2008). Understanding the molecular basis of HBV drug resistance by molecular modeling. Antiviral Res *80*, 339-353.

171. Shaw, T., Bartholomeusz, A., and Locarnini, S. (2006). HBV drug resistance: Mechanisms, detection and interpretation. J Hepatol *44*, 593 - 606.

172. Sheldon, J., Camino, N., Rodés, B., Bartholomeusz, A., Kuiper, M., Tacke, F., Núñez, M., Mauss, S., Lutz, T., Klausen, G., Locarnini, S., and Soriano, V. (2005). Selection of hepatitis B virus polymerase mutations in HIV-coinfected patients treated with tenofovir. Antivir Ther *10*, 727-734.

173. Shin-I, T., Tanaka, Y., Tateno, Y., and Mizokami, M. (2008). Development and public release of a comprehensive hepatitis virus database. Hepatol Res *38*, 234-243.

174. Simmonds, P., and Midgley, S. (2005). Recombination in the genesis and evolution of hepatitis B virus genotypes. J Virol *79*, 15467-476.

175. Smith, S., and Waterman, W. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195 - 197.

176. Soussan, P., Garreau, F., Zylberberg, H., Ferray, C., Brechot, C., and Kremsdorf, D. (2000). In vivo expression of a new hepatitis B virus protein encoded by a spliced RNA. J Clin Invest *105*, 55-60.

177. Soussan, P., Tuveri, R., Nalpas, B., Garreau, F., Zavala, F., Masson, A., Pol, S., Brechot, C., and Kremsdorf, D. (2003). The expression of hepatitis B spliced protein (HBSP) encoded by a spliced hepatitis B virus RNA is associated with viral replication and liver fibrosis. J Hepatol *38*, 343-48.

178. Stuyver, L., De Gendt, S., Van Geyt, C., Zoulim, F., Fried, M., Schinazi, R.F., and Rossau, R. (2000). A new genotype of hepatitis B virus: complete genome and phylogenetic relatedness. J Gen Virol *81*, 67-74.

179. Stuyver, L.J., Locarnini, S.A., Lok, A., Richman, D.D., Carman, W.F., Dienstag, J.L., and Schinazi, R.F. (2001). Nomenclature for antiviral-resistant human hepatitis B virus mutations in the polymerase region. Hepatology *33*, 751-57.

180. Sugauchi, F., Orito, E., Ichida, T., Kato, H., Sakugawa, H., Kakumu, S., Ishida, T., Chutaputti, A., Lai, C.L., Gish, R.G., Ueda, R., Miyakawa, Y., and Mizokami, M. (2003). Epidemiologic and virologic characteristics of hepatitis B virus genotype B having the recombination with genotype C. Gastroenterology *124*, 925-932.

181. Summers, J., and Mason, W.S. (1982). Replication of the genome of a hepatitis B--like virus by reverse transcription of an RNA intermediate. Cell *29*, 403-415.

182. Tabor, E., Gerety, R.J., Smallwood, L.A., and Barker, L.F. (1977). Coincident hepatitis B surface antigen and antibodies of different subtypes in human serum. J Immunol *118*, 369-370.

183. Tadokoro, T., and Kanaya, S. (2009). Ribonuclease H: molecular diversities, substrate binding domains, and catalytic mechanism of the prokaryotic enzymes. FEBS Journal *276*, 1482-493.

184. Tang, H., Oishi, N., Kaneko, S., and Murakami, S. (2006). Molecular functions and biological roles of hepatitis B virus x protein. Cancer Sci *97*, 977-983.

185. Tenney, D.J., Levine, S.M., Rose, R.E., Walsh, A.W., Weinheimer, S.P., Discotto, L., Plym, M., Pokornowski, K., Yu, C.F., Angus, P., Ayres, A., Bartholomeusz, A., Sievert, W., Thompson, G., Warner, N., Locarnini, S., and Colonno, R.J. (2004). Clinical emergence of entecavir-resistant hepatitis B virus requires additional substitutions in virus already resistant to Lamivudine. Antimicrob Agents Chemother *48*, 3498-3507.

186. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-680.

187. Tong, S.P., Li, J.S., Vitvitski, L., and Trépo, C. (1990). Active hepatitis B virus replication in the presence of anti-HBe is associated with viral variants containing an inactive pre-C region. Virology *176*, 596-603.

188. Tuske, S., Sarafianos, S.G., Clark, A.D., Ding, J., Naeger, L.K., White, K.L., Miller, M.D., Gibbs, C.S., Boyer, P.L., Clark, P., Wang, G., Gaffney, B.L., Jones, R.A., Jerina, D.M., Hughes, S.H., and Arnold, E. (2004). Structures of HIV-1 RT-DNA complexes before and after incorporation of the anti-AIDS drug tenofovir. Nat Struct Mol Biol *11*, 469-474.

189. UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res *40*, D71-75.

190. Villeneuve, J.-P., Durantel, D., Durantel, S., Westland, C., Xiong, S., Brosgart, C.L., Gibbs, C.S., Parvaz, P., Werle, B., Trépo, C., and Zoulim, F. (2003). Selection of a hepatitis B virus strain resistant to adefovir in a liver transplantation patient. J Hepatol *39*, 1085 - 1089.

191. Villet, S., Billioud, G., Pichoud, C., Lucifora, J., Hantz, O., Sureau, C., Dény, P., and Zoulim, F. (2009). In Vitro Characterization of Viral Fitness of Therapy-Resistant Hepatitis B Variants. Gastroenterology *136*, 168 - 176.e2.

192. Villet, S., Pichoud, C., Billioud, G., Barraud, L., Durantel, S., Trépo, C., and Zoulim, F. (2008). Impact of hepatitis B virus rtA181V/T mutants on hepatitis B treatment failure. J Hepatol *48*, 747-755.

193. Villet, S., Pichoud, C., Villeneuve, J.P., Trépo, C., and Zoulim, F. (2006). Selection of a multiple drug-resistant hepatitis B virus strain in a liver-transplanted patient. Gastroenterology *131*, 1253-261.

194. Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory *13*, 260-69.

195. Wallace, A.C., Laskowski, R.A., and Thornton, J.M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng *8*, 127-134.

196. Wang, G.H., and Seeger, C. (1992). The reverse transcriptase of hepatitis B virus acts as a protein primer for viral DNA synthesis. Cell *71*, 663-670.

197. Warner, N., and Locarnini, S. (2008). The antiviral drug selected hepatitis B virus rtA181T/sW172* mutant has a dominant negative secretion defect and alters the typical profile of viral rebound. Hepatology *48*, 88-98.

198. Wilbur, W.J., and Lipman, D.J. (1983). Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci U S A *80*, 726-730.

199. Wynne, S.A., Crowther, R.A., and Leslie, A.G.W. (1999). The Crystal Structure of the Human Hepatitis B Virus Capsid. Molecular Cell *3*, 771 - 780.

200. Yadav, V., and Chu, C.K. (2004). Molecular mechanisms of adefovir sensitivity and resistance in HBV polymerase mutants: a molecular dynamics study. Bioorg Med Chem Lett *14*, 4313-17.

201. Yeh, C.T., Chien, R.N., Chu, C.M., and Liaw, Y.F. (2000). Clearance of the original hepatitis B virus YMDD-motif mutants with emergence of distinct lamivudine-resistant mutants during prolonged lamivudine therapy. Hepatology *31*, 1318-326.

202. Yim, H.J., Hussain, M., Liu, Y., Wong, S.N., Fung, S.K., and Lok, A.S.F. (2006). Evolution of multi-drug resistant hepatitis B virus during sequential therapy. Hepatology *44*, 703-712.

203. Yuen, L.K., Ayres, A., Littlejohn, M., Colledge, D., Edgely, A., Maskill, W.J., Locarnini, S.A., and Bartholomeusz, A. (2007). SeqHepB: a sequence analysis program and relational database system for chronic hepatitis B. Antiviral Res *75*, 64-74.

204. Yuh, C.H., Chang, Y.L., and Ting, L.P. (1992). Transcriptional regulation of precore and pregenomic RNAs of hepatitis B virus. J Virol *66*, 4073-084.

205. Zheng, J., Schödel, F., and Peterson, D.L. (1992). The structure of hepadnaviral core antigens. Identification of free thiols and determination of the disulfide bonding pattern. J Biol Chem *267*, 9422-29.

206. Zhou, S., and Standring, D.N. (1992). Hepatitis B virus capsid particles are assembled from core-protein dimer precursors. Proc Natl Acad Sci U S A *89*, 10046-050.

207. Zlotnick, A., Cheng, N., Conway, J.F., Booy, F.P., Steven, A.C., Stahl, S.J., and Wingfield, P.T. (1996). Dimorphism of hepatitis B virus capsids is strongly influenced by the C-terminus of the capsid protein. Biochemistry *35*, 7412-421.

208. Zoulim, F. (2004). Mechanism of viral persistence and resistance to nucleoside and nucleotide analogs in chronic hepatitis B virus infection. Antiviral Res *64*, 1-15.

209. Zoulim, F., and Locarnini, S. (2009). Hepatitis B virus resistance to nucleos(t)ide analogues. Gastroenterology *137*, 1593-608.e1-2.

210. Zoulim, F., and Seeger, C. (1994). Reverse transcription in hepatitis B viruses is primed by a tyrosine residue of the polymerase. J Virol *68*, 6-13.

Références personnelles

Publications

HBVdb: a knowledge database for Hepatitis B Virus.
Hayer J., Jadeau F., Deléage G., Kay A., Zoulim F., Combet C.
Nucleic Acid Res. 2012 (Advance access published November 3, 2012)

Mutations that alter use of hepatitis C virus cell entry factors mediate escape from neutralizing antibodies.
Fofana I., Fafi-Kremer S., Carolla P., Fauvelle C., Zahid M.N., Turek M., Heydmann L., Cury K., Hayer J., Combet C., Cosset F.L., Pietschmann T., Hiet M.S., Bartenschlager R., Habersetzer F., Doffoël M., Keck Z.Y., Foung S.K., Zeisel M.B., Stoll-Keller F., Baumert T.F.
Gastroenterology 2012 July 143:1, 223-233.e9

Communications Orales

HBVdb: Applications to structural and functional analysis of HBV Polymerase
Hayer J., Jadeau F., Zoulim F., Deléage G., Kay A., Durantel D., Combet C.
12ème réunion du réseau national hépatites de l'ANRS, 2012, January 26-27, Paris, France.

Variabilité génétique de la RNase H du virus de l'hépatite B par pyroséquençage haut débit appliquée à la confirmation et au perfectionnement d'un modèle structural.
Rodriguez C., Hayer J., Germanidis G., Combet C., Pawlotsky JM.
12ème Réunion du réseau national hépatites de l'ANRS, **2012**, 26-27 janvier, Paris, France.

Molecular modeling of the Hepatitis B Virus polymerase reverse transcriptase and ribonuclease H domains.
Hayer J., Durantel D., Deléage G., Zoulim F., Combet C.
International Meeting on Molecular Biology of Hepatitis B Viruses, 2011, October 9-12, Lake Buena Vista, Florida, USA.

Association de l'échappement viral à l'utilisation des facteurs cellulaires d'entrée au cours de l'infection par le virus de l'hépatite C in vivo.
Fafi-Kremer S., Fofana I., Zahid M.N., Fauvelle C., Turek M., Heydmann L., Bastien-Valle M., Hayer J., Combet C., Cosset F.L., Pietschmann T., Bartenschlager R., Habersetzer F., Keck Z.-Y., Foung S.K.H., Zeisel M. B., Stoll-Keller F., Baumert T. *69èmes Journées scientifiques de l'AFEF, 2011, 28 septembre-1 octobre, Paris, France.*

HBVdb: a Knowledge Database for the Hepatitis B Virus.
Hayer J., Jadeau F., Deléage G., Combet C.
16ème journée scientifique de l'EDISS (Ecole Doctorale Interdisciplinaire Sciences Santé), 2011, March 24, Lyon, France.

HBVdb: une base de connaissances du virus de l'hépatite B.
Hayer J., Jadeau F., Deléage G., Combet C.
11ème réunion du réseau national hépatites de l'ANRS, 2011, January 27-28, Paris, France.

Posters

HBVdb: A knowledge database for the Hepatitis B Virus.
Hayer J., Jadeau F., Deléage G., Kay A., Zoulim F., Combet C.
International Meeting on Molecular Biology of Hepatitis B Viruses, 2012, September 22-25, University of Oxford, Oxford, UK.

A single residue in hepatitis C virus glycoprotein E2 modulates CD81 receptor dependency and confers resistance to neutralizing antibodies in vivo.
Fafi-Kremer S., Fofana I., Zahid M.N., Fauvelle C., Turek M., Heydmann L., Bastien-Valle M., Hayer J., Combet C., Cosset F.L., Pietschmann T., Bartenschlager R., Habersetzer F., Doffoel M., Keck Z.-Y., Foung S.K.H., Zeisel M. B., Stoll-Keller F., Baumert T.
18th Internatitional Symposium on Hepatitis C Virus and Related Viruses, 2011, September 8-12th, Seattle, USA.

HBVdb: une base de connaissances du virus de l'hépatite B.
Hayer J., Jadeau F., Deléage G., Combet C.
XIIIèmes Journées Francophones de Virologie, 2011, April 28-29, Paris, France.

HBVdb: a Knowledge Database for the Hepatitis B Virus.
Hayer J., Jadeau F., Deléage G., Combet C.
GDRE comparative genomics, 2010, November 16-17, Barcelona, Spain.

HBVdb: a Knowledge Database for the Hepatitis B Virus.
Hayer J., Jadeau F., Deléage G., Combet C.
International Meeting on Molecular Biology of Hepatitis B Viruses, 2010, October 9-13, Taipei, Taiwan.

HBVdb: a Knowledge Database for the Hepatitis B Virus.
Hayer J., Jadeau F., Deléage G., Combet C.
JOBIM, 2010, September 7-9, Montpellier, France.

Annexes

Annexe 1

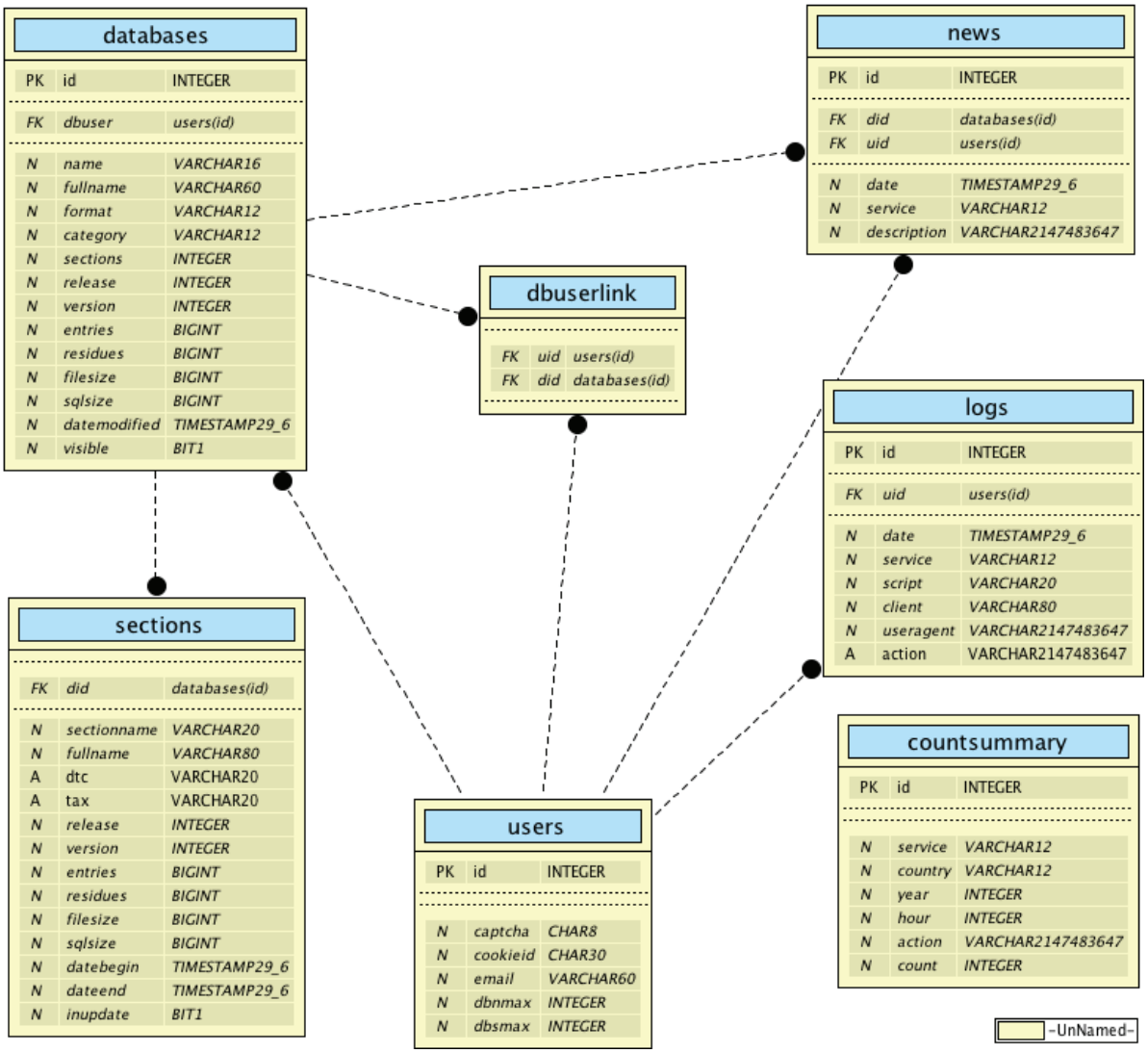
```
ID   X02763; SV 1; circular; genomic DNA; STD; VRL; 3221 BP.
XX
AC   X02763;
XX
DT   18-NOV-1986 (Rel. 10, Created)
DT   18-APR-2005 (Rel. 83, Last updated, Version 10)
XX
DE   Hepatitis b virus genome (serotype adw2)
XX
KW   antigen; core antigen; genome; overlapping genes; surface antigen;
KW   unidentified reading frame.
XX
OS   Hepatitis B virus
OC   Viruses; Retro-transcribing viruses; Hepadnaviridae; Orthohepadnavirus.
XX
RN1
RP   1-3221
RA   Valenzuela P., Quiroga M., Zalvidar J., Gray P., Rutter W.J.;
RT   "The nucleotide sequence of the hepatitis B viral genome and the
RT   identification of the major viral genes";
RL   (in) Fields B.N., Jaenisch R., Fox C.F. (Eds.);
RL   ANIMAL VIRUS GENETICS:57-70;
RL   Academic Press, New York (1980)
XX
DR   GOA; P0C625.
DR   InterPro; IPR002006; Viral_capsid_core_Hepatitis.
DR   InterPro; IPR013195; Hepatitis_B_virus_capsid_N.
DR   RFAM; RF01047; HBV_epsilon.
DR   UniProtKB/Swiss-Prot; P0C625; HBEAG_HBVA3.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..3221
FT                   /organism="Hepatitis B virus"
FT                   /strain="subtype adw2"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:10407"
FT   misc_feature    8..8
FT                   /note="pot. alternative transcription start site for core
FT                   antigen"
FT   misc_feature    54..82
FT                   /note="palindrome sequence"
FT   RBS             70..73
FT                   /note="pot. ribosome binding site"
FT   CDS             89..646
FT                   /note="core antigen (aa 1-185)"
FT                   /db_xref="GOA:P03148"
FT                   /db_xref="InterPro:IPR002006"
FT                   /db_xref="UniProtKB/Swiss-Prot:P03148"
FT                   /protein_id="CAA26537.1"
FT                   /translation="MDIDPYKEFGATVELLSFLPSDFFPSVRDLLDTASALYREALESP
FT                   EHCSPHHTALRQAILCWGELMTLATWVGNNLEDPASRDLVVNYVNTNVGLKIRQLLWFH
FT                   ISCLTFGRETVLEYLVSFGVWIRTPPAYRPPNAPILSTLPETTVVRRRDRGRSPRRRTP
FT                   SPRRRRSPSPRRRRSQSRESQC"
FT   promoter        104..110
FT                   /note="TATA-like sequence, pot. altern. promoter for A"
FT   misc_feature    308..313
FT                   /note="pot. polyadenylation signal for transcript A"
FT   promoter        400..404
FT                   /note="pot. TATA-box; altern. promoter for A and surface
FT                   antigen"
FT   CDS             495..3032
FT                   /note="unidentified reading frame A (aa 1-845)"
FT                   /db_xref="GOA:P03159"
FT                   /db_xref="InterPro:IPR000201"
FT                   /db_xref="InterPro:IPR000477"
FT                   /db_xref="InterPro:IPR001462"
```

```
FT                   /db_xref="UniProtKB/Swiss-Prot:P03159"
FT                   /protein_id="CAA26538.1"
FT                   /translation="MPLSYQHFRKLLLLDDGTEAGPLEEELPRLADADLHRRVAEDLNL
FT                   GNLNVSIPWTHKVGNFTGLYSSTVPIFNPEWQTPSFPKIHLQEDIINRCQQFVGPLTVN
FT                   EKRRLKLIMPARFYPTHTKYLPLDKGIKPYYPDQVVNHYFQTRHYLHTLWKAGILYKRE
FT                   TTRSASFCGSPYSWEQELQHGRLVIKTSQRHGDESFCSQSSGILSRSSVGPCIRSQLKQ
FT                   SRLGLQPRQGRLASSQPSRSGSIRAKAHPSTRRYFGVEPSGSGHIDHSVNNSSSCLHQS
FT                   AVRKAAYSHLSTSKRQSSSGHAVEFHCLPPNSAGSQSQGSVSSCWWLQFRNSKPCSEYC
FT                   LSHLVNLREDWGPCDEHGEHHIRIPRTPARVTGGVFLVDKNPHNTAESRLVVDFSQFSR
FT                   GISRVSWPKFAVPNLQSLTNLLSSNLSWLSLDVSAAFYHIPLHPAAMPHLLIGSSGLSR
FT                   YVARLSSNSRINNNQYGTMQNLHDSCSRQLYVSLMLLYKTYGWKLHLYSHPIVLGFRKI
FT                   PMGVGLSPFLLAQFTSAICSVVRRAFPHCLAFSYMDDVVLGAKSVQHRESLYTAVTNFL
FT                   LSLGIHLNPNKTKRWGYSLNFMGYIIGSWGTLPQDHIVQKIKHCFRKLPVNRPIDWKVC
FT                   QRIVGLLGFAAPFTQCGYPALMPLYACIQAKQAFTFSPTYKAFLSKQYMNLYPVARQRP
FT                   GLCQVFADATPTGWGLAIGHQRMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSV
FT                   VLSRKYTSFPWLLGCTANWILRGTSFVYVPSALNPADDPSRGRLGLSRPLLRLPFQPTT
FT                   GRTSLYAVSPSVPSHLPVRVHFASPLHVAWRPP"
FT   misc_feature    822..828
FT                   /note="pot. polyadenylation signal for core antigen
FT                   transcript"
FT   promoter        970..977
FT                   /note="pot. TATA-box, altern. promoter for surface antigen"
FT   CDS             1042..2244
FT                   /note="surface antigen (aa 1-400)"
FT                   /db_xref="GOA:P03141"
FT                   /db_xref="InterPro:IPR000349"
FT                   /db_xref="UniProtKB/Swiss-Prot:P03141"
FT                   /protein_id="CAA26539.1"
FT                   /translation="MGGWSSKPRKGMGTNLSVPNPLGFFPDHQLDPAFGANSNNPDWDF
FT                   NPVKDDWPAANQVGVGAFGPRLTPPHGGILGWSPQAQGILTTVSTIPPPASTNRQSGRQ
FT                   PTPISPPLRDSHPQAMQWNSTAFHQTLQDPRVRGLYLPAGGSSSGTVNPAPNIASHISS
FT                   ISARTGDPVTNMENITSGFLGPLLVLQAGFFLLTRILTIPQSLDSWWTSLNFLGGSPVC
FT                   LGQNSQSPTSNHSPTSCPPICPGYRWMCLRRFIIFLFILLLCLIFLLVLLDYQGMLPVC
FT                   PLIPGSTTTSTGPCKTCTTPAQGNSMFPSCCCTKPTDGNCTCIPIPSSWAFAKYLWEWA
FT                   SVRFSWLSLLVPFVQWFVGLSPTVWLSAIWMMWYWGPSLYSIVSPFIPLLPIFFCLWVY
FT                   I"
FT   misc_feature    2250..2254
FT                   /note="pot. altern. polyadenylation signal for surface
FT                   antigen transcript"
FT   CDS             join(2783..3221,1..26)
FT                   /note="unidentified reading frame B"
FT                   /db_xref="GOA:P69713"
FT                   /db_xref="InterPro:IPR000236"
FT                   /db_xref="UniProtKB/Swiss-Prot:P69713"
FT                   /protein_id="CAA26540.1"
FT                   /translation="MAARLYCQLDPSRDVLCLRPVGAESRGRPLSGPLGTLSSPSPSAV
FT                   PADHGAHLSLRGLPVCAFSSAGPCALRFTSARCMETTVNAHQILPKVLHKRTLGLPAMS
FT                   TTDLEAYFKDCVFKDWEELGEEIRLKVFVLGGCRHKLVCAPAPCNFFTSA"
FT   misc_feature    2767..2772
FT                   /note="pot. altern. polyadenylation signal for surface
FT                   antigen transcript"
FT   promoter        3060..3067
FT                   /note="TATA-like sequence, pot. promoter for core antigen"
XX
SQ   Sequence 3221 BP; 740 A; 868 C; 709 G; 904 T; 0 other;
     catgcaactt tttcacctct gcctaatcat ctcttgtaca tgtcccactg ttcaagcctc        60
     caagctgtgc cttgggtggc tttggggcat ggacattgac ccttataaag aatttggagc       120
     tactgtggag ttactctcgt ttttgccttc tgacttcttt ccttccgtca gagatctcct       180
     agacaccgcc tcagctctgt atcgagaagc cttagagtct cctgagcatt gctcacctca       240
     ccatactgca ctcaggcaag ccattctctg ctgggggtgaa ttgatgactc tagctacctg      300
     ggtgggtaat aatttggaag atccagcatc tagggatctt gtagtaaatt atgttaatac       360
     taacgtgggt ttaaagatca ggcaactatt gtggtttcat atatcttgcc ttacttttgg       420
     aagagagact gtacttgaat atttggtctc tttcggagtg tggattcgca ctcctccagc       480
     ctatagacca ccaaatgccc ctatcttatc aacacttccg gaaactactg ttgttagacg       540
     acgggaccga ggcaggtccc ctagaagaag aactccctcg cctcgcagac gcagatctcc       600
     atcgccgcgt cgcagaagat ctcaatctcg ggaatctcaa tgttagtatt ccttggactc       660
     ataaggtggg aaactttacg gggctttatt cctctacagt acctatcttt aatcctgaat       720
     ggcaaactcc ttcctttcct aagattcatt tacaagagga cattattaat aggtgtcaac       780
     aatttgtggg ccctctcact gtaaatgaaa agagaagatt gaaattaatt atgcctgcta       840
     gattctatcc tacccacact aaatatttgc ccttagacaa aggaattaaa ccttattatc       900
```

```
  cagatcaggt agttaatcat tacttccaaa ccagacatta tttacatact ctttggaagg       960
  ctggtattct atataagcgg gaaaccacac gtagcgcatc attttgcggg tcaccatatt      1020
  cttgggaaca agagctacag catgggaggt tggtcatcaa aacctcgcaa aggcatgggg      1080
  acgaatcttt ctgttcccaa tcctctggga ttctttcccg atcatcagtt ggaccctgca      1140
  ttcggagcca actcaaacaa tccagattgg gacttcaacc ccgtcaagga cgactggcca      1200
  gcagccaacc aagtaggagt gggagcattc gggccaaggc tcacccctcc acacggcggt      1260
  attttggggt ggagccctca ggctcagggc atattgacca cagtgtcaac aattcctcct      1320
  cctgcctcca ccaatcggca gtcaggaagg cagcctactc ccatctctcc acctctaaga      1380
  gacagtcatc ctcaggccat gcagtggaat tccactgcct tccaccaaac tctgcaggat      1440
  cccagagtca ggggtctgta tcttcctgct ggtggctcca gttcaggaac agtaaaccct      1500
  gctccgaata ttgcctctca catctcgtca atctccgcga ggactgggga ccctgtgacg      1560
  aacatggaga acatcacatc aggattccta ggacccctgc tcgtgttaca ggcggggttt      1620
  ttcttgttga caagaatcct cacaataccg cagagtctag actcgtggtg gacttctctc      1680
  aattttctag ggggatctcc cgtgtgtctt ggccaaaatt cgcagtcccc aacctccaat      1740
  cactcaccaa cctcctgtcc tccaatttgt cctggttatc gctggatgtg tctgcggcgt      1800
  tttatcatat tcctcttcat cctgctgcta tgcctcatct tcttattggt tcttctggat      1860
  tatcaaggta tgttgcccgt ttgtcctcta attccaggat caacaacaac cagtacggga      1920
  ccatgcaaaa cctgcacgac tcctgctcaa ggcaactcta tgtttccctc atgttgctgt      1980
  acaaaaccta cggatggaaa ttgcacctgt attcccatcc catcgtcctg ggctttcgca      2040
  aaatacctat gggagtgggc ctcagtccgt ttctcttggc tcagtttact agtgccattt      2100
  gttcagtggt tcgtagggct ttcccccact gtttggcttt cagctatatg gatgatgtgg      2160
  tattgggggc caagtctgta cagcatcgtg agtcccttta taccgctgtt accaattttc      2220
  ttttgtctct gggtatacat ttaaacccta acaaaacaaa aagatggggt tattccctaa      2280
  acttcatggg ctacataatt ggaagttggg gaactttgcc acaggatcat attgtacaaa      2340
  agatcaaaca ctgttttaga aaacttcctg ttaacaggcc tattgattgg aaagtatgtc      2400
  aaagaattgt gggtcttttg ggctttgctg ctccatttac acaatgtgga tatcctgcct      2460
  taatgccttt gtatgcatgt atacaagcta aacaggcttt cactttctcg ccaacttaca      2520
  aggcctttct aagtaaacag tacatgaacc tttaccccgt tgctcggcaa cggcctggtc      2580
  tgtgccaagt gtttgctgac gcaaccccca ctggctgggg cttggccata ggccatcagc      2640
  gcatgcgtgg aacctttgtg gctcctctgc cgatccatac tgcggaactc ctagccgctt      2700
  gttttgctcg cagccggtct ggagcaaagc tcatcggaac tgacaattct gtcgtcctct      2760
  cgcggaaata tacatcgttt ccatggctgc taggctgtac tgccaactgg atccttcgcg      2820
  ggacgtcctt tgtttacgtc ccgtcggcgc tgaatcccgc ggacgacccc tctcggggcc      2880
  gcttgggact ctctcgtccc cttctccgtc tgccgttcca gccgaccacg gggcgcacct      2940
  ctctttacgc ggtctccccg tctgtgcctt ctcatctgcc ggtccgtgtg cacttcgctt      3000
  cacctctgca cgttgcatgg agaccaccgt gaacgcccat cagatcctgc ccaaggtctt      3060
  acataagagg actcttggac tcccagcaat gtcaacgacc gaccttgagg cctacttcaa      3120
  agactgtgtg tttaaggact gggaggagct gggggaggag attaggttaa aggtctttgt      3180
  attaggaggc tgtaggcaca aattggtctg cgcaccagca c                         3221
//
```

Annexe 1 : Exemple d'entrée ENA (AC=X02763).

Annexe 2



Annexe 2 : Schéma relationnel de la base de données *metadb*.

Annexe 3

Genotype	Accession number	Genome length	Isolate	Reference
Genotype A				
A	EU054331	3221	n.a.	Makuwa M, et al.(2008)
A	X02763	3221	n.a.	Valenzuela P, et al.(1980)
Genotype B				
B	AB073853	3215	HBV-Ry24	Sugauchi F, et al.(2002)
B	DQ463795	3215	928-9_2004	Osiowy C, et al.(2006)
Genotype C				
C	GQ924620	3215	M38	Meldal BH, et al.(2011)
C	X75665	3215	HMA	Norder H, et al.(1994)
Genotype D				
D	AB330367	3182	TAJ14HBV	Khan A, et al.(2008)
D	GQ205378	3182	T1181	Ghosh S, et al.(2010)
Genotype E				
E	AB106564	3212	HBV-GA325	Unpublished
E	EU239220	3212	PW5	Zahn A, et al.(2008)
Genotype F				
F	AY090458	3215	n.a.	Arauz-Ruiz P, et al.(2002)
F	X75658	3215	Fou	Norder H, et al.(1994)
Genotype G				
G	AB064313	3248	n.a.	Kato H, et al.(2002)
G	AF160501	3248	n.a.	Stuyver L, et al.(2000)
Genotype H				
H	AY090454	3215	n.a.	Arauz-Ruiz P, et al.(2002)
H	FJ356716	3215	CL150171	Flichman DM, et al.(2009)

Annexe 3 : Tableau des 16 génomes complets de référence (image de HBVdb).

Annexe 4

PRABI_name* des FT *CDS	***PRABI_name* des FT *mat_peptide***	**Noms habituellement employés**
PreC	HBe	Precore, HBe antigen, PreC
C	HBc	Core, Capsid, HBc antigen, C
X	HBx	X protein, HBx, X
PreS1	LHBs	Large surface protein, PreS1, LHBs
PreS2	MHBs	Middle surface protein, PreS2, MHBs
S	SHBs	Small surface protein, S, HBs antigen
P	Pol	Polymerase, Reverse Transcriptase
SP	HBSP	Hepatitis B Spliced Protein, HBSP

Annexe 4 : Tableau des *PRABI_name* utilisés pour annoter les CDS et les *mat_peptides* dans HBVdb

Annexe 5

```
ID   X02763; SV 1; circular; genomic DNA; STD; VRL; 3221 BP.
XX
AC   X02763;
XX
XX
DT   18-NOV-1986 (Rel. 10, Created)
DT   29-NOV-2012 (Rel. 6, Last updated, Version 0)
XX
DE   Hepatitis B Virus genotype A, complete genome.
XX
KW   HBe; HBc; HBx; LHBs; MHBs; SHBs; Pol; HBSP; complete genome.
XX
OS   Hepatitis B Virus genotype A
OC   Viruses; Retro-transcribing viruses; Hepadnaviridae; Orthohepadnavirus.
XX
RN1
RP   1-3221
RA   Valenzuela P., Quiroga M., Zalvidar J., Gray P., Rutter W.J.;
RT   "The nucleotide sequence of the hepatitis B viral genome and the
RT   identification of the major viral genes";
RL   (in) Fields B.N., Jaenisch R., Fox C.F. (Eds.);
RL   ANIMAL VIRUS GENETICS:57-70; Academic Press, New York (1980)
XX
CC   The data provided in this entry have been computed thanks to
CC   the Hepatitis B Virus database (HBVdb) annotation alogrithms.
CC   HBVdb is available at http://hbvdb.ibcp.fr.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..3221
FT                   /PRABI_genotype="n.a.:A:A"
FT                   /db_xref="taxon:10407"
FT                   /db_xref="EMBL:X02763"
FT                   /db_xref="hbvdb:X02763"
FT   CDS             1814..2458
FT                   /PRABI_name="PreC"
FT                   /locus_tag="HBVORF02"
FT                   /codon_start="1"
FT                   /translation="MQLFHLCLIISCTCPTVQASKLCLGWLWGMDIDPYKEFGATVELL
FT                   SFLPSDFFPSVRDLLDTASALYREALESPEHCSPHHTALRQAILCWGELMTLATWVGNN
FT                   LEDPASRDLVVNYVNTNVGLKIRQLLWFHISCLTFGRETVLEYLVSFGVWIRTPPAYRP
FT                   PNAPILSTLPETTVVRRRDRGRSPRRRTPSPRRRRSPSPRRRRSQSRESQC"
FT   mat_peptide     1814..2458
FT                   /function="HBe/External core antigen coding sequence"
FT                   /locus_tag="HBVORF02"
FT                   /product="External core antigen"
FT                   /PRABI_name="HBe"
FT                   /PRABI_prodft=(pos:1..214, chain, "HBe antigen")
FT   CDS             1901..2458
FT                   /PRABI_name="C"
FT                   /locus_tag="HBVORF12"
FT                   /codon_start="1"
FT                   /db_xref="UniProtKB/Swiss-Prot:P03148"
FT                   /translation="MDIDPYKEFGATVELLSFLPSDFFPSVRDLLDTASALYREALESP
FT                   EHCSPHHTALRQAILCWGELMTLATWVGNNLEDPASRDLVVNYVNTNVGLKIRQLLWFH
FT                   ISCLTFGRETVLEYLVSFGVWIRTPPAYRPPNAPILSTLPETTVVRRRDRGRSPRRRTP
FT                   SPRRRRSPSPRRRRSQSRESQC"
FT   mat_peptide     1901..2458
FT                   /function="Core protein coding sequence"
FT                   /locus_tag="HBVORF12"
FT                   /product="Core protein"
FT                   /PRABI_name="HBc"
FT                   /PRABI_prodft=(pos:1..185, chain, "Core protein")
```

```
FT   CDS             1374..1838
FT                   /PRABI_name="X"
FT                   /locus_tag="HBVORF03"
FT                   /codon_start="1"
FT                   /db_xref="UniProtKB/Swiss-Prot:P69713"
FT                   /translation="MAARLYCQLDPSRDVLCLRPVGAESRGRPLSGPLGTLSSPSPSAV
FT                   PADHGAHLSLRGLPVCAFSSAGPCALRFTSARCMETTVNAHQILPKVLHKRTLGLPAMS
FT                   TTDLEAYFKDCVFKDWEELGEEIRLKVFVLGGCRHKLVCAPAPCNFFTSA"
FT   mat_peptide     1374..1838
FT                   /function="X protein coding sequence"
FT                   /locus_tag="HBVORF03"
FT                   /product="X protein"
FT                   /PRABI_name="HBx"
FT                   /PRABI_prodft=(pos:1..154, chain, "X protein")
FT   CDS             join(2854..3221,1..835)
FT                   /PRABI_name="PreS1"
FT                   /locus_tag="HBVORF01"
FT                   /codon_start="1"
FT                   /db_xref="UniProtKB/Swiss-Prot:P03141"
FT                   /translation="MGGWSSKPRKGMGTNLSVPNPLGFFPDHQLDPAFGANSNNPDWDF
FT                   NPVKDDWPAANQVGVGAFGPRLTPPHGGILGWSPQAQGILTTVSTIPPPASTNRQSGRQ
FT                   PTPISPPLRDSHPQAMQWNSTAFHQTLQDPRVRGLYLPAGGSSSGTVNPAPNIASHISS
FT                   ISARTGDPVTNMENITSGFLGPLLVLQAGFFLLTRILTIPQSLDSWWTSLNFLGGSPVC
FT                   LGQNSQSPTSNHSPTSCPPICPGYRWMCLRRFIIFLFILLLCLIFLLVLLDYQGMLPVC
FT                   PLIPGSTTTSTGPCKTCTTPAQGNSMFPSCCCTKPTDGNCTCIPIPSSWAFAKYLWEWA
FT                   SVRFSWLSLLVPFVQWFVGLSPTVWLSAIWMMWYWGPSLYSIVSPFIPLLPIFFCLWVY
FT                   I"
FT   mat_peptide     join(2854..3221,1..835)
FT                   /function="PreS1/Large Surface protein coding sequence"
FT                   /locus_tag="HBVORF01"
FT                   /product="PreS1 Surface protein"
FT                   /PRABI_name="LHBs"
FT                   /PRABI_prodft=(pos:1..400, chain, "Large Surface protein")
FT   CDS             join(3211..3221,1..835)
FT                   /PRABI_name="PreS2"
FT                   /locus_tag="HBVORF11"
FT                   /codon_start="1"
FT                   /translation="MQWNSTAFHQTLQDPRVRGLYLPAGGSSSGTVNPAPNIASHISSI
FT                   SARTGDPVTNMENITSGFLGPLLVLQAGFFLLTRILTIPQSLDSWWTSLNFLGGSPVCL
FT                   GQNSQSPTSNHSPTSCPPICPGYRWMCLRRFIIFLFILLLCLIFLLVLLDYQGMLPVCP
FT                   LIPGSTTTSTGPCKTCTTPAQGNSMFPSCCCTKPTDGNCTCIPIPSSWAFAKYLWEWAS
FT                   VRFSWLSLLVPFVQWFVGLSPTVWLSAIWMMWYWGPSLYSIVSPFIPLLPIFFCLWVYI
FT                   "
FT   mat_peptide     join(3211..3221,1..835)
FT                   /function="PreS2/Middle Surface protein coding sequence"
FT                   /locus_tag="HBVORF11"
FT                   /product="PreS2 Surface protein"
FT                   /PRABI_name="MHBs"
FT                   /PRABI_prodft=(pos:1..281, chain, "Middle Surface protein")
FT   CDS             155..835
FT                   /PRABI_name="S"
FT                   /locus_tag="HBVORF21"
FT                   /codon_start="1"
FT                   /translation="MENITSGFLGPLLVLQAGFFLLTRILTIPQSLDSWWTSLNFLGGS
FT                   PVCLGQNSQSPTSNHSPTSCPPICPGYRWMCLRRFIIFLFILLLCLIFLLVLLDYQGML
FT                   PVCPLIPGSTTTSTGPCKTCTTPAQGNSMFPSCCCTKPTDGNCTCIPIPSSWAFAKYLW
FT                   EWASVRFSWLSLLVPFVQWFVGLSPTVWLSAIWMMWYWGPSLYSIVSPFIPLLPIFFCL
FT                   WVYI"
FT   mat_peptide     155..835
FT                   /function="S protein coding sequence"
FT                   /locus_tag="HBVORF21"
FT                   /product="Surface protein S"
FT                   /PRABI_name="SHBs"
FT                   /PRABI_prodft=(pos:1..226, chain, "Small Surface protein")
FT   CDS             join(2307..3221,1..1623)
FT                   /PRABI_name="P"
FT                   /locus_tag="HBVORF13"
FT                   /codon_start="1"
FT                   /db_xref="UniProtKB/Swiss-Prot:P03159"
FT                   /translation="MPLSYQHFRKLLLLDDGTEAGPLEEELPRLADADLHRRVAEDLNL
```

```
FT                   GNLNVSIPWTHKVGNFTGLYSSTVPIFNPEWQTPSFPKIHLQEDIINRCQQFVGPLTVN
FT                   EKRRLKLIMPARFYPTHTKYLPLDKGIKPYYPDQVVNHYFQTRHYLHTLWKAGILYKRE
FT                   TTRSASFCGSPYSWEQELQHGRLVIKTSQRHGDESFCSQSSGILSRSSVGPCIRSQLKQ
FT                   SRLGLQPRQGRLASSQPSRSGSIRAKAHPSTRRYFGVEPSGSGHIDHSVNNSSSCLHQS
FT                   AVRKAAYSHLSTSKRQSSSGHAVEFHCLPPNSAGSQSQGSVSSCWWLQFRNSKPCSEYC
FT                   LSHLVNLREDWGPCDEHGEHHIRIPRTPARVTGGVFLVDKNPHNTAESRLVVDFSQFSR
FT                   GISRVSWPKFAVPNLQSLTNLLSSNLSWLSLDVSAAFYHIPLHPAAMPHLLIGSSGLSR
FT                   YVARLSSNSRINNNQYGTMQNLHDSCSRQLYVSLMLLYKTYGWKLHLYSHPIVLGFRKI
FT                   PMGVGLSPFLLAQFTSAICSVVRRAFPHCLAFSYMDDVVLGAKSVQHRESLYTAVTNFL
FT                   LSLGIHLNPNKTKRWGYSLNFMGYIIGSWGTLPQDHIVQKIKHCFRKLPVNRPIDWKVC
FT                   QRIVGLLGFAAPFTQCGYPALMPLYACIQAKQAFTFSPTYKAFLSKQYMNLYPVARQRP
FT                   GLCQVFADATPTGWGLAIGHQRMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSV
FT                   VLSRKYTSFPWLLGCTANWILRGTSFVYVPSALNPADDPSRGRLGLSRPLLRLPFQPTT
FT                   GRTSLYAVSPSVPSHLPVRVHFASPLHVAWRPP"
FT   mat_peptide     join(2307..3221,1..1623)
FT                   /function="DNA-polymerase/Reverse Transcriptase coding
FT                   sequence"
FT                   /locus_tag="HBVORF13"
FT                   /product="Polymerase/Reverse Transcriptase"
FT                   /PRABI_name="Pol"
FT                   /PRABI_prodft=(pos:1..845, chain, "Polymerase/Reverse
FT                   transcriptase")
FT                   /PRABI_prodft=(pos:1..183, domain, "Terminal Protein (TP)/
FT                   Primase domain")
FT                   /PRABI_prodft=(pos:1..183, domain, "Terminal Protein (TP)/
FT                   Primase domain")
FT                   /PRABI_prodft=(pos:184..348, domain, "Spacer")
FT                   /PRABI_prodft=(pos:349..692, domain, "Reverse Transcriptase
FT                   (RT) domain")
FT                   /PRABI_prodft=(pos:431..431, act_site, "RT catalytic Asp")
FT                   /PRABI_prodft=(pos:553..553, act_site, "RT catalytic Asp")
FT                   /PRABI_prodft=(pos:554..554, act_site, "RT catalytic Asp")
FT                   /PRABI_prodft=(pos:693..845, domain, "Ribonuclease H
FT                   (RNaseH) domain")
FT                   /PRABI_prodft=(pos:702..702, act_site, "RNaseH catalytic
FT                   Asp")
FT                   /PRABI_prodft=(pos:731..731, act_site, "RNaseH catalytic
FT                   Glu")
FT                   /PRABI_prodft=(pos:750..750, act_site, "RNaseH catalytic
FT                   Asp")
FT   CDS             join(2307..2453,489..680)
FT                   /PRABI_name="SP"
FT                   /locus_tag="HBVORF04"
FT                   /codon_start="1"
FT                   /translation="MPLSYQHFRKLLLLDDGTEAGPLEEELPRLADADLHRRVAEDLNL
FT                   GNLNDQQQPVRDHAKPARLLLKATLCFPHVAVQNLRMEIAPVFPSHRPGLSQNTYGSGP
FT                   QSVSLGSVY"
FT   mat_peptide     join(2307..2453,489..680)
FT                   /function="HBSP coding sequence"
FT                   /locus_tag="HBVORF04"
FT                   /product="Hepatitis B Spliced Protein"
FT                   /PRABI_name="HBSP"
FT                   /PRABI_prodft=(pos:1..113, chain, "HBV Spliced Protein")
XX
SQ   Sequence 3221 BP; 740 A; 868 C; 709 G; 904 T; 0 other;
     ttccactgcc ttccaccaaa ctctgcagga tcccagagtc aggggtctgt atcttcctgc        60
     tggtggctcc agttcaggaa cagtaaaccc tgctccgaat attgcctctc acatctcgtc       120
     aatctccgcg aggactgggg accctgtgac gaacatggag aacatcacat caggattcct       180
     aggacccctg ctcgtgttac aggcggggtt tttcttgttg acaagaatcc tcacaatacc       240
     gcagagtcta gactcgtggt ggacttctct caattttcta gggggatctc ccgtgtgtct       300
     tggccaaaat tcgcagtccc caacctccaa tcactcacca acctcctgtc ctccaatttg       360
     tcctggttat cgctggatgt gtctgcggcg ttttatcata ttcctcttca tcctgctgct       420
     atgcctcatc ttcttattgg ttcttctgga ttatcaaggt atgttgcccg tttgtcctct       480
     aattccagga tcaacaacaa ccagtacggg accatgcaaa acctgcacga ctcctgctca       540
     aggcaactct atgtttccct catgttgctg tacaaaacct acggatggaa attgcacctg       600
     tattcccatc ccatcgtcct gggctttcgc aaaataccta tgggagtggg cctcagtccg       660
     tttctcttgg ctcagtttac tagtgccatt tgttcagtgg ttcgtagggc tttcccccac       720
     tgtttggctt tcagctatat ggatgatgtg gtattggggg ccaagtctgt acagcatcgt       780
     gagtcccttt ataccgctgt taccaatttt cttttgtctc tgggtataca tttaaaccct       840
     aacaaaacaa aaagatgggg ttattcccta aacttcatgg gctacataat tggaagttgg       900
```

```
     ggaactttgc cacaggatca tattgtacaa aagatcaaac actgttttag aaaacttcct       960
     gttaacaggc ctattgattg gaaagtatgt caaagaattg tgggtctttt gggctttgct      1020
     gctccattta cacaatgtgg atatcctgcc ttaatgcctt tgtatgcatg tatacaagct      1080
     aaacaggctt tcactttctc gccaacttac aaggcctttc taagtaaaca gtacatgaac      1140
     ctttaccccg ttgctcggca acggcctggt ctgtgccaag tgtttgctga cgcaaccccc      1200
     actggctggg gcttggccat aggccatcag cgcatgcgtg gaacctttgt ggctcctctg      1260
     ccgatccata ctgcggaact cctagccgct tgttttgctc gcagccggtc tggagcaaag      1320
     ctcatcggaa ctgacaattc tgtcgtcctc tcgcggaaat atacatcgtt tccatggctg      1380
     ctaggctgta ctgccaactg gatccttcgc gggacgtcct ttgtttacgt cccgtcggcg      1440
     ctgaatcccg cggacgaccc ctctcggggc cgcttgggac tctctcgtcc ccttctccgt      1500
     ctgccgttcc agccgaccac ggggcgcacc tctctttacg cggtctcccc gtctgtgcct      1560
     tctcatctgc cggtccgtgt gcacttcgct tcacctctgc acgttgcatg gagaccaccg      1620
     tgaacgccca tcagatcctg cccaaggtct tacataagag gactcttgga ctcccagcaa      1680
     tgtcaacgac cgaccttgag gcctacttca aagactgtgt gtttaaggac tgggaggagc      1740
     tgggggagga gattaggtta aaggtctttg tattaggagg ctgtaggcac aaattggtct      1800
     gcgcaccagc accatgcaac tttttcacct ctgcctaatc atctcttgta catgtcccac      1860
     tgttcaagcc tccaagctgt gccttgggtg gctttggggc atggacattg acccttataa      1920
     agaatttgga gctactgtgg agttactctc gtttttgcct tctgacttct ttccttccgt      1980
     cagagatctc ctagacaccg cctcagctct gtatcgagaa gccttagagt ctcctgagca      2040
     ttgctcacct caccatactg cactcaggca agccattctc tgctgggggg aattgatgac      2100
     tctagctacc tgggtgggta ataatttgga agatccagca tctagggatc ttgtagtaaa      2160
     ttatgttaat actaacgtgg gtttaaagat caggcaacta ttgtggtttc atatatcttg      2220
     ccttactttt ggaagagaga ctgtacttga atatttggtc tctttcggag tgtggattcg      2280
     cactcctcca gcctatagac caccaaatgc ccctatctta tcaacacttc cggaaactac      2340
     tgttgttaga cgacgggacc gaggcaggtc ccctagaaga agaactccct cgcctcgcag      2400
     acgcagatct ccatcgccgc gtcgcagaag atctcaatct cgggaatctc aatgttagta      2460
     ttccttggac tcataaggtg ggaaacttta cggggcttta ttcctctaca gtacctatct      2520
     ttaatcctga atggcaaact ccttcctttc ctaagattca tttacaagag gacattatta      2580
     ataggtgtca acaatttgtg ggccctctca ctgtaaatga aaagagaaga ttgaaattaa      2640
     ttatgcctgc tagattctat cctacccaca ctaaatattt gcccttagac aaaggaatta      2700
     aaccttatta tccagatcag gtagttaatc attacttcca aaccagacat tatttacata      2760
     ctctttggaa ggctggtatt ctatataagc gggaaaccac acgtagcgca tcattttgcg      2820
     ggtcaccata ttcttgggaa caagagctac agcatgggag gttggtcatc aaaacctcgc      2880
     aaaggcatgg ggacgaatct ttctgttccc aatcctctgg gattctttcc cgatcatcag      2940
     ttggaccctg cattcggagc caactcaaac aatccagatt gggacttcaa ccccgtcaag      3000
     gacgactggc cagcagccaa ccaagtagga gtgggagcat tcgggccaag gctcacccct      3060
     ccacacggcg gtattttggg gtggagccct caggctcagg gcatattgac cacagtgtca      3120
     acaattcctc ctcctgcctc caccaatcgg cagtcaggaa ggcagcctac tcccatctct      3180
     ccacctctaa gagacagtca tcctcaggcc atgcagtgga a                          3221
//
```

Annexe 5 : Entrée annotée HBVdb X02763.

Annexe 6

```
P RGLLNu1Gv1Xw provisional genotype RF_GA
C Informative positions: 100.00 %
N 3248
ttccactgccgtccaccaagctctgcaggatcccagagtcaggggtctgaattttcctgctggtggctccagttc
aggaacagtaaaccctgctccgaatattgcctctcacatctcgtcaatctccgcgaggactggggaccctgtgac
gaacatggagaacatcacatcaggattcctaggacccctgctcgtgttacaggcggggtttttcttgttgacaag
aatcctcacaataccgcagagtctagactcgtggtggacttctctcaattttctagggggagtgcccgtgtgtcc
tggcctaaattcgcagtccccaacctccaatcactcaccaacctcctgtcctccaacttgtcctggctatcgctg
gatgtgtctgcggcgttttatcatattcctcttcatcctgctgctatgcctcatcttcttgttggttcttctgga
ctatcaaggtatgttgcccgtttgtcctctgattccaggatcctcgaccaccagcacgggaccctgcaaaacctg
cacgactcctgctcaaggcaactctatgtatccctcatgttgctgtacaaaacctacggacggaaattgcacctg
tattcccatcccatcgtcctgggctttcgcaaaatacctatgggagtgggcctcagtccgtttctcatggctcag
tttactagtgccatttgttcagtggttcgtagggctttcccccactgtttggctttcagctatattgatgatatt
gtattgggggccaaatctgtacagcatcgtgagtccctttataccgctgttaccaattttcttttgtctttgggt
atacatctaaaccctaacaaaacaaaaagatggggttattccttaaattttatgggatatgtaattggaagttgg
ggtactttgccacaagaacacatcacacagaaaattaagcaatgttttcggaaactccccgttaacaggccaatt
gattggaaagtctgtcaacgaataactggtctgttgggtttcgctgctccttttacccaatgtggttaccctgcc
ttaatgcctttatatgcatgtatacaagctaagcaggcttttactttctcgccaacctataaggcctttctctgt
aaacaatacatgaaccttttaccccgttgctaggcaacggcccggtctatgccaagtgtttgctgacgcaacccccc
actggttggggcttggccaccggccatcagcgcatgcgtggaacctttgtggctcctctgccgatccatactgcg
gaactcctagctgcttgttttgctcgcagccggtctggagcaaaactcattgggactgacaattctgtcgtcctt
tctcggaaatatacatcctttccatggctgctaggctgtgctgccaactggatccttcgcgggacgtcctttgtt
tacgtcccgtcagcgctgaatccagcggacgacccctcccggggccgtttggggctctgtcgcccccttctccgt
ctgccgttcctgccgaccacggggcgcacctctctttacgcggtctccccgtctgttccttctcatctgccggac
cgtgtgcacttcgcttcacctctgcacgttacatggaaaccgccatgaacacctctcatcatctgccaaggcagt
tatataagaggactcttggactgtttgttatgtcaacaaccggggtggagaaatacttcagggactgtgtttttg
ctgagtgggaagaattaggcaatgagtccaggttaatgacctttgtattaggaggctgtaggcataaattggtct
gcgcaccagcaccaagcaacttttcacctctgcctaatcatctcttgttcatgtcctactgttcaagcctccaa
gctgtgcctggggtggctttagggcatggatagaataactttgccgtatggctttttggcttagacattgaccc
ttataaagaatttggagctactgtggagttgctctcgttttgccttctgacttttcccgtctgttcgtgatct
tctcgacaccgcttcagctttgtaccgggaatccttagaatcctctgatcattgttcgcctcaccatacagcact
caggcaagcaatcctgtgctggggtgagttgatgactctagctacctgggtgggtaataatttggaagatccagc
atccagagatttggtggtcaattatgttaatactaatatgggtttaaaaatcaggcaactattgtggtttcacat
ttcctgtcttacttttggaagagaaaccgttcttgagtatttggtgtcttttggagtgtggattcgcactcctcc
tgcttatagaccaccaaatgcccctatcttatcaacacttccggagactactgttgttagacgaagaggcaggtc
ccctcgaagaagaactccctcgcctcgcagacgaagatctcaatcgccgcgtcgcagaagatctgcatctccagc
ttcccgatgttagtattccttggactcacaaggtgggaaactttacggggctgtattcttctactactcctgtct
ttaatcctgattggcaaactccttcttttccaaatatccatttgcatcaagacattataactaaatgtgaacaat
ttgtgggccctctcacagtaaatgagaaacgaagattaaaactagttatgcctgccagattttcccaaactcta
ctaaatatttaccattagacaaaggtatcaaaccgtattatccagaacatgtagttaatcattacttccagacca
gacattatttacataccctttggaaggcgggtattctatataagagagaaacatcccgtagcgcttcattttgtg
ggtcaccatatacttgggaacaagatctacagcatggggctttcttggacggtccctctcgagtggggaaagaac
ctttccaccagcaatcctctaggattccttcccgatcaccagttggacccagcattcagagcaaataccaacagt
ccagattgggacttcaatcccaaaaaggacccttggccagaggccaacaaggtaggagtgggagcattcgggcca
gggctcacccctccacacggaggccttttggggtggagccctcaggctcagggcatattgaccacagtgtcaaca
attcctcctcctgcctccaccaatcggcagtcaggaaggcagcctactcccatctctccacctctaagagacagt
catcctcaggccatgcagtggaa
G
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

```
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAA
I Window size: 301, Window step: 1, Min. confidence: 75.00, Min. window
identity: 90.00 , Min. nucleotide: 0180
```

Annexe 6 : Fichier de sortie du processus de genotypage *GenotypeHBV*.

Le génotype prédit (ligne P) est présenté (provisional), ainsi que le pourcentage de positions informatives (C), le nombre de positions utilisées (N), la séquence génotypée, le génotype prédit à chaque position (G), et les informations sur les paramètres du programmes (I).

Annexe 7

```
Resistant
adefovir 5     I    1    M204I     M549I
entecavir 26   I    1    M204I     M549I
lamivudine 2   R    1    M204I     M549I
telbivudine 1  R    1    M204I     M549I

>RGLLNu1Gv1Xw_P
MPLSYQHFRRLLLLDEEAGPLEEELPRLADEDLNRRVAEDLHLQLPDVSIPWTHKVGNFTGLYSSTTPVFNPDWQTPSFP
NIHLHQDIITKCEQFVGPLTVNEKRRLKLVMPARFFPNSTKYLPLDKGIKPYYPEHVVNHYFQTRHYLHTLWKAGILYKR
ETSRSASFCGSPYTWEQDLQHGAFLDGPSRVGKEPFHQQSSRIPSRSPVGPSIQSKYQQSRLGLQSQKGPLARGQQGRSG
SIRARAHPSTRRPFGVEPSGSGHIDHSVNNSSSCLHQSAVRKAAYSHLSTSKRQSSSGHAVEFHCRPPSSAGSQSQGSEF
SCWWLQFRNSKPCSEYCLSHLVNLREDWGPCDEHGEHHIRIPRTPARVTGGVFLVDKNPHNTAESRLVVDFSQFSRGSAR
VSWPKFAVPNLQSLTNLLSSNLSWLSLDVSAAFYHIPLHPAAMPHLLVGSSGLSRYVARLSSDSRILDHQHGTLQNLHDS
CSRQLYVSLMLLYKTYGRKLHLYSHPIVLGFRKIPMGVGLSPFLMAQFTSAICSVVRRAFPHCLAFSYIDDIVLGAKSVQ
HRESLYTAVTNFLLSLGIHLNPNKTKRWGYSLNFMGYVIGSWGTLPQEHITQKIKQCFRKLPVNRPIDWKVCQRITGLLG
FAAPFTQCGYPALMPLYACIQAKQAFTFSPTYKAFLCKQYMNLYPVARQRPGLCQVFADATPTGWGLATGHQRMRGTFVA
PLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVYVPSALNPADDPSRGRLGLCRPL
LRLPFLPTTGRTSLYAVSPSVPSHLPDRVHFASPLHVTWKPP
```

Annexe 7 : Fichier de sortie du processus de détection des profils de résistance *FindHBVRMut*.
La première ligne donne le statut général de résistance, les lignes suivantes donnent, pour chaque ligne : le nom du profil (drogue + numéro), le statut de résistance (R, I ou S), le nombre de mutations pour le profil, la mutation en numérotation RT et la mutation avec la numérotation de la séquence requête, qui est présentée en dessous.

Annexe 8

	HBc	HBe	HBx	LHBs	MHBs	SHBs	Pol	HBSP
A	963	725	740	925	971	1569	714	885
B	134		200	233	336	1753	199	199
C	1794	1270	1385	1227	1341	4363	1244	48
D	416	67	308	394	423	1911	279	252
E	525	342	235	227	229	595	223	233
F	119	102	192	111	112	290	103	
G	72		36	64	64	112	32	71
H	26	25	25	28	29	69	24	25

Annexe 8 : Nombre de séquences protéiques (*mat_peptides*) complètes pour chaque génotype.

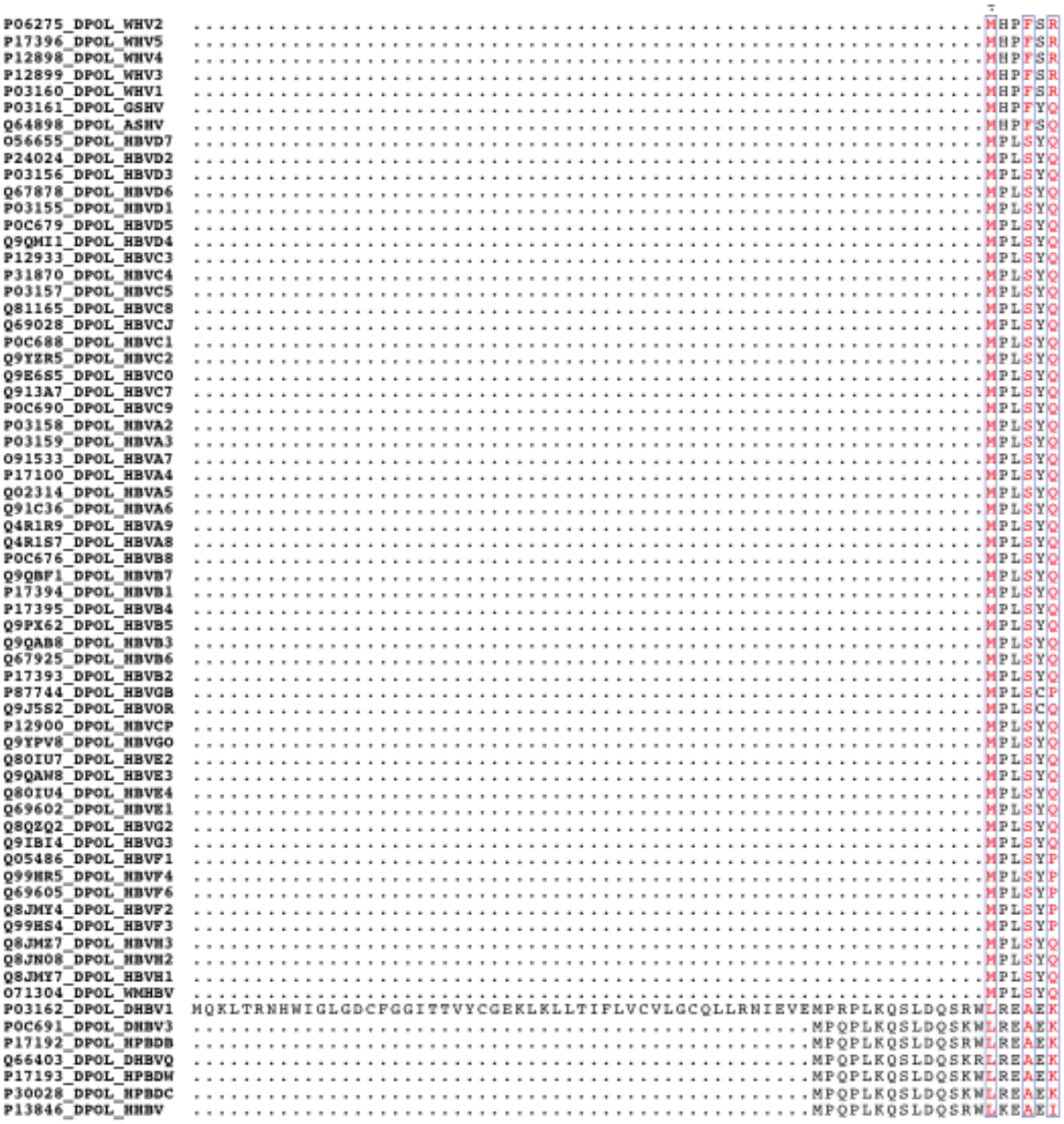
Annexe 9

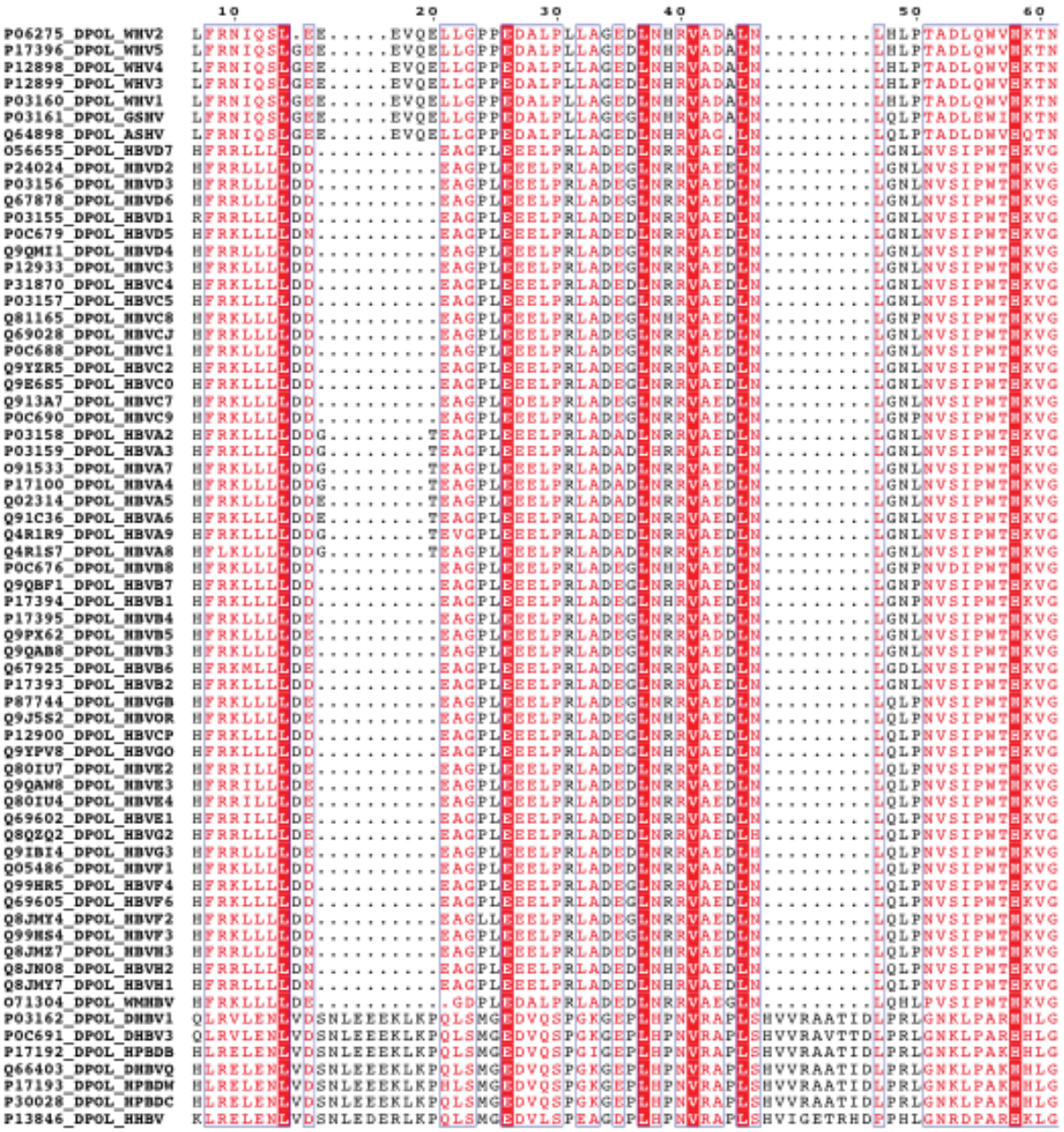
Position	Residue	All	NotD	Dpop	Dudps
681	S	0,217	0,22	0,097	0,031
698	V	0,08	0,003	0,121	0,092
699	M	0,136	0,078	0,095	0,066
702	Q	0,184	0,191	0,017	0,015
709	Q	0,156	0,041	0,463	0,449
732	N	0,22	0,182	0,01	0,086
733	I	0,111	0,051	0,115	0,094
734	I	0,076	0,022	0,203	0,214
741	V	0,004	0,005	0	0,104
742	L	0,001	0	0,01	0,101
743	S	0,004	0,004	0	0,103
744	R	0,001	0,001	0	0,111
745	K	0,008	0,008	0	0,125
747	T	0,004	0,004	0	0,113
748	S	0,005	0,005	0	0,126
749	F	0,03	0,006	0,151	0,205
750	P	0,001	0,001	0	0,115
751	W	0,001	0,001	0	0,123
753	L	0,006	0,006	0,01	0,109
756	A	0,185	0,189	0,041	0,056
786	L	0,132	0,137	0,056	0,303
787	C	0,25	0,216	0,313	0,345
792	R	0,233	0,238	0,076	0,012
795	F	0,109	0,115	0,024	0,043
796	R	0,225	0,232	0,036	0,042
807	D	0,18	0,143	0,041	0,085
817	D	0,257	0,265	0,034	0,234
830	R	0,145	0,146	0,109	0,124

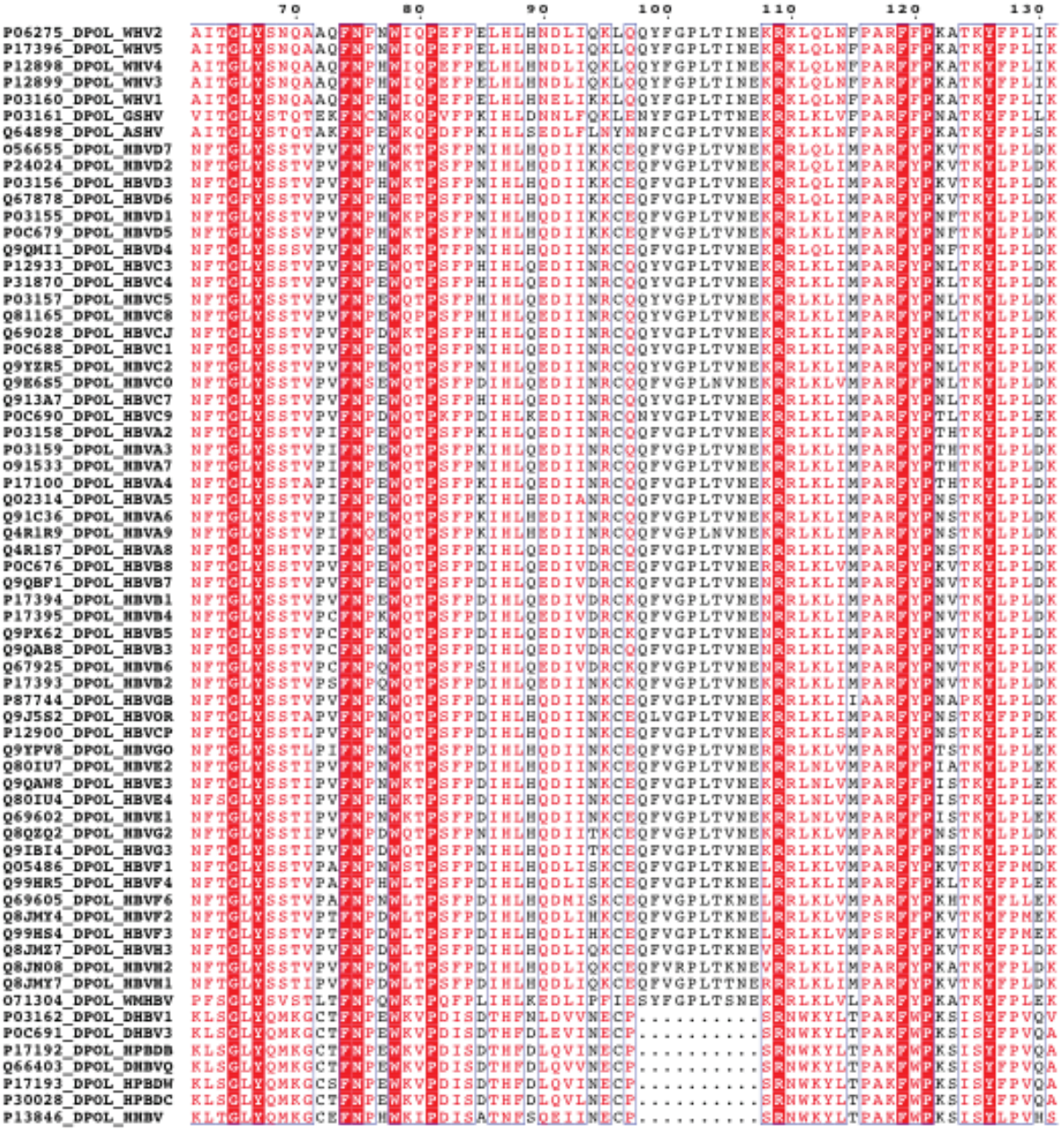
Annexe 9 : Liste des positions de la RNase H avec une entropie de Shannon normalisée > 0,1
(parmi les 4 jeux de données : all, notD, Dpop et Dudps)

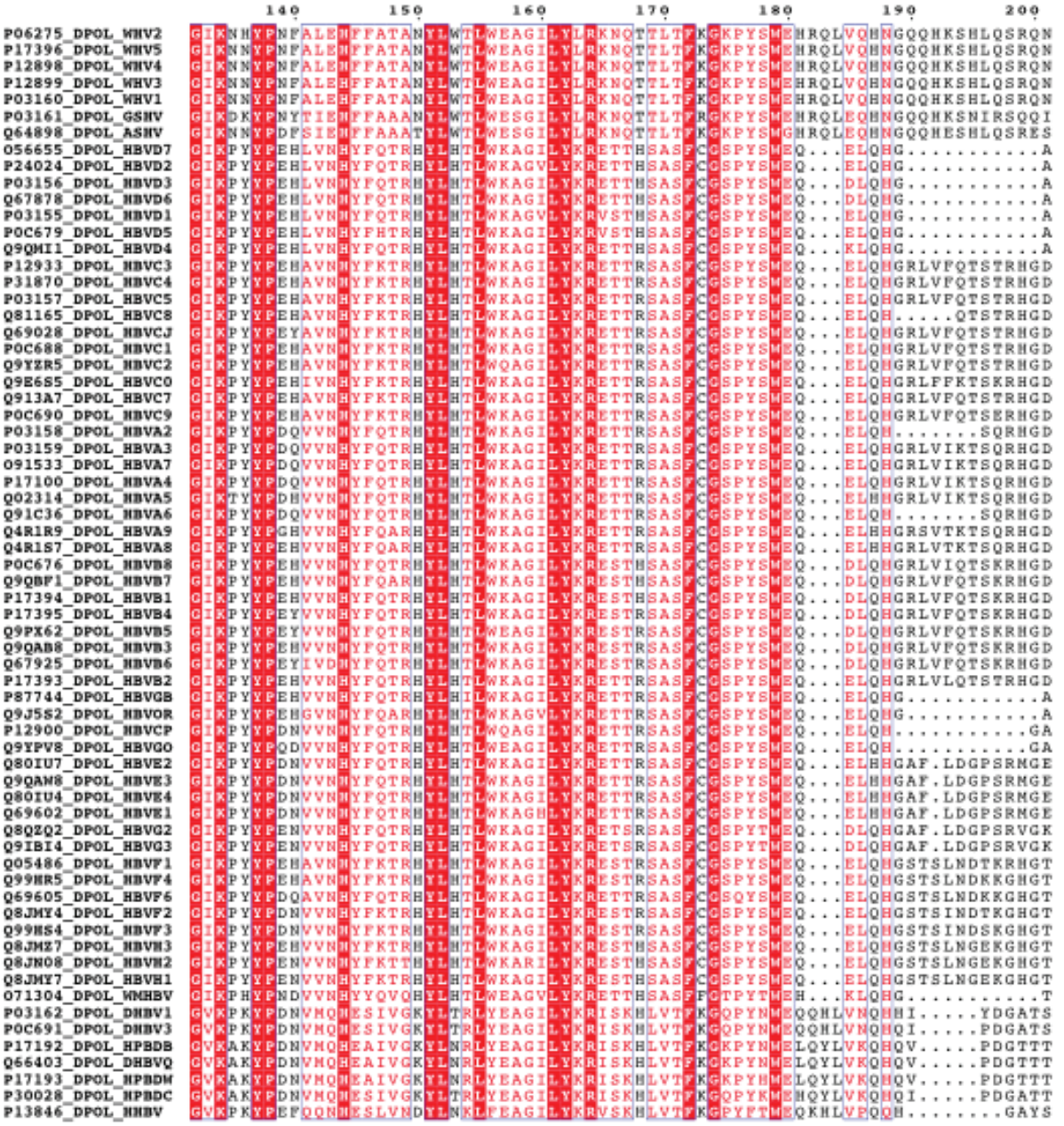
Annexe 10

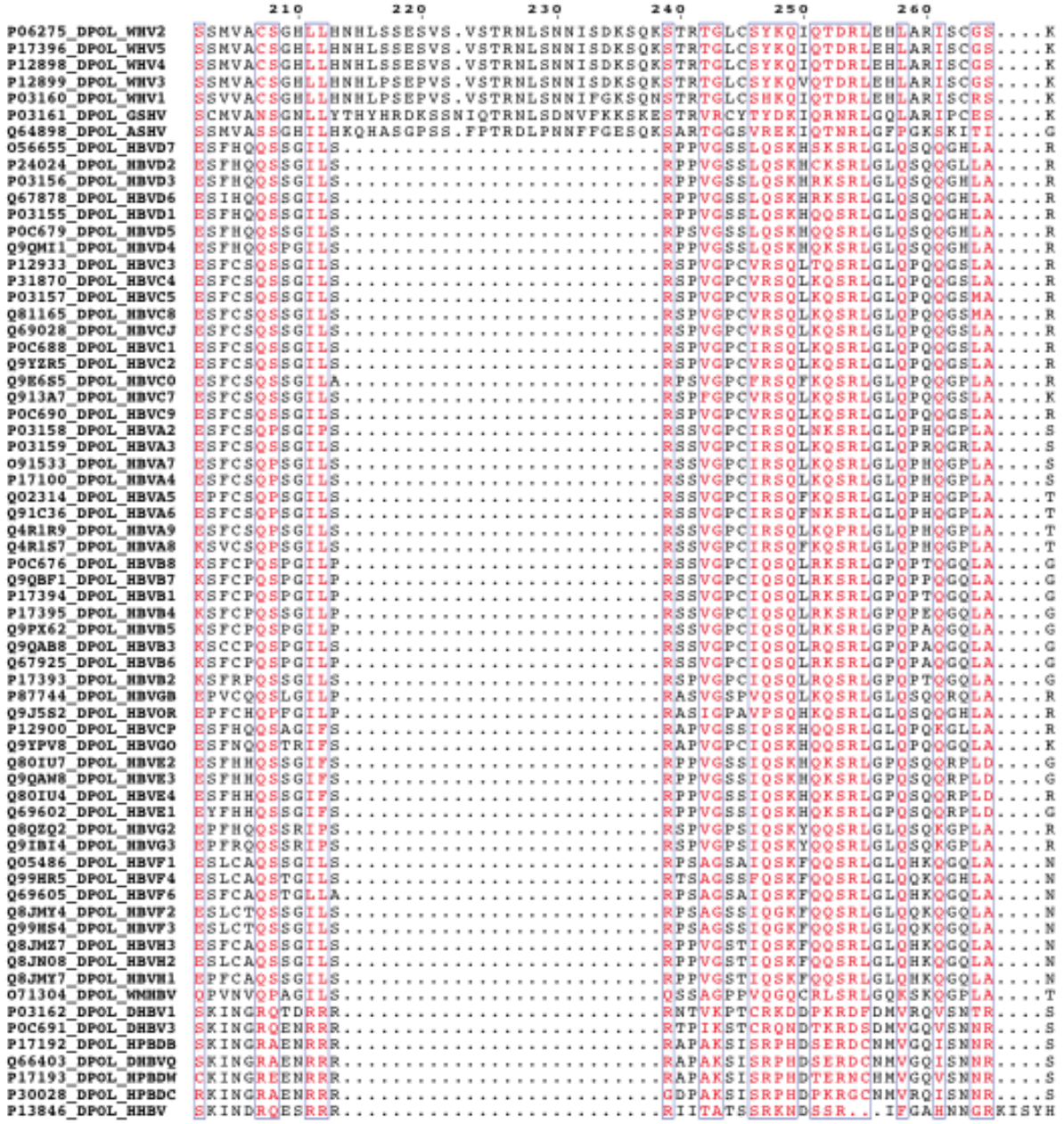
Annexe 10 : Alignement multiple des 66 séquences de polymérase d'hepadnavirus

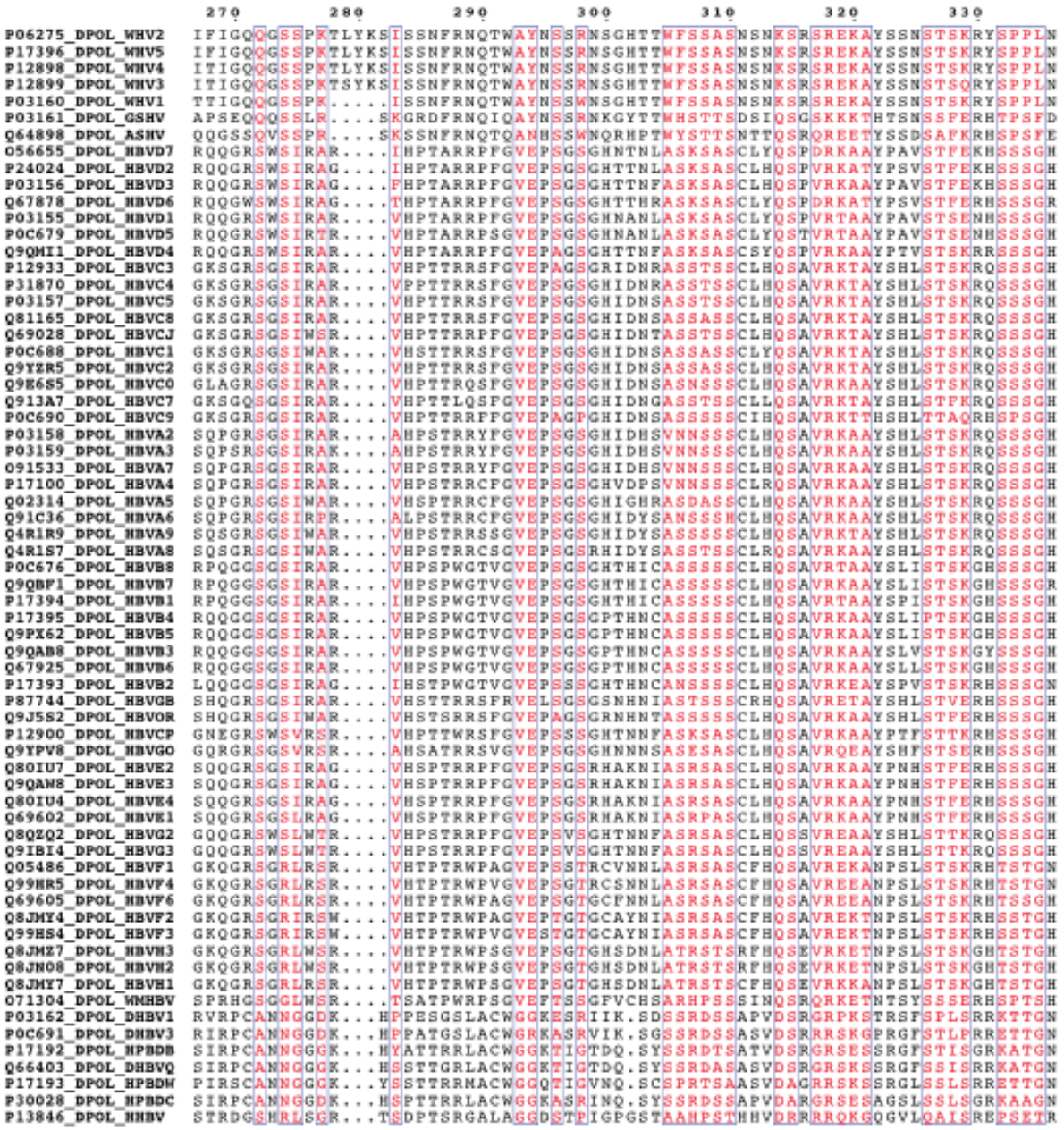


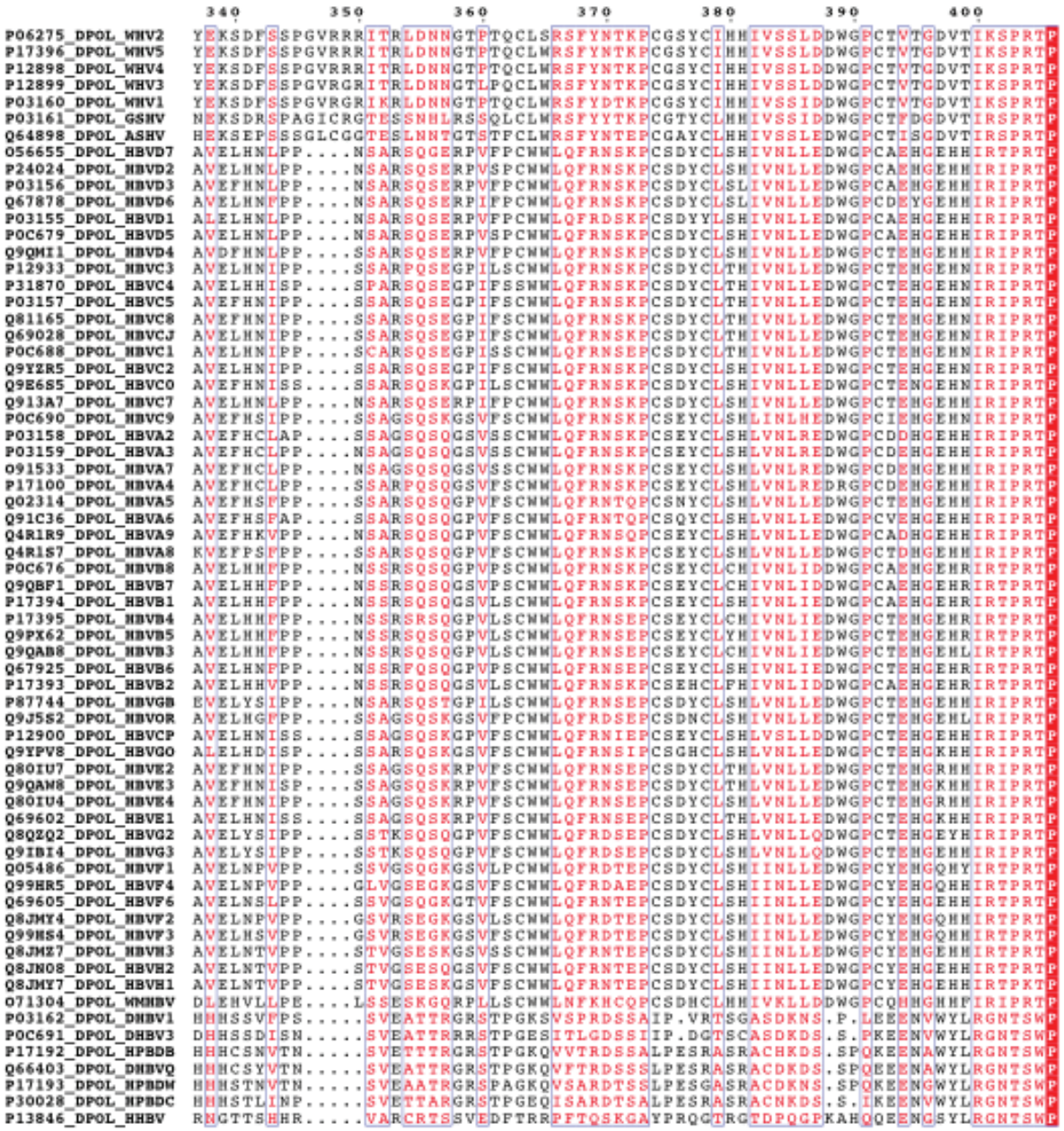

10
20
30
40
50
60
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVC0
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAN8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JNO8_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WMHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_HPBDB
Q66403_DPOL_DHBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_HHBV


70
80
90
100
110
120
130
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVC0
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAW8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JNO8_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WNHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_HPBDB
Q66403_DPOL_DHBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_HNBV

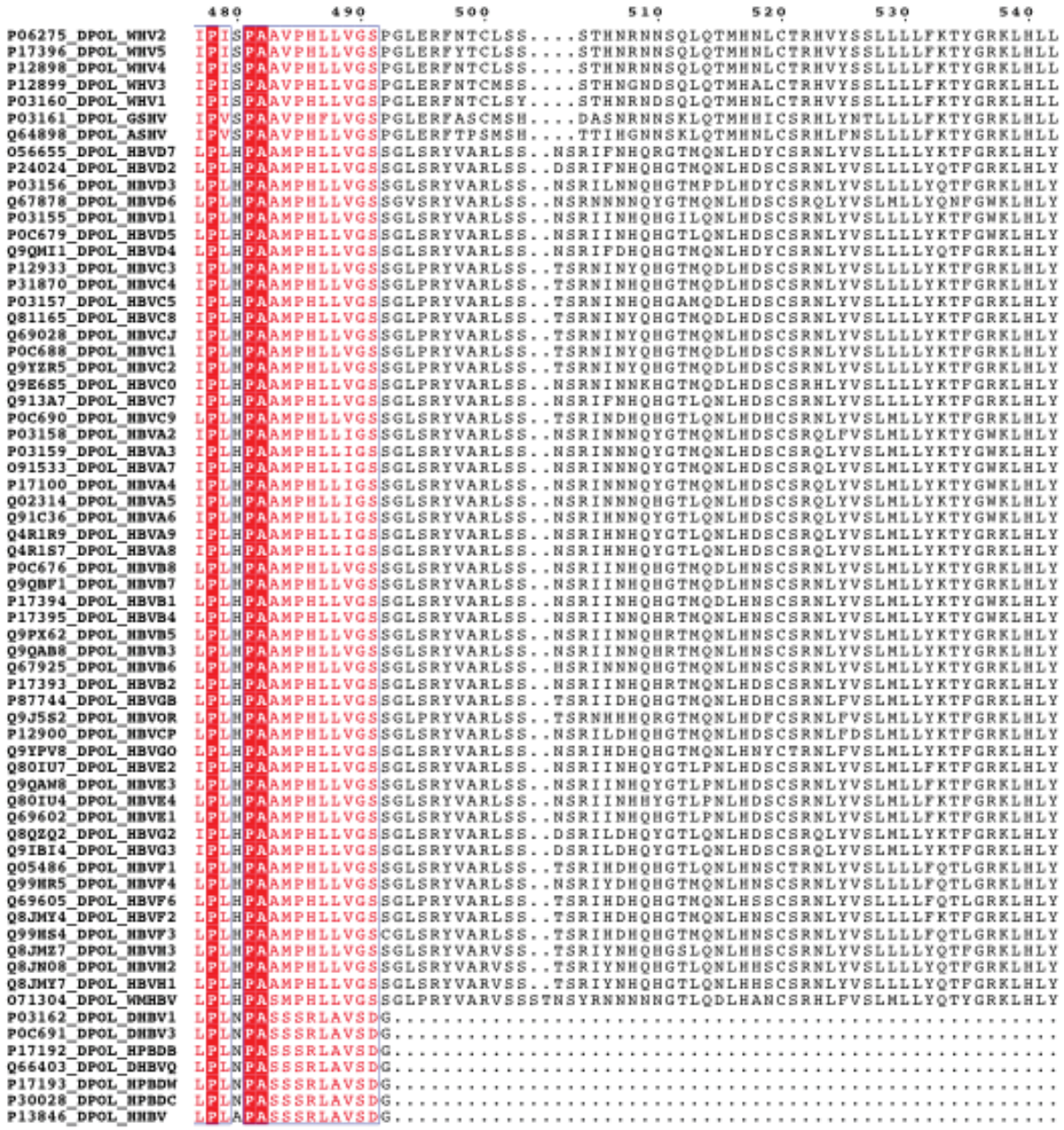

140
150
160
170
180
190
200
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVC0
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAW8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JNO8_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WMHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_HPBDB
Q66403_DPOL_DHBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_HHBV


210
220
230
240
250
260
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVC0
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAN8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JNO8_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WMHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_HPBDB
Q66403_DPOL_DHBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_HHBV

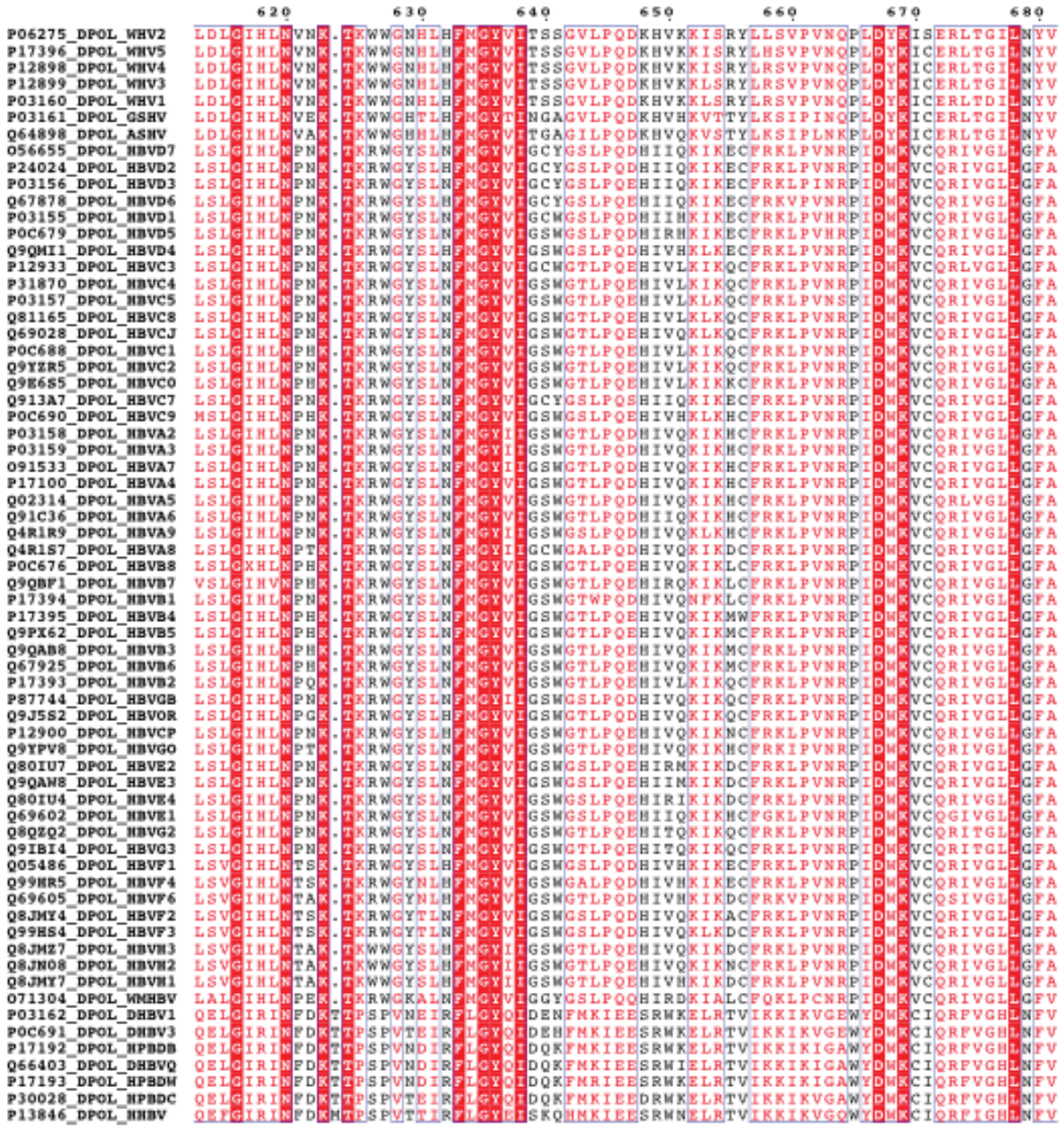

270
280
290
300
310
320
330
P06275_DPOL_WHV2 IFIGQQGSSPKTLYKSISSNFRNQTWAYNSSRNSGHTTMFSSASNSNKSRSREKAYSSNSTSKRYSPPLN
P17396_DPOL_WHV5 IFIGQQGSSPKTLYKSISSNFRNQTWAYNSSRNSGHTTMFSSASNSNKSRSREKAYSSNSTSKRYSPPLN
P12898_DPOL_WHV4 ITIGQQGSSPKTLYKSISSNFRNQTWAYNSSRNSGHTTMFSSASNSNKSRSREKAYSSNSTSKRYSPPLN
P12899_DPOL_WHV3 ITIGQQGSSPKTSYKSISSNFRNQTWAYNSSRNSGHTTMFSSASNSNKSRSREKAYSSNSTSQRYSPPLN
P03160_DPOL_WHV1 TTIGQQGSSPX.....ISSNFRNQTWAYNSSWNSGHTTMFSSASNSNKSRSREKAYSSNSTSKRYSPPLN
P03161_DPOL_GSHV APSEQQQSSLR....SKGRDFRNQIQAYNSSRNKGYTTMHSTTSDSIQSGSKKKTHTSNSSPERHTPSFD
Q64898_DPOL_ASHV QQGSSQVSSPR....SKSSNFRNQTQANHSSWNQRHPTMYSTTSNTTQSRQREETYSSDSAFKRHSPSFE
O56655_DPOL_HBVD7 RQQGRSWSIRAR....IHPTARRPFGVEPSGSGHNTNLASKSASCLYQSPDRKAAYPAVSTFEKHSSSGH
P24024_DPOL_HBVD2 RQQGRSWSIRAG....IHPTARRPFGVEPSGSGHTTNLASKSASCLHQSPVRKATYPSVSTFEKHSSSGH
P03156_DPOL_HBVD3 RQQGRSWSIRAG....FHPTARRPFGVEPSGSGHTTNFASKSASCLHQSPVRKAAYPAVSTFEKHSSSGH
Q67878_DPOL_HBVD6 RQQGWSWSIRAG....THPTARRPFGVEPSGSGHTTHRASKSASCLYQSPDRKATYPSVSTFERHSSSGR
P03155_DPOL_HBVD1 RQQGRSWSIRAR....VHPTARRPFGVEPSGSGHNANLASKSASCLYQSPVRTAAYPAVSTSENHSSSGH
P0C679_DPOL_HBVD5 RQQGRSWSIRTR....VHPTARRPSGVEPSGSGHNANLASKSASCLYQSTVRTAAYPAVSTSENHSSSGH
Q9QMI1_DPOL_HBVD4 RQQGRSWSIRAR....VHPTARRPFGVEPAGSGHTTNFASKSASCSYQSPVRKAAYPTVSTSKRRSSSGH
P12933_DPOL_HBVC3 GKSGRSGSIRAR....VHPTTRRSFGVEPAGSGRIDNRASSTSSCLHQSAVRKTAYSHLSTSKRQSSSGH
P31870_DPOL_HBVC4 GKSGRSGSIRAR....VPPTTRRSFGVEPSGSGHIDNRASSTSSCLHQSAVRKTAYSHLSTSKRQSSSGH
P03157_DPOL_HBVC5 GKSGRSGSIRAR....VHPTTRRSFGVEPSGSGHIDNSASSTSSCLHQSAVRKTAYSHLSTSKRQSSSGH
Q81165_DPOL_HBVC8 GKSGRSGSIRAR....VHPTTRRSFGVEPSGSGHIDNSASSASSCLHQSAVRKTAYSHLSTSKRQSSSGH
Q69028_DPOL_HBVCJ GKSGRSGSIWSR....VHPTTRRPFGVEPSGSGHIDNTASSTSSCLHQSAVRKTAYSHLSTSKRQSSSGH
P0C688_DPOL_HBVC1 GKSGRSGSIWAR....VHSTTRRSFGVEPSGSGHIDNSASSASSCLYQSAVRKTAYSHLSTSKRQSSSGH
Q9YZR5_DPOL_HBVC2 GKSGRSGSIRAR....VHPTTRRSFGVEPSGSGHIDNSASSASSCLHQSAVRKTAYSHLSTSKRQSSSGH
Q9E6S5_DPOL_HBVC0 GLAGRSGSIRAR....VHPTTRQSFGVEPSGSGHIDNSASNSSCLHQSAVRKAAYSHLSTSKRQSSSGH
Q913A7_DPOL_HBVC7 GKSGQSGSIRAR....VHPTTLQSFGVEPSGSGHIDNGASSTSSCLLQSAVRKTAYSHLSTFKRQSSSGH
P0C690_DPOL_HBVC9 GKSGRSGSIRAR....VHPTTRRFFGVEPAGPGHIDNSASSSSSCIHQSAVRKTTHSHLTTAQRHSPSGH
P03158_DPOL_HBVA2 SQPGRSGSIRAR....AHPSTRRYFGVEPSGSGHIDHSVNNSSSCLHQSAVRKAAYSHLSTSKRQSSSGH
P03159_DPOL_HBVA3 SQPSRSGSIRAK....AHPSTRRYFGVEPSGSGHIDHSVNNSSSCLHQSAVRKAAYSHLSTSKRQSSSGH
O91533_DPOL_HBVA7 SQPGRSGSIRAR....VHPSTRRYFGVEPSGSGHIDHSVNNSSSCLHQSAVRKAAYSHLSTSKRQSSSGH
P17100_DPOL_HBVA4 SQPGRSGSIRAR....VHPSTRRCFGVEPSGSGHVDPSVNNSSSCLRQSAVRKAAYSHLSTSKRQSSSGH
Q02314_DPOL_HBVA5 SQPGRSGSIWAR....VHSPTRRCFGVEPSGSGHIGHRASDASSCLHQSAVRKAAYSHLSTSKRQSSSGH
Q91C36_DPOL_HBVA6 SQPGRSGSIRPR....ALPSTRRCFGVEPSGSGHIDYSANSSSMCLHQSAVRKAAYSHLSTSKRQSSSGH
Q4R1R9_DPOL_HBVA9 SQSGRSGSIWAR....VHPSTRRSSGVEPSGSGHIDYSASSSSSCLHQSAVRKTAYSHLSTSKRQSSSGH
Q4R1S7_DPOL_HBVA8 SQSGRSGSIWAR....VHPSTRRCSGVEPSGSRHIDYSASSTSSCLRQSAVRKAAYSHLSTSKRQSSSGH
P0C676_DPOL_HBVB8 RPQGGSGSIRAR....VHPSPWGTVGVEPSGSGHTHICASSSSSCLHQSAVRTAAYSLISTSKGHSSSGH
Q9QBF1_DPOL_HBVB7 RPQGGSGSIRAR....VHPSPWGTVGVEPSGSGHTHICASSSSSCLHQSAVRKAAYSLISTSKGHSSSGR
P17394_DPOL_HBVB1 RPQGGSGSIRAR....IHPSPWGTVGVEPSGSGHTHICASSSSSCLHQSAVRTAAYSPISTSKGHSSSGH
P17395_DPOL_HBVB4 RQQGGSGSIRAR....VHPSPWGTVGVEPSGSGPTHNCASSSSSCLHQSAVRKAAYSLIPTSKGHSSSGH
Q9PX62_DPOL_HBVB5 RQQGGSGSIRAR....VHPSPWGTVGVEPSGSGPTHNCASSSSSCLHQSAVRKAAYSLISTSKGHSSSGH
Q9QAB8_DPOL_HBVB3 RQQGGSGSIRAR....VHPSPWGTVGVEPSGSGPTHNCASSSSSCLHQSAVRKAAYSLVSTSKGYSSSGH
Q67925_DPOL_HBVB6 RQQGGSGSIRAR....VHPSPWGTVGVEPSGSGPTHNCASSSSSCLHQSAVRKAAYSLLSTSKGHSSSGH
P17393_DPOL_HBVB2 LQQGGSGSIRAG....IHSTPWGTVGVEPSSSGHTHNCANSSSSCLHQSAVRKEAYSPVSTSKRHSSSGN
P87744_DPOL_HBVGB SHQGRSGSIRAR....VHSTTRRSFRVELSGSGSNHNIASTSSSCRHQSAVRETAYSHLSTVERHSSSGH
Q9J5S2_DPOL_HBVOR SHQGRSGSIWAR....VHSTSRRSFGVEPAGSGRNHNTASSSSSCLHQSAVRKAAYSHLSTFERHSSSGH
P12900_DPOL_HBVCP GNEGRSWSVRSR....VHPTTWRSFGVEPSSSGHTNNFASKSASCLHQSAVRKAAYPTFSTTKRHSSSGH
Q9YPV8_DPOL_HBVGO GQRGRSGSVRSR....AHSATRRSVGVEPSGSGHNNNSASESASCLHQSAVRQEAYSHFSTSERHSSSGH
Q80IU7_DPOL_HBVE2 SQQGRSGSIRAG....VHSPTRRPFGVEPSGSRHAKNIASRSASCLHQSAVRKAAYPNHSTFERHSSSGH
Q9QAW8_DPOL_HBVE3 SQQGRSGSIRAG....VHSPTRRPFGVEPSGSRHAKNIASRSASCLHQSAVRKAAYPNHSTFERHSSSGH
Q80IU4_DPOL_HBVE4 SQQGRSGSIRAG....VHSPTRRPFGVEPSGSRHAKNIASRSASCLHQSAVRKAAYPNHSTFERHSSSGH
Q69602_DPOL_HBVE1 SQQGRSGSLRAG....VHSPTRRPFGVEPSGSRHAKNIASRPASCLHQSAVRKAAYPNHSTFERHSSSGH
Q8QZQ2_DPOL_HBVG2 GQQGRSWSLWTR....VHPSTRRPFGVEPSVSGHTNNFASRSASCLHQSSVREAAYSHLSTTKRQSSSGH
Q9IBI4_DPOL_HBVG3 GQQGRSWSLWTR....VHPSTRRPFGVEPSVSGHTNNFASRSASCLHQSSVREAAYSHLSTTKRQSSSGH
Q05486_DPOL_HBVF1 GKQGRSGRLRSR....VHTPTRWPAGVEPSSTRCVNNLASRSASCFHQSAVREKANPSLSTSKRHTSTGN
Q99HR5_DPOL_HBVF4 GKQGRSGRLRSR....VHTPTRWPVGVEPSGTRCSNNLASRSASCFHQSAVREEANPSLSTSKRHTSTGN
Q69605_DPOL_HBVF6 GKQGRSGRLRSR....VHTPTRWPAGVEPSGTGCFNNLASRSASCFHQSAVREEANPSLSTSKRHTSSGH
Q8JMY4_DPOL_HBVF2 GKQGRSGRIRSW....VHTPTRWPAGVEPTGTGCAYNIASRSASCFHQSAVREKTNPSLSTSKRHSSTGH
Q99HS4_DPOL_HBVF3 GKQGRSGRIRSW....VHTPTRWPVGVESTGTGCAYNIASRSASCFHQSAVREKTNPSLSTSKRHSSTGH
Q8JMZ7_DPOL_HBVH3 GKQGRSGRLWSR....VHTPTRWPSGVEPSGTGHSDNLATRSTSRFHQSEVRKETNPSLSTSKGHTSTGH
Q8JNO8_DPOL_HBVH2 GKQGRSGRLWSR....VHTPTRWPSGVEPSGTGHSDNLATRSTSRFHQSEVRKETNPSLSTSKGHTSTGH
Q8JMY7_DPOL_HBVH1 GKQGRSGRLRSR....VHTPTRWPSGVEPSGTGHSDNLATRSTSCFHQSEVRKKANPSLSTSKGHTSTGH
O71304_DPOL_WMHBV SPRHGSGGLWSR....TSATPWRPSGVEFTSSGFVCHSARHPSSSINQSRQRKETNTSYSSSERHSPTSH
P03162_DPOL_DHBV1 RVRPCANNGGDK...HPPESGSLACWGGKESRIIK.SDSSRDSSAPVDSRGRPKSTRSFSPLSRRKTTGN
P0C691_DPOL_DHBV3 RIRPCANNGGDK...HPPATGSLACWGRKASRVIK.SGSSRDSSASVDSRRRSKGPRGFSTLPRRETTGN
P17192_DPOL_HPBDB SIRPCANNGGGK...HYATTRRLACWGGKTIGTDQ.SYSSRDTSATVDSRGRSESSRGFSTISGRKATGN
Q66403_DPOL_DHBVQ SIRPCANNGGGK...HSSTTGRLACWGGKTIGTDQ.SYSSRDASASVDSRGRSKSSRGFSSISRRKATGN
P17193_DPOL_HPBDW PIRSCANNGGGK...YSSTTRRNACWGGQTIGVNQ.SCSPRTSAASVDAGRRSKSSRGLSSLSRRETTGN
P30028_DPOL_HPBDC SIRPCANNGGDK...HSPTTRRLACWGGKASRINQ.SYSSRDSSAPVDARGRSESAGSLSSLSGRKAAGN
P13846_DPOL_HNBV STRDGSHRLSGR...TSDPTSRGALAGGDSTPIGPGSTAAHPSTHHVDRRRRQKGQGVLQAISREPSETR


340
350
360
370
380
390
400
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVC0
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAW8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JNO8_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WMHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_HPBDB
Q66403_DPOL_DHBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_HHBV

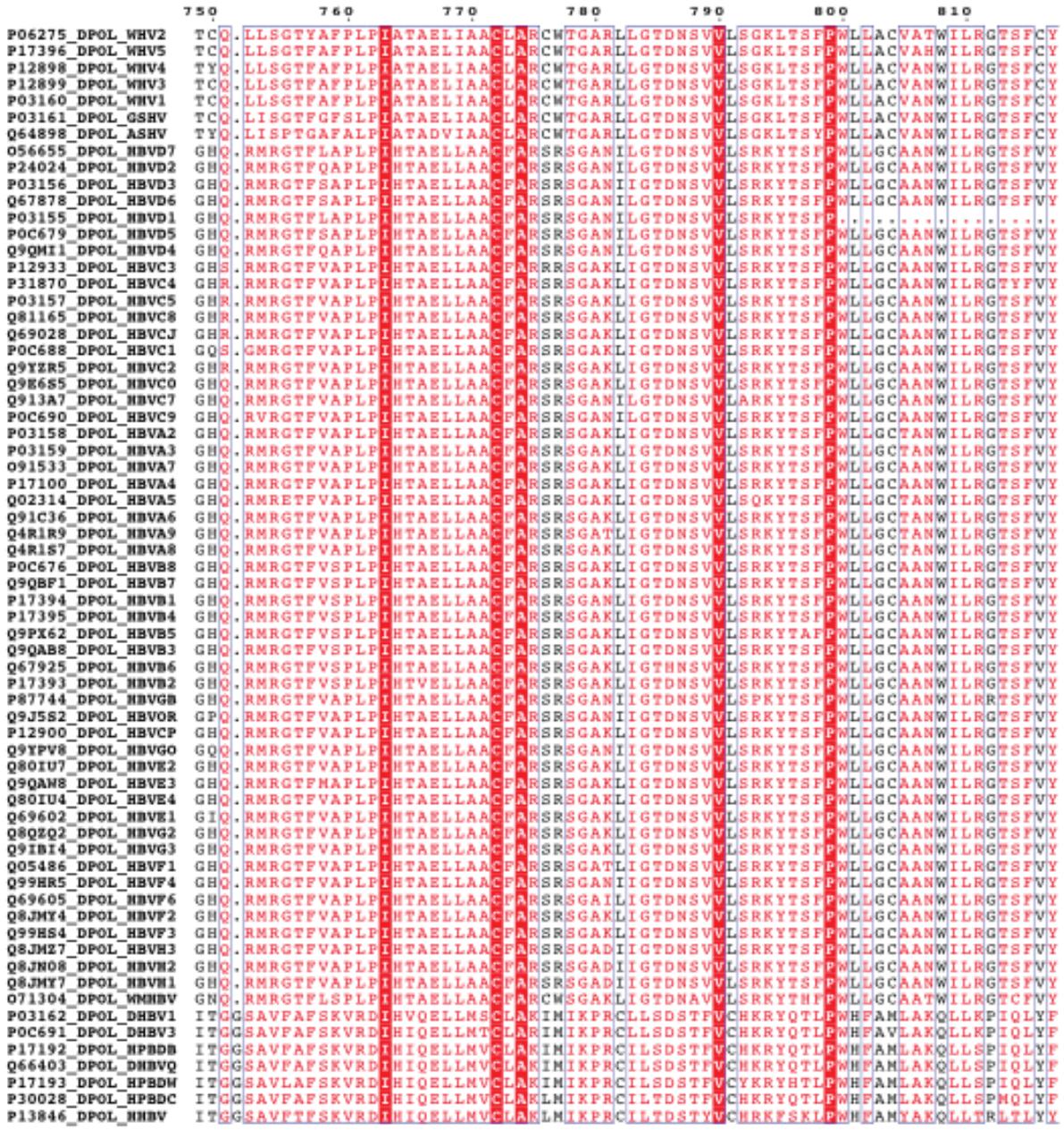

410
420
430
440
450
460
470
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVC0
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAW8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JN08_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WMHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_NPBDB
Q66403_DPOL_DNBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_MNBV


480
490
500
510
520
530
540
P06275_DPOL_WHV2 IPISPAAVPHLLVGSPGLERFNTCLSS....STHNRNNSQLQTMHNLCTRHVYSSLLLLFKTYGRKLHLL
P17396_DPOL_WHV5 IPISPAAVPHLLVGSPGLERFYTCLSS....STHNRNNSQLQTMHNLCTRHVYSSLLLLFKTYGRKLHLL
P12898_DPOL_WHV4 IPISPAAVPHLLVGSPGLERFNTCLSS....STHNRNNSQLQTMHNLCTRHVYSSLLLLFKTYGRKLHLL
P12899_DPOL_WHV3 IPISPAAVPHLLVGSPGLERFNTCMSS....STHNGNDSQLQTMHALCTRHVYSSLLLLFKTYGRKLHLL
P03160_DPOL_WHV1 IPISPAAVPHLLVGSPGLERFNTCLSY....STHNRNDSQLQTMHNLCTRHVYSSLLLLFKTYGRKLHLL
P03161_DPOL_GSHV IPVSPAAVPHFLVGSPGLERFASCMSH....DASNRNNSKLQTMHHICSRHLYNTLLLLFKTYGRKLHLL
Q64898_DPOL_ASHV IPVSPAAVPHLLVGSPGLERFTPSMSH....TTIHGNNSKLQTMHNLCSRHLFNSLLLLFKTYGRKLHLL
O56655_DPOL_HBVD7 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIFNHQRGTMQNLHDYCSRNLYVSLLLLYKTFGRKLHLY
P24024_DPOL_HBVD2 LPLHPAAMPHLLVGSSGLSRYVARLSS..DSRIFNHQHGTMQNLHDSCSRNLYVSLLLLYQTFGRKLHLY
P03156_DPOL_HBVD3 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRILNNQHGTMPDLHDYCSRNLYVSLLLLYQTFGRKLHLY
Q67878_DPOL_HBVD6 LPLHPAAMPHLLVGSSGVSRYVARLSS..NSRNNNNQYGTMQNLHDSCSRQLYVSLMLLYQNFGWKLHLY
P03155_DPOL_HBVD1 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQHGILQNLHDSCSRNLYVSLLLLYKTFGWKLHLY
P0C679_DPOL_HBVD5 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQHGTLQNLHDSCSRNLYVSLLLLYKTFGWKLHLY
Q9QMI1_DPOL_HBVD4 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIFDHQHGTMQNLHDYCSRNLYVSLLLLYQTFGRKLHLY
P12933_DPOL_HBVC3 IPLHPAAMPHLLVGSSGLPRYVARLSS..TSRNINYQHGTMQDLHDSCSRNLYVSLLLLYKTFGRKLHLY
P31870_DPOL_HBVC4 IPLHPAAMPHLLVGSSGLPRYVARLSS..TSRNINHQHGTMQDLHDSCSRNLYVSLLLLYKTFGRKLHLY
P03157_DPOL_HBVC5 IPLHPAAMPHLLVGSSGLPRYVARLSS..TSRNINHQHGAMQDLHDSCSRNLYVSLLLLYKTFGRKLHLY
Q81165_DPOL_HBVC8 IPLHPAAMPHLLVGSSGLPRYVARLSS..TSRNINYQHGTMQDLHDSCSRNLYVSLLLLYKTFGRKLHLY
Q69028_DPOL_HBVCJ IPLHPAAMPHLLVGSSGLPRYVARLSS..TSRNINYQHGTMQNLHDSCSRNLYVSLLLLYKTFGRKLHLY
P0C688_DPOL_HBVC1 IPLHPAAMPHLLVGSSGLPRYVARLSS..TSRNINYQHGTMQDLHDSCSRNLYVSLLLLYKTFGRKLHLY
Q9YZR5_DPOL_HBVC2 IPLHPAAMPHLLVGSSGLPRYVARLSS..TSRNINYQHGTMQDLHDSCSRNLYVSLLLLYKTFGRKLHLY
Q9E6S5_DPOL_HBVC0 IPLHPAAMPHLLVGSSGLPRYVARLSS..NSRNINNKHGTMQDLHDSCSRHLYVSLLLLYKTFGRKLHLY
Q913A7_DPOL_HBVC7 IPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIFNHQHGTLQNLHDSCSRNLYVSLLLLYKTFGRKLHLY
P0C690_DPOL_HBVC9 LPLHPAAMPHLLVGSSGLSRYVARLSS..TSRINDHQHGTLQNLHDHCSRNLYVSLMLLYKTFGRKLHLY
P03158_DPOL_HBVA2 IPLHPAAMPHLLIGSSGLSRYVARLSS..NSRINNNQYGTMQNLHDSCSRQLFVSLMLLYKTYGWKLHLY
P03159_DPOL_HBVA3 IPLHPAAMPHLLIGSSGLSRYVARLSS..NSRINNNQYGTMQNLHDSCSRQLYVSLMLLYKTYGWKLHLY
O91533_DPOL_HBVA7 IPLHPAAMPHLLIGSSGLSRYVARLSS..NSRINNNQYGTMQNLHDSCSRQLYVSLMLLYKTYGWKLHLY
P17100_DPOL_HBVA4 IPLHPAAMPHLLIGSSGLSRYVARLSS..NSRINNNQYGTMQNLHDSCSRQLYVSLMLLYKTYGWKLHLY
Q02314_DPOL_HBVA5 IPLHPAAMPHLLIGSSGLSRYVARLSS..NSRINNNQHGTLQNLHDSCSRQLYVSLMLLYKTYGWKLHLY
Q91C36_DPOL_HBVA6 IPLHPAAMPHLLIGSSGLSRYVARLSS..NSRIHNNQYGTLQNLHDSCSRQLYVSLMLLYKTYGWKLHLY
Q4R1R9_DPOL_HBVA9 IPLHPAAMPHLLIGSSGLSRYVARLSS..NSRIHNHQYGTLQNLHDSCSRQLYVSLMLLYKTYGRKLHLY
Q4R1S7_DPOL_HBVA8 IPLHPAAMPHLLIGSSGLSRYVARLSS..NSRIHNHQYGTLQNLHDSCSRQLYVSLMLLYKTYGRKLHLY
P0C676_DPOL_HBVB8 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQHGTMQDLHNSCSRNLYVSLMLLYKTYGRKLHLY
Q9QBF1_DPOL_HBVB7 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQHGTMQDLHNSCSRNLYVSLMLLYKTYGRKLHLY
P17394_DPOL_HBVB1 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQHGTMQDLHNSCSRNLYVSLMLLYKTYGWKLHLY
P17395_DPOL_HBVB4 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINNQHRTMQNLHNSCSRNLYVSLMLLYKTYGWKLHLY
Q9PX62_DPOL_HBVB5 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINNQHRTMQNLHNSCSRNLYVSLMLLYKTYGRKLHLY
Q9QAB8_DPOL_HBVB3 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINNQHRTMQNLHNSCSRNLYVSLMLLYKTYGRKLHLY
Q67925_DPOL_HBVB6 LPLHPAAMPHLLVGSSGLSRYVARLSS..HSRINNNQHGTMQNLHNSCSRNLYVSLMLLYKTYGRKLHLY
P17393_DPOL_HBVB2 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQHRTMQNLHDSCSRNLYVSLMLLYKTYGRKLHLY
P87744_DPOL_HBVGB LPLHPAAMPHLLVGSSGLSRYVARLSS..TSRIIDHQHGTMQNLHDHCSRNLFVSLMLLYKTFGRKLHLY
Q9J5S2_DPOL_HBVOR LPLHPAAMPHLLVGSSGLPRYVARLSS..TSRNHHHQRGTMQNLHDFCSRNLFVSLMLLYKTFGRKLHLY
P12900_DPOL_HBVCP LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRILDHQHGTMQNLHDSCSRNLFDSLMLLYKTFGRKLHLY
Q9YPV8_DPOL_HBVGO LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIHDHQHGTMQNLHNYCTRNLFVSLMLLYKTFGRKLHLY
Q80IU7_DPOL_HBVE2 IPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQYGTLPNLHDSCSRNLYVSLMLLFKTFGRKLHLY
Q9QAN8_DPOL_HBVE3 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQYGTLPNLHDSCSRNLYVSLMLLFKTFGRKLHLY
Q80IU4_DPOL_HBVE4 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHHYGTLPNLHDSCSRNLYVSLMLLFKTFGRKLHLY
Q69602_DPOL_HBVE1 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIINHQHGTLPNLHDSCSRNLYVSLMLLFKTFGRKLHLY
Q8QZQ2_DPOL_HBVG2 IPLHPAAMPHLLVGSSGLSRYVARLSS..DSRILDHQYGTLQNLHDSCSRQLYVSLMLLYKTFGRKLHLY
Q9IBI4_DPOL_HBVG3 IPLHPAAMPHLLVGSSGLSRYVARLSS..DSRILDHQYGTLQNLHDSCSRQLYVSLMLLYKTFGRKLHLY
Q05486_DPOL_HBVF1 LPLHPAAMPHLLVGSSGLSRYVARLSS..TSRIHDHQHGTLQNLHNSCTRNLYVSLLLLFQTLGRKLHLY
Q99HR5_DPOL_HBVF4 LPLHPAAMPHLLVGSSGLSRYVARLSS..NSRIYDHQHGTMQNLHNSCSRNLYVSLLLLFQTLGRKLHLY
Q69605_DPOL_HBVF6 LPLHPAAMPHLLVGSSGLPRYVARLSS..TSRIHDHQHGTMQNLHSSCSRNLYVSLLLLFQTLGRKLHLY
Q8JMY4_DPOL_HBVF2 LPLHPAAMPHLLVGSSGLSRYVARLSS..TSRIHDHQHGTMQNLHNSCSRNLYVSLLLLFKTFGRKLHLY
Q99HS4_DPOL_HBVF3 LPLHPAAMPHLLVGSCGLSRYVARLSS..TSRIHDHQHGTMQNLHNSCSRNLYVSLLLLFQTLGRKLHLY
Q8JMZ7_DPOL_HBVH3 LPLHPAAMPHLLVGSSGLSRYVARVSS..TSRIYNHQHGSLQNLHHSCSRNLYVSLLLLYQTFGRKLHLY
Q8JN08_DPOL_HBVH2 LPLHPAAMPHLLVGSSGLSRYVARVSS..TSRIYNHQHGTLQNLHHSCSRNLYVSLLLLYQTFGRKLHLY
Q8JMY7_DPOL_HBVH1 LPLHPAAMPHLLVGSSGLSRYVARVSS..TSRIYNHQHGTLQNLHHSCSRNLYVSLLLLYQTFGRKLHLY
O71304_DPOL_WMHBV LPLHPASMPHLLVGSSGLPRYVARVSSSTNSYRNNNNNGTLQDLHANCSRHLFVSLMLLYQTYGRKLHLY
P03162_DPOL_DHBV1 LPLNPASSSRLAVSDG.....
P0C691_DPOL_DHBV3 LPLNPASSSRLAVSDG.....
P17192_DPOL_HPBDB LPLNPASSSRLAVSDG.....
Q66403_DPOL_DHBVQ LPLNPASSSRLAVSDG.....
P17193_DPOL_HPBDW LPLNPASSSRLAVSDG.....
P30028_DPOL_HPBDC LPLNPASSSRLAVSDG.....
P13846_DPOL_HNBV LPLAPASSSRLAVSDG.....


550
560
570
580
590
600
610
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVC0
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAN8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JN08_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WMHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_HPBDB
Q66403_DPOL_DHBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_HHBV

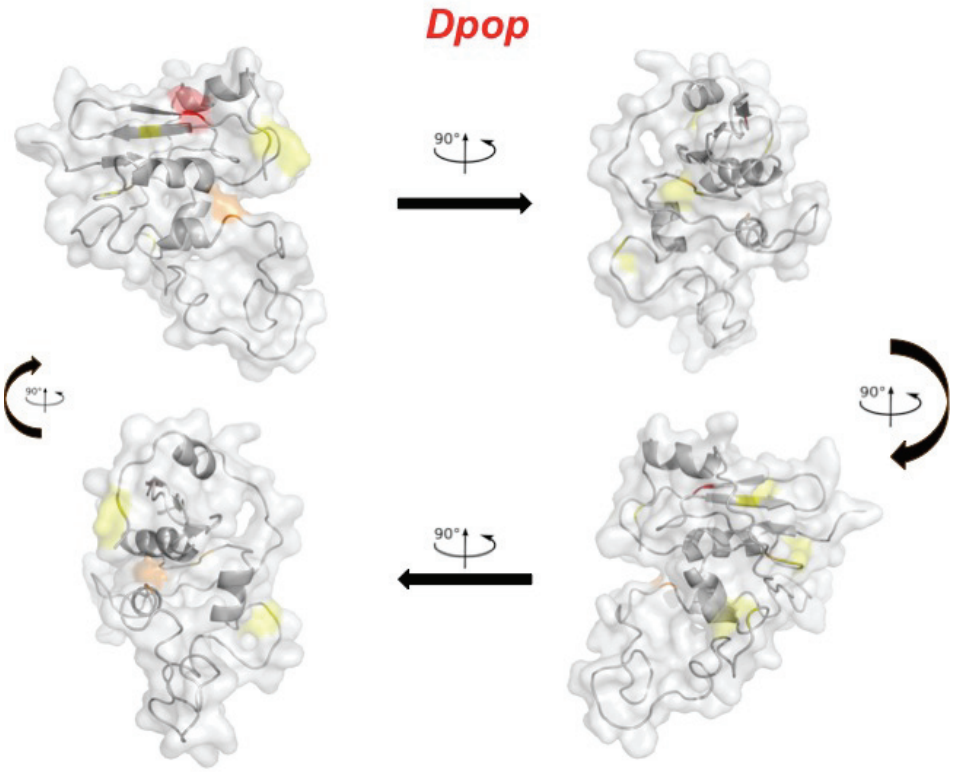

620
630
640
650
660
670
680
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVC0
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAN8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JN08_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WHHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_HPBDB
Q66403_DPOL_DHBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_HHBV

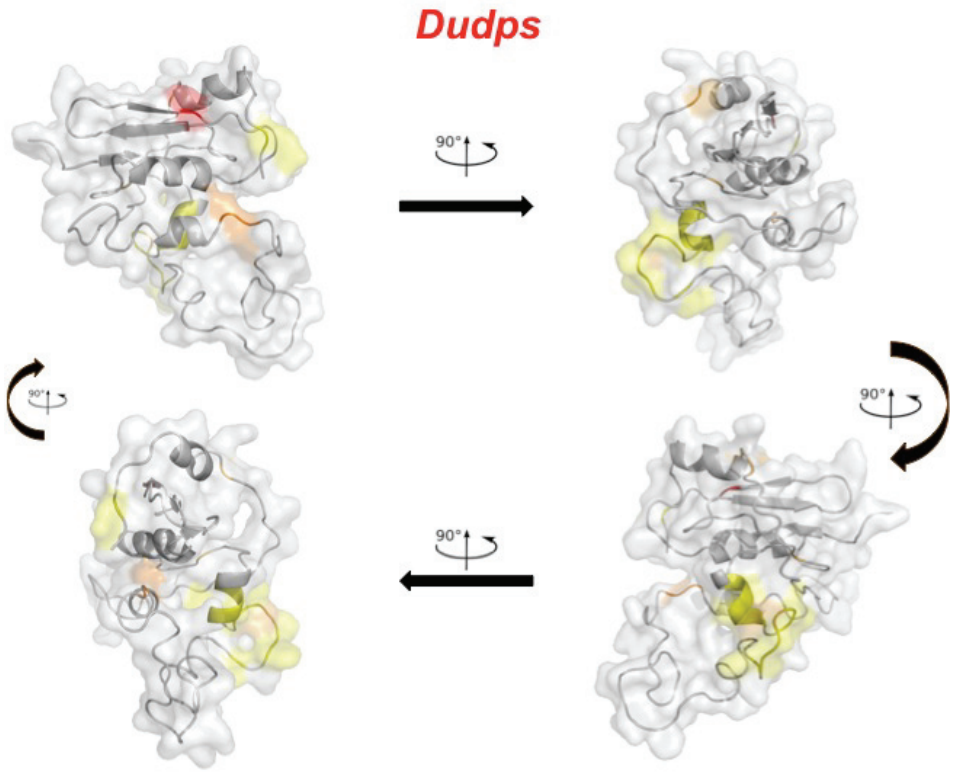

690
700
710
720
730
740
P06275_DPOL_WHV2
P17396_DPOL_WHV5
P12898_DPOL_WHV4
P12899_DPOL_WHV3
P03160_DPOL_WHV1
P03161_DPOL_GSHV
Q64898_DPOL_ASHV
O56655_DPOL_HBVD7
P24024_DPOL_HBVD2
P03156_DPOL_HBVD3
Q67878_DPOL_HBVD6
P03155_DPOL_HBVD1
P0C679_DPOL_HBVD5
Q9QMI1_DPOL_HBVD4
P12933_DPOL_HBVC3
P31870_DPOL_HBVC4
P03157_DPOL_HBVC5
Q81165_DPOL_HBVC8
Q69028_DPOL_HBVCJ
P0C688_DPOL_HBVC1
Q9YZR5_DPOL_HBVC2
Q9E6S5_DPOL_HBVCO
Q913A7_DPOL_HBVC7
P0C690_DPOL_HBVC9
P03158_DPOL_HBVA2
P03159_DPOL_HBVA3
O91533_DPOL_HBVA7
P17100_DPOL_HBVA4
Q02314_DPOL_HBVA5
Q91C36_DPOL_HBVA6
Q4R1R9_DPOL_HBVA9
Q4R1S7_DPOL_HBVA8
P0C676_DPOL_HBVB8
Q9QBF1_DPOL_HBVB7
P17394_DPOL_HBVB1
P17395_DPOL_HBVB4
Q9PX62_DPOL_HBVB5
Q9QAB8_DPOL_HBVB3
Q67925_DPOL_HBVB6
P17393_DPOL_HBVB2
P87744_DPOL_HBVGB
Q9J5S2_DPOL_HBVOR
P12900_DPOL_HBVCP
Q9YPV8_DPOL_HBVGO
Q80IU7_DPOL_HBVE2
Q9QAN8_DPOL_HBVE3
Q80IU4_DPOL_HBVE4
Q69602_DPOL_HBVE1
Q8QZQ2_DPOL_HBVG2
Q9IBI4_DPOL_HBVG3
Q05486_DPOL_HBVF1
Q99HR5_DPOL_HBVF4
Q69605_DPOL_HBVF6
Q8JMY4_DPOL_HBVF2
Q99HS4_DPOL_HBVF3
Q8JMZ7_DPOL_HBVH3
Q8JN08_DPOL_HBVH2
Q8JMY7_DPOL_HBVH1
O71304_DPOL_WMHBV
P03162_DPOL_DHBV1
P0C691_DPOL_DHBV3
P17192_DPOL_HPBDB
Q66403_DPOL_DHBVQ
P17193_DPOL_HPBDW
P30028_DPOL_HPBDC
P13846_DPOL_HHBV

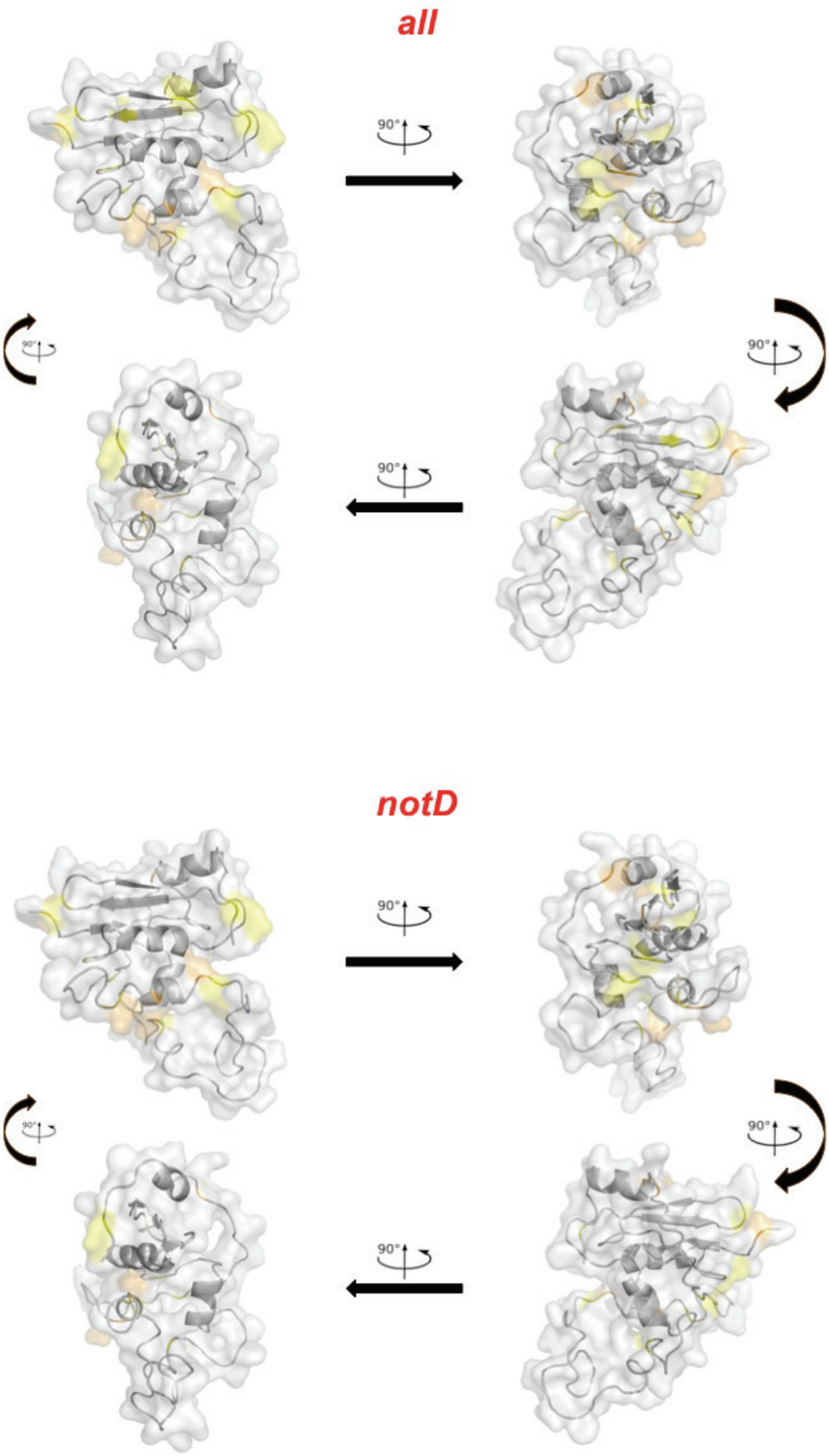

750
760
770
780
790
800
810
P06275_DPOL_WHV2 TCQ.LLSGTYAFPLPIATAELIAACLARCWTGARLLGTDNSVVLSGKLTSFPWLLACVATWILRGTSFCY
P17396_DPOL_WHV5 TCQ.LLSGTFAFPLPIATAELIAACLARCWTGARLLGTDNSVVLSGKLTSFPWLLACVAHWILRGTSFCY
P12898_DPOL_WHV4 TYQ.LLSGTFAFPLPIATAELIAACLARCWTGARLLGTDNSVVLSGKLTSFPWLLACVANWILRGTSFCY
P12899_DPOL_WHV3 TCQ.LLSGTFAFPLPIATAELIAACLARCWTGARLLGTDNSVVLSGKLTSFPWLLACVANWILRGTSFCY
P03160_DPOL_WHV1 TCQ.LLSGTFAFPLPIATAELIAACLARCWTGARLLGTDNSVVLSGKLTSFPWLLACVANWILRGTSFCY
P03161_DPOL_GSHV TCQ.LISGTFGFSLPIATAELIAACLARCWTGARLLGTDNSVVLSGKLTSFPWLLACVANWILRGTSFCY
Q64898_DPOL_ASHV TYQ.LISPTGAFALPIATADVIAACLARCWTGARLLGTDNSVVLSGKLTSYPWLLACVANWILRGTSFCY
O56655_DPOL_HBVD7 GHQ.RMRGTFLAPLPIHTAELLAACFARSRSGANILGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P24024_DPOL_HBVD2 GHQ.RMRGTFQAPLPIHTAELLAACFARSRSGANILGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P03156_DPOL_HBVD3 GHQ.RMRGTFSAPLPIHTAELLAACFARSRSGANIIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q67878_DPOL_HBVD6 GHQ.RMRGTFSAPLPIHTAELLAACFARSRSGANILGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P03155_DPOL_HBVD1 GHQ.RMRGTFLAPLPIHTAELLAACFARSRSGANILGTDNSVVLSRKYTSFP..............
P0C679_DPOL_HBVD5 GHQ.RMRGTFSAPLPIHTAELLAACFARSRSGANILGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q9QMI1_DPOL_HBVD4 GHQ.RMRGTFQAPLPIHTAELLAACFARSRSGANILGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P12933_DPOL_HBVC3 GHS.RMRGTFVAPLPIHTAELLAACFARRRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P31870_DPOL_HBVC4 GHR.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTYFVY
P03157_DPOL_HBVC5 GHR.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q81165_DPOL_HBVC8 GHR.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q69028_DPOL_HBVCJ GHR.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P0C688_DPOL_HBVC1 GQS.GMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q9YZR5_DPOL_HBVC2 GHR.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q9E6S5_DPOL_HBVC0 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q913A7_DPOL_HBVC7 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGANILGTDNSVVLARKYTSFPWLLGCAANWILRGTSFVY
P0C690_DPOL_HBVC9 GHQ.RVRGTFVAPLPIHTAELLAACFARSRSGANILGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P03158_DPOL_HBVA2 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCTANWILRGTSFVY
P03159_DPOL_HBVA3 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCTANWILRGTSFVY
O91533_DPOL_HBVA7 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P17100_DPOL_HBVA4 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q02314_DPOL_HBVA5 GHQ.RMRETFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSQKYTSFPWLLGCTANWILRGTSFVY
Q91C36_DPOL_HBVA6 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCTANWILRGTSFVY
Q4R1R9_DPOL_HBVA9 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGATLIGTDNSVVLSRKYTSFPWLLGCTANWILRGTSFVY
Q4R1S7_DPOL_HBVA8 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCTANWILRGTSFVY
P0C676_DPOL_HBVB8 GHQ.RMRGTFVSPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q9QBF1_DPOL_HBVB7 GHQ.RMRGTFVSPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P17394_DPOL_HBVB1 GHQ.RMRGTFVSPLPIHTAELLAACFARSRSGANLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P17395_DPOL_HBVB4 GHQ.RMRGTFVSPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q9PX62_DPOL_HBVB5 GHQ.RMRGTFVSPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTAFPWLLGCAANWILRGTSFVY
Q9QAB8_DPOL_HBVB3 GHQ.RMRGTFVSPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q67925_DPOL_HBVB6 GHQ.RMRGTFVSPLPIHTAELLAACFARSRSGAKLIGTHNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P17393_DPOL_HBVB2 GHQ.RMRGTFVSPLPIHTVELLAACWARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P87744_DPOL_HBVGB GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGANIIGTDNSVVLSPKYTSFPWLLGCAANWILRRTSFVY
Q9J5S2_DPOL_HBVOR GPQ.RMRGTFVAPLPIHTAELLAACFARSRSGANIIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
P12900_DPOL_HBVCP GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q9YPV8_DPOL_HBVGO GQQ.RMRGTFVAPLPIHTAELLAACFARSRSGANIIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q80IU7_DPOL_HBVE2 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q9QAN8_DPOL_HBVE3 GHQ.RMRGTFMAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q80IU4_DPOL_HBVE4 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q69602_DPOL_HBVE1 GIQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q8QZQ2_DPOL_HBVG2 GHQ.RMRGTFVAPLPIHTAELLAACWARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q9IBI4_DPOL_HBVG3 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q05486_DPOL_HBVF1 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGATLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q99HR5_DPOL_HBVF4 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGANIIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q69605_DPOL_HBVF6 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAILIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q8JMY4_DPOL_HBVF2 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q99HS4_DPOL_HBVF3 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGAKLIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q8JMZ7_DPOL_HBVH3 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGADIIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q8JNO8_DPOL_HBVH2 GHQ.RMRGTFVAPLPIHTAELLAACFARSRSGADIIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
Q8JMY7_DPOL_HBVH1 GHQ.RMRGTFVAPLPIHTAELLAACWARSRSGADIIGTDNSVVLSRKYTSFPWLLGCAANWILRGTSFVY
O71304_DPOL_WMHBV GNQ.RMRGTFLSPLPIHTAELLAACFARCWSGAKLIGTDNAVVLSRKYTHFPWLLGCAATWILRGTCFVY
P03162_DPOL_DHBV1 ITGGSAVFAFSKVRDIHVQELLMSCLAKIMIKPRCLLSDSTFVCHKRYQTLPWHFAMLAKQLLKPIQLYF
P0C691_DPOL_DHBV3 ITGGSAVFAFSKVRDIHIQELLMTCLARIMIKPRCLLSDSTFVCHKRYQTLPWHFAVLAKQLLKPIQLYF
P17192_DPOL_HPBDB ITGGSAVFAFSKVRDIHIQELLMVCLAKIMIKPRCILSDSTFVCHKRYQTLPWHFAMLAKQLLSPIQLYF
Q66403_DPOL_DHBVQ ITGGSAVFAFSKVRDIHIQELLMVCLAKIMIKPRCILSDSTFVCHKRYQTLPWHFAMLAKQLLSPIQLYF
P17193_DPOL_HPBDW ITGGSAVLAFSKVRDIHIQELLMVCLAKIMIKPRCILSDSTFVCYKRYHTLPWHFAMLAKQLLSPIQLYF
P30028_DPOL_HPBDC ITGGSAVFAFSKVRDIHIQELLMVCLAKLMIKPRCILTDSTFVCHKRYQTLPWHFAMLAKQLLSPMQLYF
P13846_DPOL_HNBV ITGGSAVFTFSKVRDIHIQELLMVCLAKLMIKPRCILTDSTYVCHRKFSKLPWHFAMYAKQLLTRLTLYY


820 830 840 850 860 870 880
P06275_DPOL_WHV2 VPSALNPADLPSRGLLPALRPLPRLRLRPQTTRISLWAASPPVSPRAPVRVAWSSPVQTCEPWIPP
P17396_DPOL_WHV5 VPSALNPADLPSRGLLPALRPLPRLRLRPQTSRISLWAASPPVSPRRPVRVAWSSPVQTCEPWIPP
P12898_DPOL_WHV4 VPSALNPADLPSRGLLPVLRPLPRLRLRPQTSRISLWAASPPVSPRRPVRVAWSSPVQNCEPWIPP
P12899_DPOL_WHV3 VPSALNPADLPSRGLLPVLRPLPRLRLRPQTSRISLWAASPPVSPRRPVRVAWSSPVQTCEPWIPP
P03160_DPOL_WHV1 VPSALNPADLPSRGLLPVLRPLPRLRFRPPTSRISLWAASPPVSPRRPVRVAWSSPVQNCEPWIPP
P03161_DPOL_GSHV VPSADNPADLPSRGLLPALRPLPLLRFRPVTKRISLWAASPPVSTRRPVRVAWASPVQTCEPWIPP
Q64898_DPOL_ASHV VPSAANPADLPSRGLLPALHPVPTLRFRPQLSRISLWAASPPVSPRRPVRVAWASPVQNSEPWFPP
O56655_DPOL_HBVD7 VPSALNPADDPSRGRLGLSRPLLRLPFRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
P24024_DPOL_HBVD2 VPSALNPADDPSRGRLGLSRPLLRLPFRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
P03156_DPOL_HBVD3 VPSALNPADDPSRGRLGLSRPLLRLPFRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
Q67878_DPOL_HBVD6 VPSALNPADDPSRGRLGLSRPLLCLPFRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
P03155_DPOL_HBVD1 ....
P0C679_DPOL_HBVD5 VPSALNPADDPSRGRLGPCRPLLHLPFRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
Q9QMI1_DPOL_HBVD4 VPSALNPADDPSRGRLGLCRPLLRLPFRPTTGRTSLYAVSPSVPSHLPDHVHFASPLH..VAWRPP
P12933_DPOL_HBVC3 VPSALNPADDPSRGRLGLYRPLLHLPFRPTTGRTSLYAVFLSVPSHLPVRVHFASPLH..VAWRPP
P31870_DPOL_HBVC4 VPSALNPADDPSRGRLGLIRPLLHLRFRPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
P03157_DPOL_HBVC5 VPSALNPADDPSRGRLGLYRPLLLLPFRPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
Q81165_DPOL_HBVC8 VPSALNPADDPSRGRLGLYRPLLHLPFRPTTGRTSLYAVSPSVPSHLPSRVHFASPLH..VAWRPP
Q69028_DPOL_HBVCJ VPSALNPADDPSRGRLGLYRPLLHLPFRPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
P0C688_DPOL_HBVC1 VPSALNPADDPSRGRLGLYRPLLHLPFRPTTGRASLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
Q9YZR5_DPOL_HBVC2 VPSALNPADDPSRGRLGLYRPLLRLPFQPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWKPP
Q9E6S5_DPOL_HBVC0 VPSALNPADDPSRGRLGLYRPLLRLPFRPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
Q913A7_DPOL_HBVC7 VPSALNPADDPSRGRLGLYRPLLHLPFRPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWKPP
P0C690_DPOL_HBVC9 VPSALNPADDPSRGRLGIYRPLLRLPFRPSTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
P03158_DPOL_HBVA2 VPSALNPADDPSRGRLGLSRPLLRLPFQPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
P03159_DPOL_HBVA3 VPSALNPADDPSRGRLGLSRPLLRLPFQPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
O91533_DPOL_HBVA7 VPSALNPADDPSRGRLGLYRPLLRLPFQPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
P17100_DPOL_HBVA4 VPSALNPADDPSRGRLGLSRPLLRLPFQPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
Q02314_DPOL_HBVA5 VPSALNPADDPSRGRLGLYRPLLRLPYRPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
Q91C36_DPOL_HBVA6 VPSALNPADDPSRGRLGLYRPLLRLPYRPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
Q4R1R9_DPOL_HBVA9 VPSALNPADDPSRGRLGLYRPLLRLPYRPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
Q4R1S7_DPOL_HBVA8 VPSALNPADDPSRGRLGLYRPLLRLPFRPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
P0C676_DPOL_HBVB8 VPSALNPADDPSRGRLGLYRPLLRLPYRPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
Q9QBF1_DPOL_HBVB7 VPSALNPADDPSRGRLGLYRPLLRLPYRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
P17394_DPOL_HBVB1 VPSALNPADDPSRGRLGLYRPLLRLPYRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
P17395_DPOL_HBVB4 VPSALNPADDPSRGRLGLYRPLLRLLYRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
Q9PX62_DPOL_HBVB5 VPSALNPADDPSRGRLGLYRPLLRLLYRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWXPP
Q9QAB8_DPOL_HBVB3 VPSALNPADDPSRGRLGLYRPLLRLLYRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
Q67925_DPOL_HBVB6 VPSALNPADDPSRGRLGLYRPLLRLPYQPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
P17393_DPOL_HBVB2 VPSALNPADDPSRGRLGLYRPLLRLPYRPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
P87744_DPOL_HBVGB VPSALNPADDPSRGRLGLYRPLLRPWFRPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
Q9J5S2_DPOL_HBVOR VPSALNPADDPSRGRLGLYRPLLRLPFRPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
P12900_DPOL_HBVCP VPSALNPADDPSRGRLGLYRPLIRLLFQPTTGRTSLYAVSPSVPSHLPVRVHFASPLH..VAWRPP
Q9YPV8_DPOL_HBVGO VPSALNPADDPSRGRLGLSRPLCRLPFQPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
Q80IU7_DPOL_HBVE2 VPSALNPADDPSRGRLGIYRPLLRLPFQPTTGRTSLYAVSPSVPSRLPDRVHFASPLH..VAWRPP
Q9QAW8_DPOL_HBVE3 VPSALNPADDPSRGRLGIYRPLLRLPFQPSTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
Q80IU4_DPOL_HBVE4 VPSALNPADDPSRGRLGVCRPLLRLPFQPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
Q69602_DPOL_HBVE1 VPSALNPADDPSRGRLGIFRPLLRLRFRPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VAWRPP
Q8QZQ2_DPOL_HBVG2 VPSALNPADDPSRGRLGLCRPLLRLPFLPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VTWKPP
Q9IBI4_DPOL_HBVG3 VPSALNPADDPSRGRLGLCRPLLRLPFLPTTGRTSLYAVSPSVPSHLPDRVHFASPLH..VTWKPP
Q05486_DPOL_HBVF1 VPSALNPADDPSRGRLGLYRPLLRLPFQPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
Q99HR5_DPOL_HBVF4 VPSALNPADDPSRGRLGLYRPLLRLPFQPTTGRTSLYADSPSVPSHLPVRVHFASPLH..VAWRPP
Q69605_DPOL_HBVF6 VPSALNPADEPSRGRLGLYRPLLRLPFQPTTGRTSLYADSPSVPSHLPDRVHFASPLH..VAWRPP
Q8JMY4_DPOL_HBVF2 VPSALNPADDPSRGRLGLYRPLLRLPFQPTTGRTSLYAASPSVPSHLPDRVHFASPLH..VAWRPP
Q99HS4_DPOL_HBVF3 VPSALNPADDPSRGRLGLYRPLLRLPFQPTTGRTSLYAASPSVPSHLPDRVHFASPLH..VAWRPP
Q8JMZ7_DPOL_HBVH3 VPSALNPADDPSRGRLGLCRPLLRLPFRPTTGRTSLYADSPPVPSHLPARVHFASPLH..VAWRPP
Q8JN08_DPOL_HBVH2 VPSALNPADDPSRGRLGLCRPLLRLPFRPTTGRTSLYADSPPVPFHQPARVHFGSPLH..VAWRPP
Q8JMY7_DPOL_HBVH1 VPSALNPADDPSRGRLGLCRPLLRLPFRPTTGRTSLYADSPPVPSHLPARVHFASPLH..VAWRPP
O71304_DPOL_WMHBV VPSKLNPADDPSRGCLGLLKPLPRLLFQPSTGRTSLYAVSPPVPFHRPGRVLFASPLQPGDAWRPP
P03162_DPOL_DHBV1 VPSKYNPADGPSR.....HKPP..DWTAFPYTPLSKAIYIPHRLCGT.....
P0C691_DPOL_DHBV3 VPSKYNPADGPSR.....HRPP..DWTAFPYTPLSKAIYIPHRLCGT.....
P17192_DPOL_HPBDB VPSKYNPADGPSR.....HKPP..DWTALTYTPLSKAIYIPHRLCGT.....
Q66403_DPOL_DHBVQ VPSKYNPADGPSR.....HRPP..DWTALTYTPLSKAIYIPHRLCGT.....
P17193_DPOL_HPBDW VPSKYNPADGPSR.....HKPP..DWTALTYTPLSKAIYIPHRLCGT.....
P30028_DPOL_HPBDC VPSKYNPADGPSR.....HKPP..DWTALTYTPLSKAIYIPHRLCGT.....
P13846_DPOL_HNBV VPSKYNPADGPTR.....HKPP..DWTAVTYTPLSKHIYIPHRLCGL.....

Annexe 11







Annexe 11 : Cartographie sur le modèle de RNase H de l'entropie de Shannon normalisée pour les 4 jeux de données.
Les valeurs supérieures à 0,1 sont représentées selon le gradient de couleur défini dans la partie matériel et méthodes (4.2.4). Qautre vues sont présentées pour les 4 jeux de données.

Annexe 12

Alignement de Das *et al.*

```
pdb1t05:A       EKEG---------KISKIGPENPYNTPVFAIKKKDS---TKWRKLVDFRELNKR--TQDF
Q05486          EDWGPCYEHGQHYIRTPRT--PARVTGGVFLVDKNPHNTTESRLVVDFSQFSRGTTRVSW
                *. *           :      .  *  . : .*:.   *: * :*** ::.:     .:

pdb1t05:A       WEVQL-GIPHPAGLKKK-KSVTVLDVGDAYFSVPLD------------------------
Q05486          PKFAVPNLQSLTNLLSSNLSWLSLDVSAAFYHLPLHPAAMPHLLVGSSGLSRYVARLSST
                 :. : .:   :.* ..  *   ***. *:: :**.

pdb1t05:A       --EDFRKYTAFTI---------------------PSINNETPGIRYQYNVLPQGWKGSPA
Q05486          SRIHDHQHGTLQNLHNSCTRNLYVSLLLLFQTLGRKLHLYSHPIILGFRKIPMGVGLSPF
                   . ::: ::                        .::  :  *   :. :* *   **

pdb1t05:A       IFQSSMTKILEPFKK-QNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWGLTTP
Q05486          LLAQ-FTSAICSVVRRAFPHCLAFSYMDDLVLGAK-SVQHLESLYTAVTNFLLSVGIHLN
                :: . :*. : .. :   *. : :.***** :*:. .: : .:    : :.**  *:

pdb1t05:A       DKKH-QKEPPFLWMGYELHPDK--WTVQPIVLPE--KDSWTVN-DI-----CKLVGKLNW
Q05486          TSKTKRWGYSLHFMGYVIGSWGSLPQDHIVHKIKECFRKLPVNRPIDWKVCQRIVGLLGF
                 .*  :   .: :*** : .       : :   :    . .**  *      ::** *.:

pdb1t05:A       ASQI--YPGIKVRQLSKLLRGTK-ALTEV-IPLTEEAELELAENREILKEPVHG
Q05486          AAPFTQCGYPALMPLYACITAKQAFVFSPTKAFLCKQYMNLYPVARQRPGLCQV
                *: :       :  *   : ..:  : .   .:  :  ::*    .      :

```

Alignement de Daga *et al.*

```
pdb1t05:A       TEM---EKEGKISKIGPENPYNTP--VFAIKKK--DSTKWRKLVDFRELNKRTQDFWEVQ
Q05486          EDWGPCYEHGQHYIRTPRTPARVTGGVFLVDKNPHNTTESRLVVDFSQFSRGTTRVSWPK
                 :     :.*:     *..* ...  ** :.*:  ::*: * :*** ::.: *  .   :

pdb1t05:A       LGIPHPAGLK----KKKSVTVLDVGDAYFSVPLD-----------EDFRKYTA-FTIPSI
Q05486          FAVPNLQSLTNLLSSNLSWLSLDVSAAFYHLPLHPAAMPHLLVGSSGLSRYVARLSSTSR
                :.:*:  .*.    .: *   ***. *:: :**.           ..: :*.* :: .*

pdb1t05:A       NN--------------------------ETPGIRYQ---------YNVLPQGWKGSPAIF
Q05486          IHDHQHGTLQNLHNSCTRNLYVSLLLLFQTLGRKLHLYSHPIILGFRKIPMGVGLSPFLL
                 :                          :* * : :         :. :* *   ** ::

pdb1t05:A       QSSMTKILEPFKKQNPDIVIYQYMDDLYVGSDLEIGQHRTKI-EELRQHLLRWGL-TTPD
Q05486          AQFTSAICSVVRRAFPHCLAFSYMDDLVLGA--KSVQHLESLYTAVTNFLLSVGIHLNTS
                 .  : * . .::  *. : :.***** :*:  :  **  .:   : :.**  *:  ...

pdb1t05:A       KKHQKEPPFLWMGYELH-----PDKWTVQPI-----VLPEKDSWTVNDICKLVGKLNWAS
Q05486          KTKRWGYSLHFMGYVIGSWGSLPQDHIVHKIKECFRKLPVNRPIDWKVCQRIVGLLGFAA
                *.::   .: :*** :      *:.  *: *      ** : .   :   ::** *.:*:

pdb1t05:A       QI----YPGIKVRQLSKLLRGTKALTEVIPLTEEAELELAENREILK-EP-VHGV
Q05486          PFTQCGYPALMPLYAC--ITAKQAFVFSPTYKAFLCKQYMNLYPVARQRPGLCQV
                 :    **.:     .  : ..:*:.   . .     :  :   : : .* :  *
```

Alignement de Bartholomeusz *et al.*

```
pdb1t05:A       ALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKD--STKWRKLVDFRELNKRTQDFWEV
Q05486          EDWGPCY--EHGQHYIRTPRTPARVTGGVFLVDKNPHNTTESRLVVDFSQFSRGTTRVSW
                     *   *:  :  :     .  *  . : .*:  .*.  :**   .  .*     .

pdb1t05:A       -QLGIPHPAG----LKKKKSVTVLDVGDAYFSVPLDEDFRKYTAFTIPSINNETP-----
Q05486          PKFAVPNLQSLTNLLSSNLSWLSLDVSAAFYHLPLHPAAMPHLLVGSSGLSRYVARLSST
                 ::.:*:  .    *..: *   ***. *:: :**.     :  .  ..:.. ..

pdb1t05:A       ----------------------------------------GIRYQYNVLPQGWKGSPA
Q05486          SRIHDHQHGTLQNLHNSCTRNLYVSLLLLFQTLGRKLHLYSHPIILGFRKIPMGVGLSPF
                                                       *   :. :* *   **

pdb1t05:A       IFQSSMTKILEPFKKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWGLTTPD
Q05486          LLAQFTSAICSVVRRAFPHCLAFSYMDDLVLGAKSVQH-LESLYTAVTNFLLSVGIHLNT
                :: .  : * . .::  *. : :.***** :*:.      .:    : :.**  *:

pdb1t05:A       KKHQK-EPPFLWMGYE--LH
Q05486          SKTKRWGYSLHFMGYVIGSW
                .* ::   .: :***
```

Annexe 12 : Alignement des séquences de RT VIH et VHB de Das, Daga et Bartholomeusz.
(Das et al., 2001; Daga et al., 2010; Bartholomeusz et al., 2004)

Annexe 13

N° du modèle	Valeur énergie (kcal/mol)	RMSD avec l'empreinte (Å)	Nombre de violations de contraintes	Résidus dans la région favorable (%)
1	-15598,2	2,17	21 (0,2 %)	84,7
2	-15653,1	2,44	19 (0,2 %)	85,6
3	-15627,9	2,30	22 (0,3 %)	85,7
4	-11807,3	5,89	64 (0,7 %)	80,9
5*	-13550,4	**1,41**	22 (0,3 %)	86,4
6	-13407,9	1,66	19 (0,2 %)	83,7
7	**-15764,6**	2,08	20 (0,2 %)	**87,3**
8	-15487,3	2,11	19 (0,2 %)	**87,3**
9	-13926,7	1,75	20 (0,2 %)	84,9
10	-13936,7	1,53	**16 (0,2 %)**	84,7

Annexe 13 : Valeurs des critères de sélection pour les 10 modèles de RT-RNase H générés.
(gras : meilleure valeur, rouge : modèle éliminé de l'analyse, vert* : modèle sélectionné.)

Articles

Article 1

HBVdb : A knowledge database for Hepatitis B Virus.

Juliette Hayer, Fanny Jadeau, Gilbert Deléage, Alan Kay, Fabien Zoulim and Christophe Combet

Nucleic Acid Research,
Advance Access published November 3, 2012

D566–D570 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 3 November 2012*
doi:10.1093/nar/gks1022

HBVdb: a knowledge database for Hepatitis B Virus

Juliette Hayer[1], Fanny Jadeau[1], Gilbert Deléage[1], Alan Kay[2], Fabien Zoulim[2,3] and Christophe Combet[1,*]

[1]Unité Bases Moléculaires et Structurales des Systèmes Infectieux; UMR 5086 CNRS - Université Claude Bernard Lyon 1; IBCP FR 3302 - 7, passage du Vercors, 69367 Lyon CEDEX 07, [2]INSERM, U1052, Viral Hepatitis Research Laboratory; Université Lyon 1, 151, cours Albert Thomas, 69003 Lyon, and [3]Hospices Civils de Lyon, Hepatology Department, 69004 Lyon, France

Received August 14, 2012; Revised September 28, 2012; Accepted October 3, 2012

ABSTRACT

We have developed a specialized database, HBVdb (http://hbvdb.ibcp.fr), allowing the researchers to investigate the genetic variability of Hepatitis B Virus (HBV) and viral resistance to treatment. HBV is a major health problem worldwide with more than 350 million individuals being chronically infected. HBV is an enveloped DNA virus that replicates by reverse transcription of an RNA intermediate. HBV genome is optimized, being circular and encoding four overlapping reading frames. Indeed, each nucleotide of the genome takes part in the coding of at least one protein. However, HBV shows some genome variability leading to at least eight different genotypes and recombinant forms. The main drugs used to treat infected patients are nucleos(t)ides analogs (reverse transcriptase inhibitors). Unfortunately, HBV mutants resistant to these drugs may be selected and be responsible for treatment failure. HBVdb contains a collection of computer-annotated sequences based on manually annotated reference genomes. The database can be accessed through a web interface that allows static and dynamic queries and offers integrated generic sequence analysis tools and specialized analysis tools (e.g. annotation, genotyping, drug resistance profiling).

INTRODUCTION

Hepatitis B virus (HBV) is a major health problem worldwide with more than 350 million people being chronic carriers. Chronic HBV infection is associated with a significantly increased risk of developing severe liver diseases, including liver cirrhosis, and hepatocellular carcinoma (HCC), one of the most common forms of human cancer. The estimated risk of HCC in chronic HBV carriers is ~100 times greater than in uninfected individuals (1). Currently available anti-HBV drugs have limitations. Indeed, interferon alpha administration is associated with adverse reactions, while nucleos(t)ide analogs are virostatic and require long-term administration (2,3).

HBV is an enveloped DNA virus that belongs to the *Hepadnaviridae*, a family of hepatotropic DNA viruses infecting certain mammalian or avian hosts (4). It contains a small (~3.2 kb), partially double-stranded relaxed-circular DNA (rcDNA) genome that replicates by reverse transcription of an RNA intermediate, the pregenomic RNA (pgRNA). The genome encodes four overlapping open reading frames (ORFs) that are translated to produce the viral core protein (5,6), the surface proteins (5,7), a polymerase/reverse transcriptase (RT) (2,4), and HBx (8,9). The HBV life cycle starts with the binding of the virus to an unknown receptor of the host cell. Then, the viral particle is internalized. The virion rcDNA is delivered to the nucleus, where it is repaired to form a covalently closed-circular cccDNA. The episomal cccDNA serves as the template for the transcription of the pgRNA and the other viral mRNAs by the host RNA polymerase II (10).

The viral genome is variable because of the spontaneous error rate of the viral polymerase and the lack of proof reading activity. There are eight genotypes of HBV designated A–H based on >8% nt variation over the entire genome. The eight HBV genotypes are distributed in distinct geographical localizations (11–13). Recombinant forms involving different genotypes have also been reported (14). However, the extensive overlap between the four encoded ORFs limits the diversity that the virus can tolerate. Indeed, every nucleotide participates in the coding of at least one viral protein attesting of an optimized small genome (15,16). Moreover, the genome variability is also constrained by environmental pressure exerted by the host immune response and the antiviral drugs for treated patients.

To allow researchers to investigate the genetic variability of HBV sequences and viral resistance to treatment, several databases and repositories (17–19) have been published to date. Moreover, tools specific to HBV genotyping (20–22) are available for virologists, as well

*To whom correspondence should be addressed. Tel: +33 4 37 65 29 47; Fax: +33 4 72 72 26 04; Email: christophe.combet@ibcp.fr

as tools aimed at drug resistance mutations analysis, among which some are freely accessible (23) and others need a registration (24). We have developed HBVdb, a database that contains a collection of computer-annotated HBV sequences thanks to manually annotated reference genomes. The sequences taken as input are the ones publicly available in the INSDC (25), including partial and complete HBV genomes. The database can be queried via a web interface and the query results can be further analysed with the numerous integrated generic and specialized (e.g. annotation, genotyping, drug resistance profiling) analysis tools.

DATABASE BUILDING

We developed a fully automated procedure to annotate all HBV sequences from the European Nucleotide Archive (ENA) (26), using a reference set of 16 manually annotated and non-recombinant complete genomes representing the 8 genotypes. The HBVdb building process starts with the retrieval of all the HBV entries in ENA. The second step is the automatic annotation of these entries in their text format (flat file ENA format). The annotated HBVdb entries are then loaded into a PostgreSQL relational database system. Finally, sequence datasets are extracted and multiple sequence alignments together with associated data are computed.

The HBVdb is updated on a monthly basis. The program for the automatic annotation, as well as for the querying and the management of the database, are implemented in Java and SQL programming languages.

ANNOTATION PROCEDURE

A standard numbering system of HBV genomes exists, defined by the (often hypothetical) EcoR1 restriction site as the origin of the genome (27). However, the circularity of the HBV genome leads to the deposit of sequences in generic public databases that do not follow this system. Such sequences will result in one or several partial pairwise alignments with the reference genomes. To circumvent this problem, we modeled the HBV by duplicating the sequence of each reference genome (from ~3.2 kb to ~6.4 kb). The first step of the automated annotation procedure is a similarity search, using the FASTA program (28) (Figure 1A) in order to identify the most similar reference genome to the sequence to annotate. The second step of the annotation procedure checks if the query sequence follows the EcoR1 numbering by looking if the query sequence is aligned to only one replicate or if it overlaps both. In the latter case, the part of the query sequence that is aligned on the second replicate is shifted to the corresponding region of the first replicate. If there are gaps between the shifted part and the fixed part, they are replaced by 'n'. This checking ensures that all the sequences follow the EcoR1 numbering system. The third step consists in optimizing the global query-reference pairwise alignment in order to avoid erroneous translation or to detect non-functional sequences. For each coding sequence (CDS) found in the query, a pairwise alignment is produced from the global one and divided into non-overlapping 3-nt windows (i.e. codons). If a window contains one or two gaps, the process tries to optimize gaps in order to have only 3 nt in the window or three gaps (codon deletion). If the optimization fails, the entry is discarded from the database. In the fourth step, the reference genome features (e.g. *CDS*, *mat_peptide*) are mapped onto the sequence to annotate when they are present. The sequence is genotyped in a fifth step. The last step corresponds to the drug resistance profiling. Finally, the annotated HBV entry is formatted as an ENA text entry for its later inclusion in the relational database.

GENOTYPING

Starting from all the pairwise query-reference alignments (Figure 1B), the algorithm computes a matrix containing the identity percentage at each position of the query for each genotype. The identity percentage is computed by summing the identity percentage over overlapping sliding windows (window length 301 nt, window step 1 nt) divided by the number of windows used at each query position. The maximum identity percentage calculated for each query position is taken from the matrix to fill an array with the corresponding genotype letter, as long as the maximum mean identity percentage is above or equal to 90%. The genotype is then computed from this array. The procedure is able to process only sequences with lengths equal or above the window size. Overall, the genotyping algorithm allows the identification of 'pure' genotype sequence as well as recombinant ones (14,29). In the result file, users can find the number of informative positions used in the genotype computation. This value can be used as a confidence value of the genotype prediction. The accuracy of our genotyping tool is similar to Oxford (20,30), jpHMM (21) and NCBI (22) HBV genotyping tools, including recombinant genomes detection. The HepSEQ genotyping tool (17) uses only polymerase/surface genes in genotype computation and is less accurate in recombinant detection (Supplementary Tables S1 and S2).

DRUG RESISTANCE PROFILING

An alignment between the query protein sequence and a reference reverse transcriptase (RT) sequence is computed. The algorithm searches, in the query sequence, for mutations defining known resistance profiles to lamivudine, telbivudine, entecavir, adefovir and tenofovir drugs (31). If all the mutations of one profile are found, the profile is reported with the associated drug and resistance status. There are three possible resistance statuses designated by 'Sensitive' (S), 'Intermediate' (I, reduced susceptibility) and 'Resistant' (R). The algorithm output provides the detected profile, the status and mutation positions in the query sequence and according to the RT numbering system (32). In its current implementation, our resistance tool does not look for antiviral drug-associated potential vaccine-escape mutants [ADAPVEM (33)] contrarily to the HepSeq polymerase annotator (17).

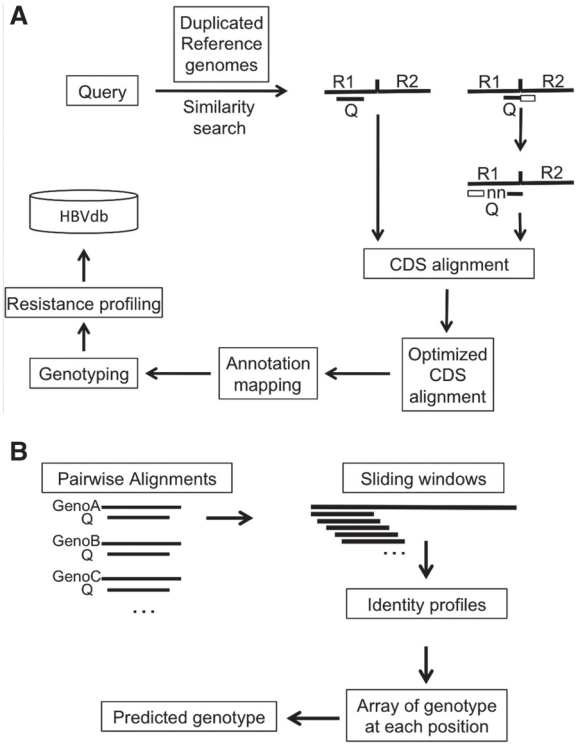


Figure 1. Annotation and genotyping processes. (**A**) The annotation procedure starts with the computation of a pairwise alignment between the query sequence (Q) and the most similar duplicated reference genome sequence (R1, R2: replicate 1, 2; nn indicates missing nucleotides between shifted and fixed parts of the query sequence). This alignment is split up into CDS alignments that are optimized before the mapping of features and the transfer of annotations. (**B**) Genotyping process. These pairwise alignments between query (Q) and reference sequences (e.g. GenoA, genotype A reference genome) are iterated using sliding windows to compute a matrix of mean identities. The matrix is used to produce the array of genotype at each query sequence position. Predicted genotype is deduced from this array.

DATABASE CONTENT

The text format of an HBVdb entry is an extension of the ENA-Annotation format as used for the euHCVdb (34). Some elements of the ENA-Annotation entry are conserved such as the accession number, the organism name, the creation date, and the references. After the annotation procedure, some elements are corrected and/or completed in the entry, mainly in the features. Indeed, a set of new qualifiers that store specific data is added to some features. The qualifier *PRABI_genotype* is added to the *source* feature to indicate a provisional genotype predicted by the genotyping tool. The qualifier *PRABI_name* is added to each feature to ensure standard names across all the database entries. Concerning the protein annotations, some qualifiers are added to the *mat_peptide* features. These qualifiers, noted *PRABI_prodft*, follow the feature table format of the UniProtKB database (35). The *PRABI_prodft* qualifiers describe the protein chains, the protein domains, some sites like active sites (*act_site*) that designate the catalytic residues of enzymes, and the resistance mutations (*res_mut*) with the drug and the resistance status.

WEB INTERFACE

The HBVdb is accessible through a website (Figure 2A) divided into two parts. In the static part, the user can find general information about HBV and the nomenclature used through genome organization, protein descriptions (Figure 2B), the reference genomes and the genotypes. The user can also access pre-computed 'Nucleotide' and 'Protein' datasets sorted according to the genotype (rows) and the protein or CDS names (columns). The datasets provide full-length sequences in Pearson/Fasta format, their multiple sequence alignments computed with Muscle (36) and displayed in Clustal W format (Figure 2C), and the corresponding residue repertoires with Shannon entropies that are useful for analysing conserved/variable alignment positions. The user can download the corresponding files for further analysis. Furthermore, the alignments can be interactively edited with the 'EditAlignment' applet developed by our team. In the dynamic part, users can extract their own dataset by combining multiple criteria (e.g. genotype and sequence length and protein name). The datasets can be exported as Pearson/Fasta sequences, accession number lists and entry flat files for further analysis with the integrated analysis tools.

The available analysis tools are either generic or specialized. The generic analysis tools (e.g. BLAST (37) or Clustal W (38)) are available through the NPS@ server (39), that is an integrated sequence analysis web server. 'Annotate' (Figure 2D), 'Genotype' and 'Resistance' (Figure 2E) specialized tools allow the analysis of one or several HBV sequences uploaded by the user. The annotation of a nucleotide sequence produces a result page listing the CDS found in the query sequence, presenting the predicted genotype, and giving the drug resistance status (with a link to resistance output file) if the nucleotide sequence contains the CDS of the RT domain. The results page also gives access to the global pairwise alignment between the query sequence and the reference genome, as well as the pairwise alignments for each CDS. The user can also access the text entry format of the annotated sequence. The annotation of a protein sequence ends up with a result page indicating the most similar sequence used for its annotation, with links to the entry and the pairwise alignment, and the resistance status if the sequence contains the RT domain. The 'Genotype' tool allows the user to genotype nucleotide sequences. It produces a result page giving the predicted genotype, and the genotype computed at each position of the query sequence. The 'Resistance' tool enables the detection of known drug resistance mutations from nucleotide or protein sequences. The output lists the drug, the resistance status (R, I, S) or 'n.a.' if the query sequence does not contain the HBV RT domain, and the identified mutations.

STATISTICS

HBVdb is available since June 2012. The release 4 (September 2012) comprises 39 289 sequences, including 3606 complete genomes.

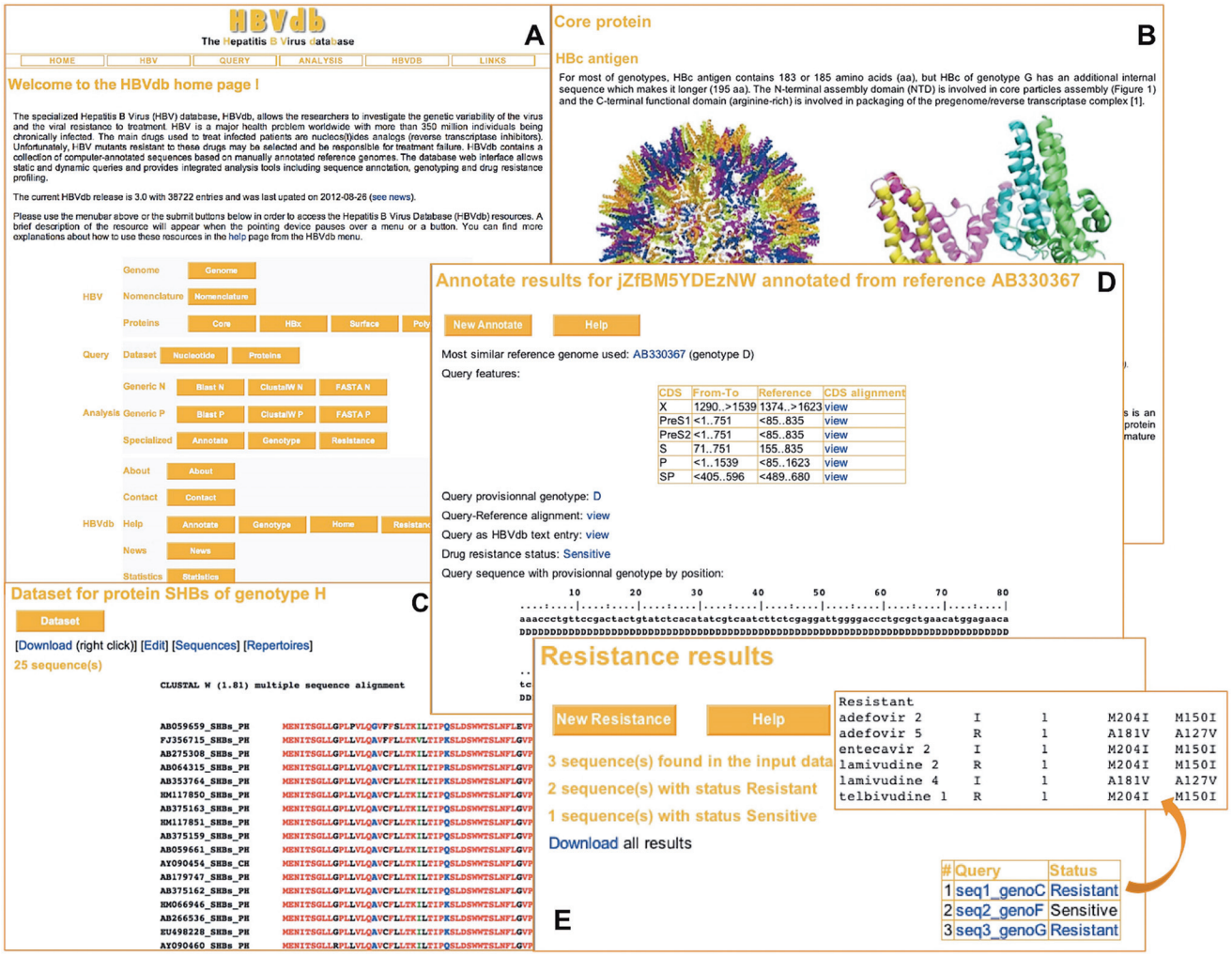


Figure 2. Snapshot of the HBVdb web interface. (**A**) The HBVdb homepage, with the menu bar and all the menus repeated as buttons for ease-of-use. (**B**) The HBV core protein information page. (**C**) The multiple sequence alignment of the genotype H SHBs proteins. (**D**) The *Annotate* tool detailed result page for an input nucleotide sequence showing the CDS positions with links to their alignment, the genotype and the resistance status. (**E**) The *Resistance* tool result page showing the resistance status and the resistance profile.

CONCLUSIONS AND PERSPECTIVES

HBVdb is a collection of human HBV sequences, annotated thanks to a computer-automated system. The automatic annotation process used to generate HBVdb entries guarantees standard annotations and allows efficient keyword searches. The HBVdb website allows researchers to access pre-computed datasets and to extract data thanks to a dynamic query system. The extracted data can be further analysed with a set of bioinformatics tools available in the NPS@ server. HBVdb website also enables users to annotate their own sequences through the HBVdb automatic annotation process. In the future, the annotation will be enriched with new data (e.g. mapping of promoters) and analysis tools (e.g. ADAPVEM detection) will be added. The HBVdb and related programs will be useful to face the analysis of the flood of sequences that are generated thanks to the Next-Generation Sequencing (NGS) technologies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2.

ACKNOWLEDGEMENTS

We acknowledge members of the French ANRS hepatitis network for stimulating discussions through the last years in order to make HBVdb a useful tool for the researchers.

FUNDING

J.H. is recipient of a doctoral fellowship from the Ministère de l'Enseignement Supérieur et de la Recherche. HBVdb is hosted on the Pôle Rhône-Alpes de BioInformatique platform funded by the Groupement d'Intérêt Scientifique Infrastructures en Biologie Sante et Agronomie (GIS IBiSA). Funding for open access

charge: Groupement d'Intérêt Scientifique Infrastructures en Biologie Sante et Agronomie (GIS IBiSA).

Conflict of interest statement. None declared.

REFERENCES

1. Nguyen,D.H., Ludgate,L. and Hu,J. (2008) Hepatitis B virus-cell interactions and pathogenesis. *J. Cell. Physiol.*, **216**, 289–294.
2. Zoulim,F. and Locarnini,S. (2009) Hepatitis B virus resistance to nucleos(t)ide analogues. *Gastroenterology*, **137**, 1593–608.e1-2.
3. European Association For The Study Of The Liver. (2009) EASL Clinical Practice Guidelines: management of chronic hepatitis B. *J. Hepatol.*, **50**, 227–242.
4. Nassal,M. (2008) Hepatitis B viruses: reverse transcription a different way. *Virus Res.*, **134**, 235–249.
5. Bruss,V. (2004) Envelopment of the Hepatitis B Virus nucleocapsid. *Virus Res.*, **106**, 199–209.
6. Wynne,S.A., Crowther,R.A. and Leslie,A.G.W. (1999) The crystal structure of the human Hepatitis B Virus capsid. *Mol. Cell*, **3**, 771–780.
7. Glebe,D. and Urban,S. (2007) Viral and cellular determinants involved in hepadnaviral entry. *World J. Gastroenterol.*, **13**, 22–38.
8. Benhenda,S., Cougot,D., Buendia,M.-A. and Neuveut,C. (2009) Chapter 4 Hepatitis B Virus X Protein: molecular functions and its role in virus life cycle and pathogenesis. In: Woude,G.F.V. and Klein,G. (eds), *Advances in Cancer Research*. Academic Press, Vol. 103, pp. 75–109.
9. Bouchard,M.J. and Schneider,R.J. (2004) The enigmatic X gene of hepatitis B virus. *J. Virol.*, **78**, 12725–12734.
10. Beck,J. and Nassal,M. (2007) Hepatitis B virus replication. *World J Gastroenterol*, **13**, 48–64.
11. Bartholomeusz,A., Tehan,B.G. and Chalmers,D.K. (2004) Comparisons of the HBV and HIV polymerase, and antiviral resistance mutations. *Antivir. Ther.*, **9**, 149–160.
12. Schaefer,S. (2007) Hepatitis B Virus taxonomy and Hepatitis B Virus genotypes. *World J. Gastroenterol.*, **13**, 14–21.
13. Norder,H., Couroucé,A.M., Coursaget,P., Echevarria,J.M., Lee,S.D., Mushahwar,I.K., Robertson,B.H., Locarnini,S. and Magnius,L.O. (2004) Genetic diversity of Hepatitis B Virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. *Intervirology*, **47**, 289–309.
14. Simmonds,P. and Midgley,S. (2005) Recombination in the genesis and evolution of Hepatitis B Virus genotypes. *J. Virol.*, **79**, 15467–15476.
15. Kay,A. and Zoulim,F. (2007) Hepatitis B Virus genetic variability and evolution. *Virus Res.*, **127**, 164–176.
16. Araujo,N.M., Waizbort,R. and Kay,A. (2011) Hepatitis B Virus infection from an evolutionary point of view: how viral, host, and environmental factors shape genotypes and subgenotypes. *Infect. Genet. Evol.*, **11**, 1199–1207.
17. Gnaneshan,S., Ijaz,S., Moran,J., Ramsay,M. and Green,J. (2007) HepSEQ: International Public Health Repository for hepatitis B. *Nucleic Acids Res.*, **35**, D367–D370.
18. Shin-I,T., Tanaka,Y., Tateno,Y. and Mizokami,M. (2008) Development and public release of a comprehensive hepatitis virus database. *Hepatol. Res.*, **38**, 234–243.
19. Panjaworayan,N., Roessner,S.K., Firth,A.E. and Brown,C.M. (2007) HBVRegDB: annotation, comparison, detection and visualization of regulatory elements in Hepatitis B Virus sequences. *Virol. J.*, **4**, 136.
20. Alcantara,L.C., Cassol,S., Libin,P., Deforche,K., Pybus,O.G., Van Ranst,M., Galvão-Castro,B., Vandamme,A.M. and de Oliveira,T. (2009) A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res.*, **37**, W634–W642.
21. Schultz,A.K., Bulla,I., Abdou-Chekaraou,M., Gordien,E., Morgenstern,B., Zoaulim,F., Dény,P. and Stanke,M. (2012) jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. *Nucleic Acids Res.*, **40**, W193–W198.
22. Rozanov,M., Plikat,U., Chappey,C., Kochergin,A. and Tatusova,T. (2004) A web-based genotyping resource for viral sequences. *Nucleic Acids Res.*, **32**, W654–W659.
23. Rhee,S.Y., Margeridon-Thermet,S., Nguyen,M.H., Liu,T.F., Kagan,R.M., Beggel,B., Verheyen,J., Kaiser,R. and Shafer,R.W. (2010) Hepatitis B Virus reverse transcriptase sequence variant database for sequence analysis and mutation discovery. *Antiviral Res.*, **88**, 269–275.
24. Yuen,L.K., Ayres,A., Littlejohn,M., Colledge,D., Edgely,A., Maskill,W.J., Locarnini,S.A. and Bartholomeusz,A. (2007) SeqHepB: a sequence analysis program and relational database system for chronic hepatitis B. *Antiviral Res.*, **75**, 64–74.
25. Karsch-Mizrachi,I., Nakamura,Y. and Cochrane,G. (2012). International Nucleotide Sequence Database Collaboration. (2012) The International Nucleotide Sequence Database collaboration. *Nucleic Acids Res.*, **40**, D33–D37.
26. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tárraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.
27. Ono,Y., Onda,H., Sasada,R., Igarashi,K., Sugino,Y. and Nishioka,K. (1983) The complete nucleotide sequences of the cloned Hepatitis B Virus DNA; subtype adr and adw. *Nucleic Acids Res.*, **11**, 1747–1757.
28. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
29. Kramvis,A., Arakawa,K., Yu,M.C., Nogueira,R., Stram,D.O. and Kew,M.C. (2008) Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. *J. Med. Virol.*, **80**, 27–46.
30. de Oliveira,T., Deforche,K., Cassol,S., Salminen,M., Paraskevis,D., Seebregts,C., Snoeck,J., van Rensburg,E.J., Wensing,A.M., Vijver,D.A. *et al.* (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**, 3797–3800.
31. European Association for the Study of the Liver. (2012) EASL Clinical Practice Guidelines: management of chronic Hepatitis B Virus infection. *J. Hepatol.*, **57**, 167–185.
32. Stuyver,L.J., Locarnini,S.A., Lok,A., Richman,D.D., Carman,W.F., Dienstag,J.L. and Schinazi,R.F. (2001) Nomenclature for antiviral-resistant human Hepatitis B Virus mutations in the polymerase region. *Hepatology*, **33**, 751–757.
33. Locarnini,S.A. and Yuen,L. (2010) Molecular genesis of drug-resistant and vaccine-escape HBV mutants. *Antivir. Ther.*, **15**, 451–461.
34. Combet,C., Garnier,N., Charavay,C., Grando,D., Crisan,D., Lopez,J., Dehne-Garcia,A., Geourjon,C., Bettler,E., Hulo,C. *et al.* (2007) euHCVdb: the European hepatitis C virus database. *Nucleic Acids Res.*, **35**, D363–D366.
35. UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
36. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
37. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
38. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
39. Combet,C., Blanchet,C., Geourjon,C. and Deleage,G. (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.*, **25**, 147.

Article 2

Mutations that alter use of hepatitis C virus cell entry factors mediate escape from neutralizing antibodies.

Isabel Fofana, Samira Fafi-Kremer, Patric Carolla, Catherine Fauvelle, Muhammad Nauman Zahid, Marine Turek, Laura Heydmann, Karine Cury, Juliette Hayer, Christophe Combet, François-Loïc Cosset, Thomas Pietschmann, Marie-Sophie Hiet, Ralf Bartenschlager, François Habersetzer, Michel Doffoël, Zhen-Yong Keck, Steven K H Foung, Mirjam B Zeisel, Françoise Stoll-Keller and Thomas F Baumert

Gastroenterology 143:1, 223-233.e9

GASTROENTEROLOGY 2012;143:223–233

Mutations That Alter Use of Hepatitis C Virus Cell Entry Factors Mediate Escape From Neutralizing Antibodies

ISABEL FOFANA,[*,‡] SAMIRA FAFI–KREMER,[*,‡,§] PATRIC CAROLLA,[*,‡] CATHERINE FAUVELLE,[*,‡] MUHAMMAD NAUMAN ZAHID,[*,‡] MARINE TUREK,[*,‡] LAURA HEYDMANN,[*,‡] KARINE CURY,[*,‡] JULIETTE HAYER,[‖] CHRISTOPHE COMBET,[‖] FRANÇOIS–LOÏC COSSET,[¶] THOMAS PIETSCHMANN,[#] MARIE–SOPHIE HIET,[**] RALF BARTENSCHLAGER,[**] FRANÇOIS HABERSETZER,[*,‡,‡‡] MICHEL DOFFOËL,[*,‡,‡‡] ZHEN–YONG KECK,[§§] STEVEN K. H. FOUNG,[§§] MIRJAM B. ZEISEL,[*,‡] FRANÇOISE STOLL–KELLER,[*,‡,§] and THOMAS F. BAUMERT[*,‡,‡‡]

[*]Inserm, U748, Strasbourg, France; [‡]Université de Strasbourg, Strasbourg, France; [§]Laboratoire de Virologie, [‡‡]Pôle Hepato-Digestif, Hôpitaux Universitaires de Strasbourg, Strasbourg, France; [‖]Bases Moléculaires et Structurales des Systèmes Infectieux, UMR 5086, Centre National de la Recherche Scientifique, Université de Lyon, Institut de Biologie et Chimie des Proteines, Lyon, France; [¶]Université de Lyon, Université Claude Bernard Lyon1, IFR 128, Inserm U758; Ecole Normale Supérieure de Lyon, 69364 Lyon, France; [#]Division of Experimental Virology, TWINCORE, Centre for Experimental and Clinical Infection Research, a joint venture between the Medical School Hannover and the Helmholtz Centre for Infection Research, Hannover, Germany; [**]The Department of Infectious Diseases, Molecular Virology, Heidelberg University, Heidelberg, Germany; and [§§]Department of Pathology, Stanford University School of Medicine, Stanford, California

BACKGROUND & AIMS: The development of vaccines and other strategies to prevent hepatitis C virus (HCV) infection is limited by rapid viral evasion. HCV entry is the first step of infection; this process involves several viral and host factors and is targeted by host-neutralizing responses. Although the roles of host factors in HCV entry have been well characterized, their involvement in evasion of immune responses is poorly understood. We used acute infection of liver graft as a model to investigate the molecular mechanisms of viral evasion. **METHODS:** We studied factors that contribute to evasion of host immune responses using patient-derived antibodies, HCV pseudoparticles, and cell culture–derived HCV that express viral envelopes from patients who have undergone liver transplantation. These viruses were used to infect hepatoma cell lines that express different levels of HCV entry factors. **RESULTS:** By using reverse genetic analyses, we identified altered use of host-cell entry factors as a mechanism by which HCV evades host immune responses. Mutations that alter use of the CD81 receptor also allowed the virus to escape neutralizing antibodies. Kinetic studies showed that these mutations affect virus–antibody interactions during postbinding steps of the HCV entry process. Functional studies with a large panel of patient-derived antibodies showed that this mechanism mediates viral escape, leading to persistent infection in general. **CONCLUSIONS: We identified a mechanism by which HCV evades host immune responses, in which use of cell entry factors evolves with escape from neutralizing antibodies. These findings advance our understanding of the pathogenesis of HCV infection and might be used to develop antiviral strategies and vaccines.**

Keywords: Virology; Liver Disease; Tissue Culture Model; Immunity.

Hepatitis C virus (HCV) infection is a major cause of liver disease.[1] A vaccine is not available and antiviral treatment is limited by resistance and adverse effects.[2] HCV-induced liver disease is a leading indication for liver transplantation (LT).[3] A major limitation of LT is the universal reinfection of the liver graft with accelerated recurrence of liver disease. A strategy to prevent reinfection is lacking.[3] Thus, there is an urgent unmet medical need for the development of efficient and safe antivirals and vaccines.

HCV entry is required for initiation, maintenance, and dissemination of infection. Viral entry is a key target for adaptive host responses and antiviral strategies.[4,5] Functional studies in clinical cohorts highlight that viral entry and escape from antibody-mediated neutralization play an important role in viral persistence and liver disease.[6–12] HCV entry is a highly orchestrated process mediated by viral envelope glycoproteins E1 and E2 and several host factors including heparan sulfate, CD81, scavenger receptor BI (SR-BI), claudin-1 (CLDN1), occludin (OCLN) (reviewed by Zeisel et al[5]), and kinases.[13] Although the role of E1E2 in antibody-mediated neutralization has been studied intensively,[4,5,14] the role of host factors for viral evasion in vivo is only poorly understood.

Acute graft infection is an established in vivo model to study viral evasion because viral infection and host-neutralizing responses can be monitored precisely.[8] Viral entry and escape from host-neutralizing responses are important determinants allowing the virus to rapidly infect the liver during transplantation.[8] However, the molecular mechanisms by which the virus evades host immunity to persistently reinfect the liver graft are unknown.

To uncover viral and host factors mediating enhanced viral entry and escape, we functionally analyzed genetically closely related prototype variants derived from a well-char-

Abbreviations used in this paper: CLDN, claudin; HCV, hepatitis C virus; HCVcc, cell culture-derived HC; HCVpp, hepatitis C virus pseudoparticles; HMAb, human monoclonal antibody; HVR, hypervariable region; LT, liver transplantation; mAb, monoclonal antibody; OCLN, occludin SR-BI, scavenger receptor class B type I; VA, variant A; VC, variant C; VL, variant L.

0016-5085/$36.00
http://dx.doi.org/10.1053/j.gastro.2012.04.006

acterized patient undergoing LT.[8] In one variant, P01VL, reinfecting the liver graft was characterized by high infectivity and escape from neutralizing antibodies present in autologous pretransplant serum.[8] The other closely related variants, P01VA and P01VC, were not selected during LT and were characterized by lower infectivity and high sensitivity to neutralization by autologous pretransplant serum.[8] Previous studies had indicated that an E2 region comprising amino acids 425–483 most likely contained mutations responsible for the phenotype of enhanced entry and viral evasion of variants reinfecting the liver graft.[8]

Materials and Methods

Patients

Evolution and functional analysis of viral variants of patient P01 have been described.[8] Anti-HCV–positive serum samples from patients undergoing transplantation and chronic HCV infection were obtained with approval from the Strasbourg University Hospital Institutional Review Board (ClinicalTrial.gov Identifiers NCT00638144 and NCT00213707).

Plasmids

Plasmids for HCV pseudoparticle (HCVpp) production of variants VL, VA, and VC have been described.[8] E1E2-encoding sequences were used as templates for individual and combinations of mutations using the QuikChange II XL site-directed mutagenesis kit (Agilent Technologies, Massy, France). Mutations were confirmed by DNA sequence analysis (GATC Biotech, Mulhouse, France) for the desired mutation and for exclusion of unexpected residue changes in the full-length E1E2 encoding sequences. Mutated constructs were designated X#Y, where # is the residue location in H77c,[15] X is the mutated amino acid, and Y is the original amino acid.

Antibodies

Monoclonal anti-E1 (11B7) and anti-E2 (AP33, IGH461, 16A6); human anti-HCV IgG[10,16]; human monoclonal antibodies (HMAbs) CBH-2, CBH-5, CBH-23, and HC-1 have been described.[9,17] Anti-CD81 (JS-81) was from BD Biosciences (Heidelberg, Germany), AP33 was from Genentech (San Francisco, CA), and 11B7, IGH461, and 16A6 were from Innogenetics (Ghent, Belgium).

Cell Lines

HEK 293T and Huh7.5.1 cells were cultured as described.[10,13,16] Huh7.5.1 cells overexpressing HCV entry factors were created by stable lentiviral gene transfer of CLDN1, OCLN, SR-BI, or CD81.[18] Huh7.5 stably transduced with retroviral vectors encoding for CD81- and CD13-specific short hairpin (sh) RNAs have been described.[19] Receptor expression was assessed by flow cytometry.[13]

HCV Pseudoparticle and Cell Culture-Derived HCV Production, Infection, and Neutralization

Lentiviral HCVpp bearing patient-derived envelope glycoproteins were produced as described.[8,10,20] The amount of HCVpp was normalized after quantification of human immunodeficiency virus p24 antigen expression (Innotest Human Immunodeficiency Virus Antigen mAb Kit; Innogenetics) and HCVpp entry was performed as described.[8,10,11,16] Chimeric HCVcc expressing patient-derived structural proteins were constructed and produced as described in the Supplementary Materials and Methods section. HCVcc infectivity was measured by determining the tissue culture infectious dose 50% ($TCID_{50}$)[21] or intracellular HCVRNA levels as described.[13,21,22] HCVpp and HCVcc neutralization were performed as described.[8,10,11,16]

Kinetic Assays

HCVpp kinetic assays were performed in Huh7.5.1 cells using anti-CD81 (JS-81) and anti-E2 (CBH-23) monoclonal antibodies (mAbs) as described.[16,23]

Statistical Analysis

Statistical analysis (repeated-measures analysis of variance) was performed using SPSS 16.0 software for Windows (SPSS, Inc, Chicago, IL).

Results

HCV E2 Residues at Positions 447, 458, and 478 Confer Enhanced Viral Entry of a High-Infectivity Variant Reinfecting the Liver Graft

To investigate the molecular mechanism of enhanced entry of the variant VL reinfecting the liver graft, we first introduced individual mutations of region $E2_{425-483}$[8] of the low-entry and neutralization-sensitive mutant VC into HCVpp expressing envelope glycoproteins of the highly infectious escape variant VL (Figure 1*A*). Previous studies indicated that this region most likely contains the mutations responsible for the high-infectivity phenotype of VL.[8] After normalization of HCVpp levels by p24 antigen expression, viral entry was quantified relative to the escape variant VL. The entry level of the nonselected variant VC was 5% compared with the escape variant VL (Figure 1*B*). By introducing the mutations S458G and R478C into VC, chimeric HCVpp showed similar viral entry level as the paternal variant VL whereas introduction of individual or a combination of other mutations only had a partial effect (Figure 1*B*, Supplementary Figure 1). To explore the impact of other positions on viral entry we introduced mutations from another nonselected variant termed VA into VL (Figure 1*A*) and identified position F447 as an additional residue relevant for enhanced entry of the escape variant VL (Figure 1*C*). These results show that residues F447L, S458G, and R478C are largely responsible for the high infectivity of the escape variant VL.

Enhanced Viral Entry by Mutations F447L, S458G, and R478C of the Escape Variant Is the Result of Altered Use of CD81

To address whether the mutations affect viral entry by different use of cell entry factors SR-BI, CD81, CLDN1, and OCLN, we studied viral entry of HCVpp derived from parental and chimeric variants in Huh7.5.1 cells stably overexpressing the 4 main entry factors individually (Figure 2*A*). Overexpression of either SR-BI, CD81, CLDN1, or OCLN did not affect the stability or proportion of other cell-surface HCV receptors (Figure 2*B* and data not shown).

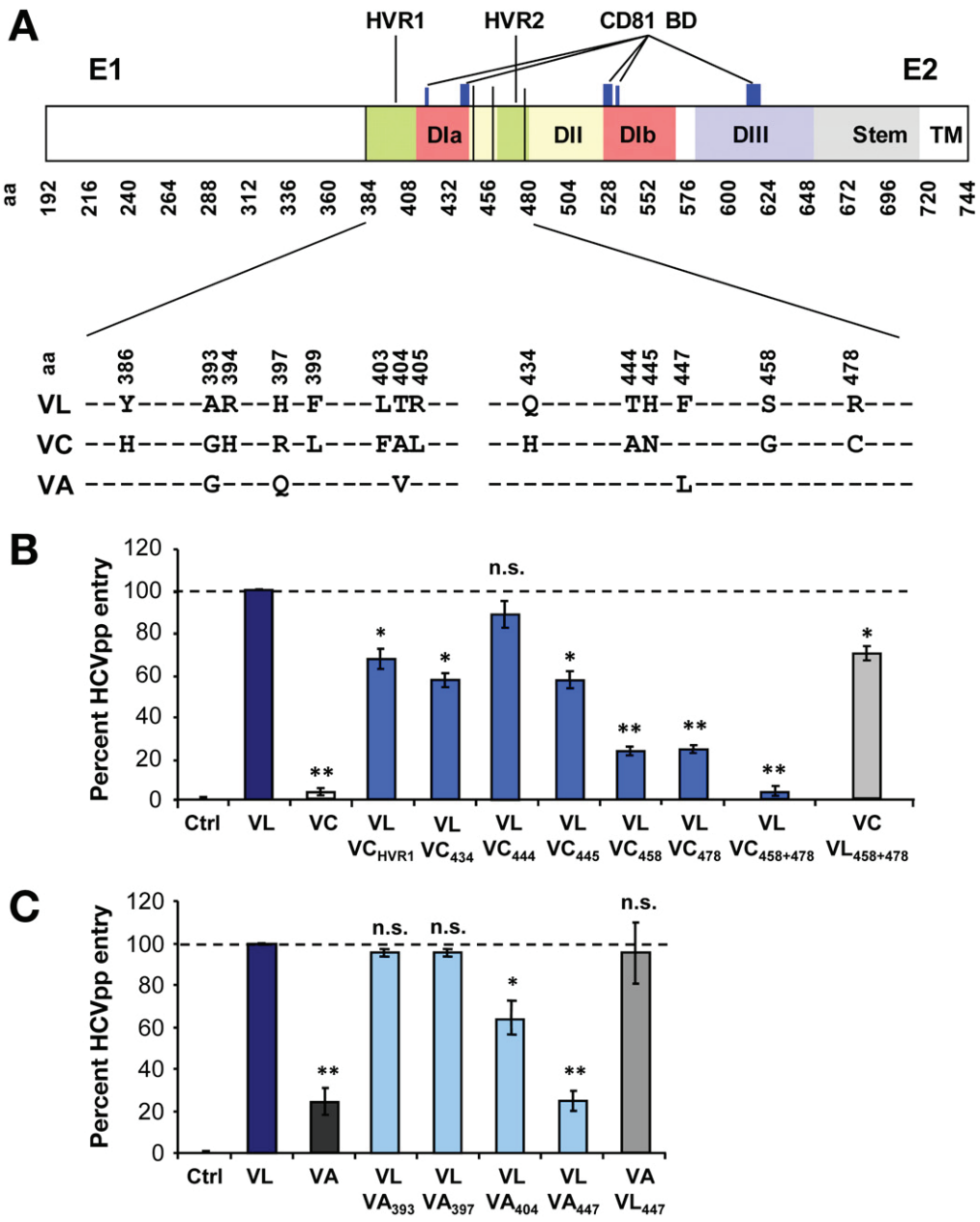


Figure 1. Positions 447, 458, and 478 confer enhanced viral entry of a high-infectivity variant reinfecting the liver graft. (*A*) Genomic organization and mutations of envelope glycoproteins of escape variant VL and nonselected variants VC and VA. HVR1 and HVR2 are depicted in *green*; E2 domains are depicted in *red* (DI), *yellow* (DII), and *blue* (DIII); and CD81 binding domains are depicted in *dark blue*.[29,33,39] Positions 447, 458, and 478 are highlighted in *black vertical lines*. Differences between VL, VC, and VA in region $E1E2_{384-483}$ are displayed. (*B* and *C*) Viral entry in Huh7.5.1 cells of the escape variant VL, the nonselected variants VC and VA, as well as chimeric variants containing defined mutations of VC and VA in VL or vice versa (Supplementary Figure 1). HCVpp infection was analyzed by luciferase reporter gene expression. Results are expressed as the percentage of viral entry compared with VL. Means ± standard deviation from at least 4 independent experiments performed in triplicate are shown. Significant differences in HCVpp entry between variants are indicated (*$P \leq .05$; **$P < .001$). aa, amino acid; BD, binding domain.

Overexpression of CD81 significantly enhanced viral entry of VL (3.2-fold) and VC (2-fold) compared with parental cells ($P < .001$) (Figure 2*C*). The fold-change in HCVpp entry was significantly higher for VL than for VC ($P < .001$). Exchanging the 2 residues at positions 458 and 478 similarly increased viral entry. This suggests that the combination of the 2 individual mutations modulates viral entry by altering CD81 dependency. Overexpression of SR-BI also increased viral entry of VL and VC, but no specific increase was observed for the chimeric strains containing substitutions at positions 458 and 478 (Figure 2*C*). These data confirm an important role for SR-BI as an entry factor for patient-derived variants, but also show that positions 458 and 478 do not significantly alter SR-BI dependency. Thus, increased entry efficiency of VL in SR-BI–overexpressing cells most likely is caused by other mutations (eg, in hypervariable region 1 [HVR1]). Viral entry enhancement was less pronounced in cells overexpressing CLDN1 or OCLN than CD81 and SR-BI (Figure 2*C*), and no specific modulation of viral entry was associated with the 2 variants or chimeric strains.

The CD81 use of viral variants VL, VC, and VA was investigated further using Huh7.5 cells with silenced CD81 expression (Figure 3*A*).[19] The escape variant VL showed the highest decrease (5.4-fold) of viral entry in shCD81-Huh7.5 cells compared with the decrease of variants VC (4.3-fold; $P < .001$) and VA (2.9-fold; $P < .001$) (Figure 3*B* and *C*). Exchange of the mapped residues into chimeric expression plasmids conferred the phenotype of decreased entry of VL (Figure 3*B* and *C*), confirming that identified residues modulate viral entry by different CD81 use. Moreover, using a relevant model system for HCV-CD81 interactions occurring in vivo consisting of cell surface–expressed CD81, we show that E1E2 complexes of

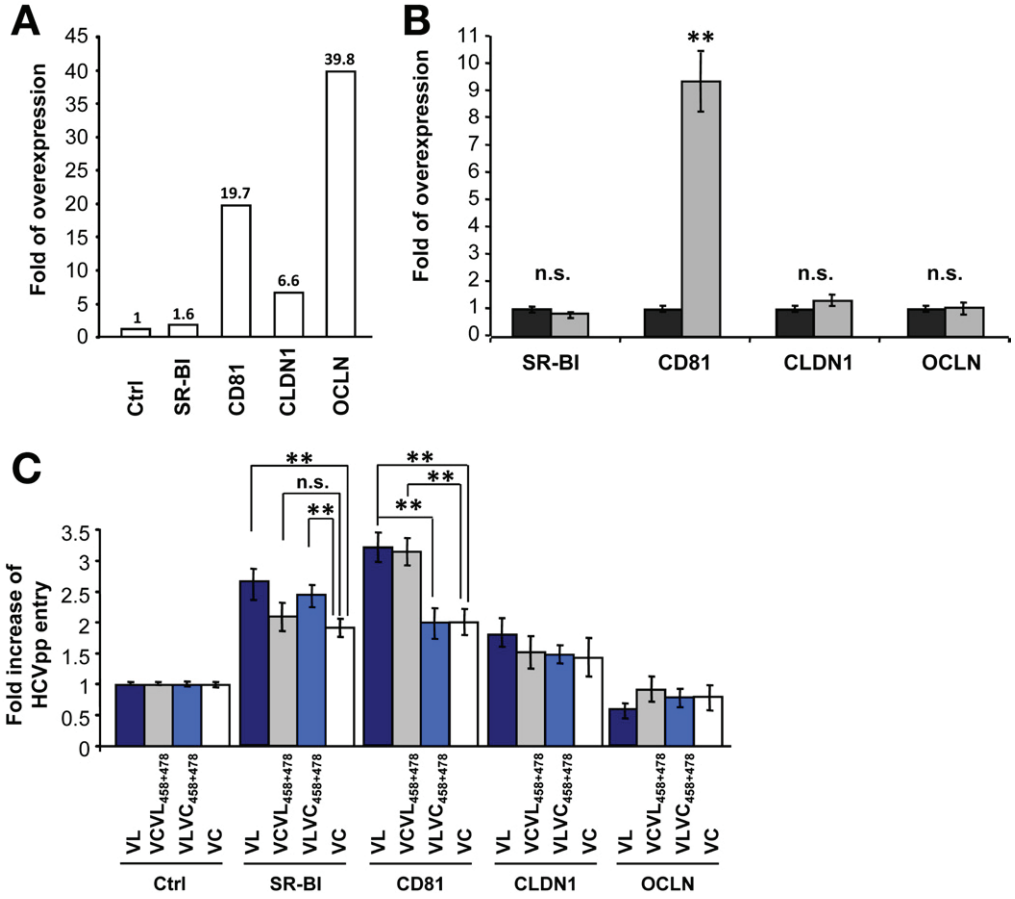


Figure 2. Altered use of CD81 is responsible for enhanced viral entry of the escape variant. (*A*) Entry factor expression in clones of SR-BI-, CD81-, CLDN1-, or OCLN-transduced Huh7.5.1 cells. The relative overexpression of each entry factor was determined by flow cytometry and is indicated as fold expression compared with parental Huh7.5.1 cells. (*B*) Entry factor expression in pools of CD81-overexpressing Huh7.5.1 cells (*grey bars*). The relative entry factor expression was determined as described in panel *A*. (*C*) Receptor dependency of patient-derived HCVpp entry. Parental and transduced Huh7.5.1 cells were incubated with parental or chimeric HCVpp and viral entry was determined as described in Figure 1. Viral entry is expressed as the fold-change of viral entry compared with parental cells. Means ± standard deviation from 3 independent experiments performed in triplicate are shown. Significant differences in HCVpp entry between variants are indicated (**$P < .001$).

the escape variant VL bound less efficiently to shCD81-Huh7.5 cells than glycoproteins of variants VC and VA (Supplementary Figure 2*A*). Exchange of the mapped residues conferred similar phenotypes as the parental glycoproteins (Supplementary Figure 2*B*), suggesting that the residues at positions 447, 458, and 478 alter E1E2 interactions with cell surface CD81.

Taken together, these data show the following: (1) the escape variant is characterized by markedly altered CD81 use, and (2) altered CD81 use of the variant is mediated by residues at positions 447, 458, and 478.

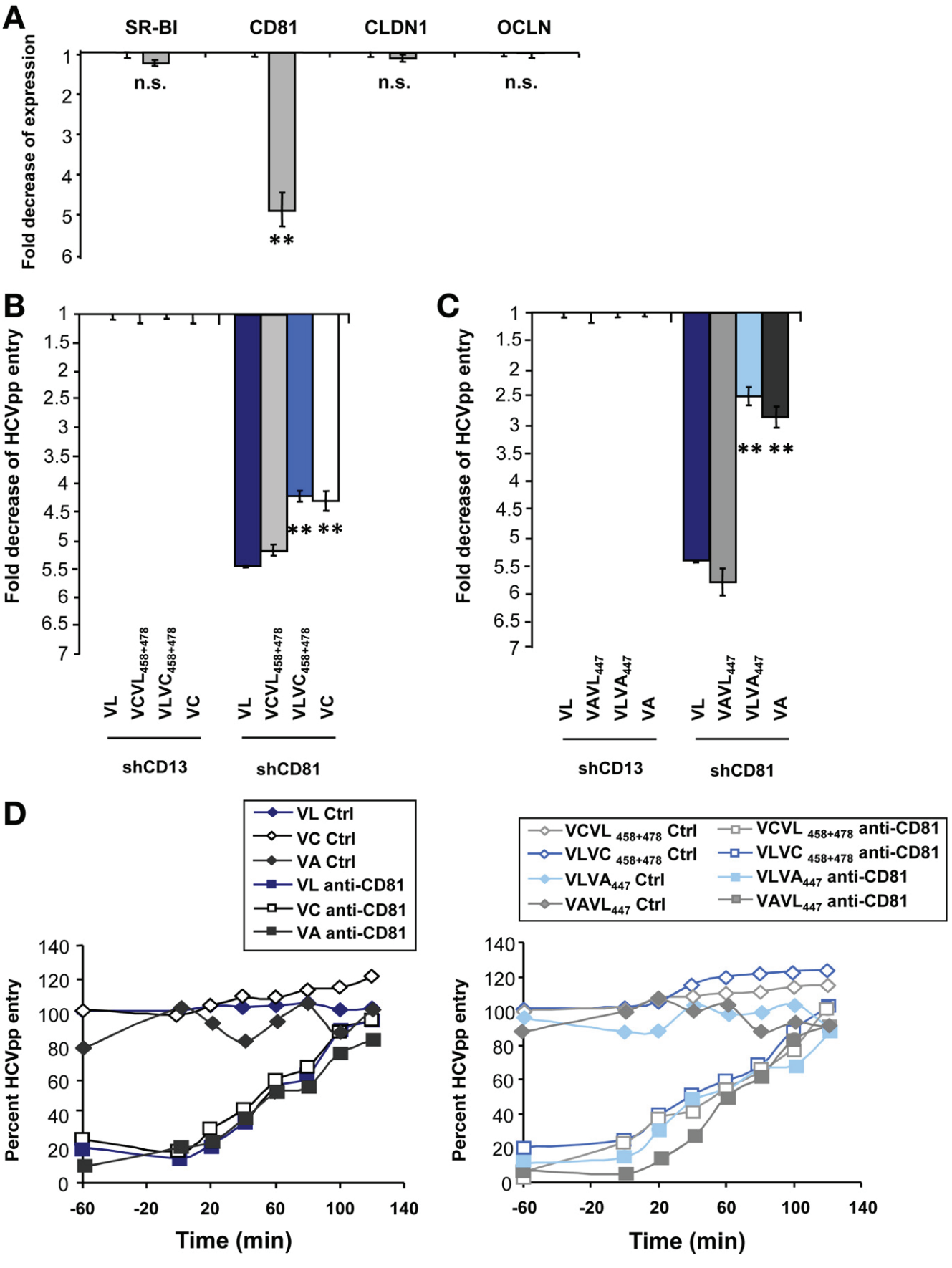
Because the levels of E1E2 incorporation into HCVpp and lentiviral p24 antigen expression were similar for all strains (Supplementary Figure 3*A–D*), it is unlikely that the differences in viral entry are the result of impaired HCVpp assembly or release.

Next, to assess whether enhanced entry is owing to more rapid internalization of viral particles, we investigated internalization kinetics of the parental and chimeric variants in the presence of anti-CD81 antibody.[16,21,23,24] Because entry kinetics of parental and chimeric variants were similar (Figure 3*D*), it is unlikely that the mutant-induced modulation of CD81 dependency alters the velocity of viral entry.

Positions 447, 458, and 478 Mediate Escape From Autologous Transplant Serum During Graft Reinfection

To assess whether the residues in region $E2_{425-483}$ influencing viral entry (Figure 1) also were responsible for escape from antibody-mediated neutralization, we studied the impact of each single and combined substitution of the nonselected variant VC on neutralization by autologous pretransplant serum. Autologous pretransplant serum only poorly neutralized the selected variant VL as well as the variants substituted at positions 434, 444, and 445, whereas individual substitution at positions 458 and 478 signifi-

Figure 3. Different CD81 use of viral variants in Huh7.5 cells with silenced CD81 expression. (*A*) Entry factor expression in Huh7.5 cells with silenced CD81 (*grey bars*) or CD13 (*black bars*) expression. CD81 expression was determined by flow cytometry and is indicated as fold expression compared with control shCD13-Huh7.5 cells. (*B* and *C*) Entry of patient-derived HCVpp VL, VC, and VA. Huh7.5 cells with silenced CD81 or CD13 expression were incubated with parental or chimeric HCVpp and viral entry was determined as described in Figure 1. Viral entry is expressed as the fold-change of viral entry compared with shCD13-Huh7.5 control cells. Means ± standard deviation from 3 independent experiments performed in triplicate are shown. Significant differences in HCVpp entry between wild-type and chimeric variants are indicated (**$P < .001$). (*D*) Entry kinetics of patient-derived variants. Kinetics of HCVpp entry was performed using anti-CD81 or isotype control antibody (5 μg/mL). HCV entry was determined as described in Figure 1. One representative experiment of 4 is shown.


A
SR-BI
CD81
CLDN1
OCLN
Fold decrease of expression
n.s.
n.s.
n.s.
**
B
Fold decrease of HCVpp entry
** **
VL
VCVL458+478
VLVC458+478
VC
shCD13
shCD81
C
Fold decrease of HCVpp entry
** **
VL
VAVL447
VLVA447
VA
shCD13
shCD81
D
VL Ctrl
VC Ctrl
VA Ctrl
VL anti-CD81
VC anti-CD81
VA anti-CD81
Percent HCVpp entry
Time (min)
VCVL 458+478 Ctrl
VLVC 458+478 Ctrl
VLVA447 Ctrl
VAVL447 Ctrl
VCVL 458+478 anti-CD81
VLVC 458+478 anti-CD81
VLVA447 anti-CD81
VAVL447 anti-CD81
Percent HCVpp entry
Time (min)

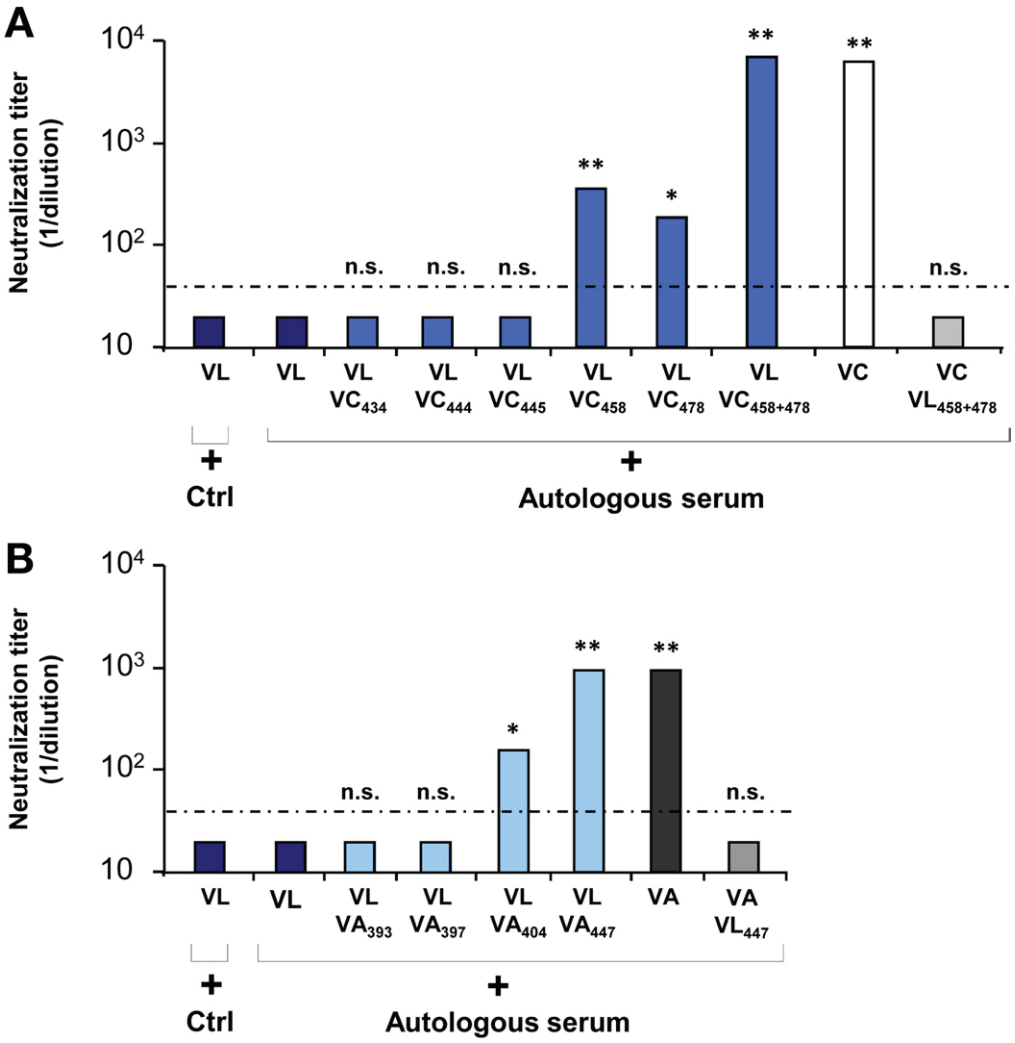


Figure 4. Positions 447, 458, and 478 mediate viral escape from neutralization by autologous transplant serum. Neutralization of the escape variant VL, variants VC and VA, and the chimeric strains. HCVpp were incubated with autologous anti-HCV–positive or control serum in serial dilutions for 1 hour at 37°C before incubation with Huh7.5.1 cells. Neutralization titers obtained by end point dilution are indicated. *Dotted line* indicates the threshold for a positive neutralization titer (1/40). Means ± standard deviation from at least 4 experiments performed in triplicate are shown. (*A*) Neutralization of variants VL, VL containing individual or combined mutations of VC, and VC with double substitutions of VL by autologous anti–HCV-positive pretransplant serum. (*B*) Neutralization of variants VL, VL containing individual mutations of VA, and VA with single substitution of VL by autologous anti-HCV–positive pretransplant serum. Significant differences in neutralization between variants are indicated (*$P \leq .05$; **$P < .001$).

cantly ($P < .001$ and $P \leq .05$, respectively) increased the sensitivity of $VLVC_{458}$ and $VLVC_{478}$ to autologous neutralizing antibodies (1:400 and 1:200, respectively) (Figure 4*A*). It is noteworthy that only the variant $VLVC_{458+478}$ showed a similar neutralization titer as the nonselected variant VC (1:6400; $P < .001$). To confirm that these mutations were indeed responsible for the phenotype of the parental variant VL, we investigated neutralization of $VCVL_{458+478}$ by autologous serum. The variant $VCVL_{458+478}$ escaped autologous neutralization similarly to the escape variant VL (Figure 4*A*). A similar phenotype was observed when mutation 447 of VA was introduced into the VL complementary DNA (Figure 4*B*). In contrast, the introduction of other residues into VL only had a minor effect on neutralization (Figure 4*B*). Taken together, these findings suggest that the residues at positions 447, 458, and 478 simultaneously are responsible for both enhanced viral entry and evasion from antibody-mediated neutralization.

Positions 447, 458, and 478 Define a Conformational Epitope Involved in Evasion From Host-Neutralizing Responses

To further elucidate the mechanism of viral evasion of the escape variant VL from patient-derived neutralizing antibodies, we investigated whether the identified mutations F447L, S458G, and R478C confer resistance or sensitivity to a panel of mAbs directed against conformational[9,17] and linear E2 epitopes.[16] The conformational HMAbs (CBH-2, CBH-5, CBH-23, and HC-1) have shown a broad cross-neutralizing activity by interfering with E2-CD81 interaction[9,17] and their epitopes only partially are defined (Supplementary Table 1). AP33 is directed against a conserved epitope comprising amino acids 412–423.[25] Although the escape variant VL was neutralized poorly by several HMAbs directed against conformational epitopes, VC and VA were neutralized efficiently by all HMAbs (Figure 5*A* and *B*). Moreover, by substituting the residues at positions 458 and 478 or 447, the well-neutralized nonselected variants VC ($VCVL_{458+478}$) and VA ($VAVL_{447}$) became neutralization resistant as the escape variant VL. Introducing the residues of VC or VA into VL ($VLVC_{458+478}$ and $VLVA_{447}$) restored neutralization by HMAbs, suggesting that these residues are part of the HMAbs epitopes. In contrast, anti-E2 antibodies (AP33, 16A6, IGH461) targeting linear epitopes similarly neutralized parental and chimeric variants (Figure 5*A* and *B* and Supplementary Table 1).

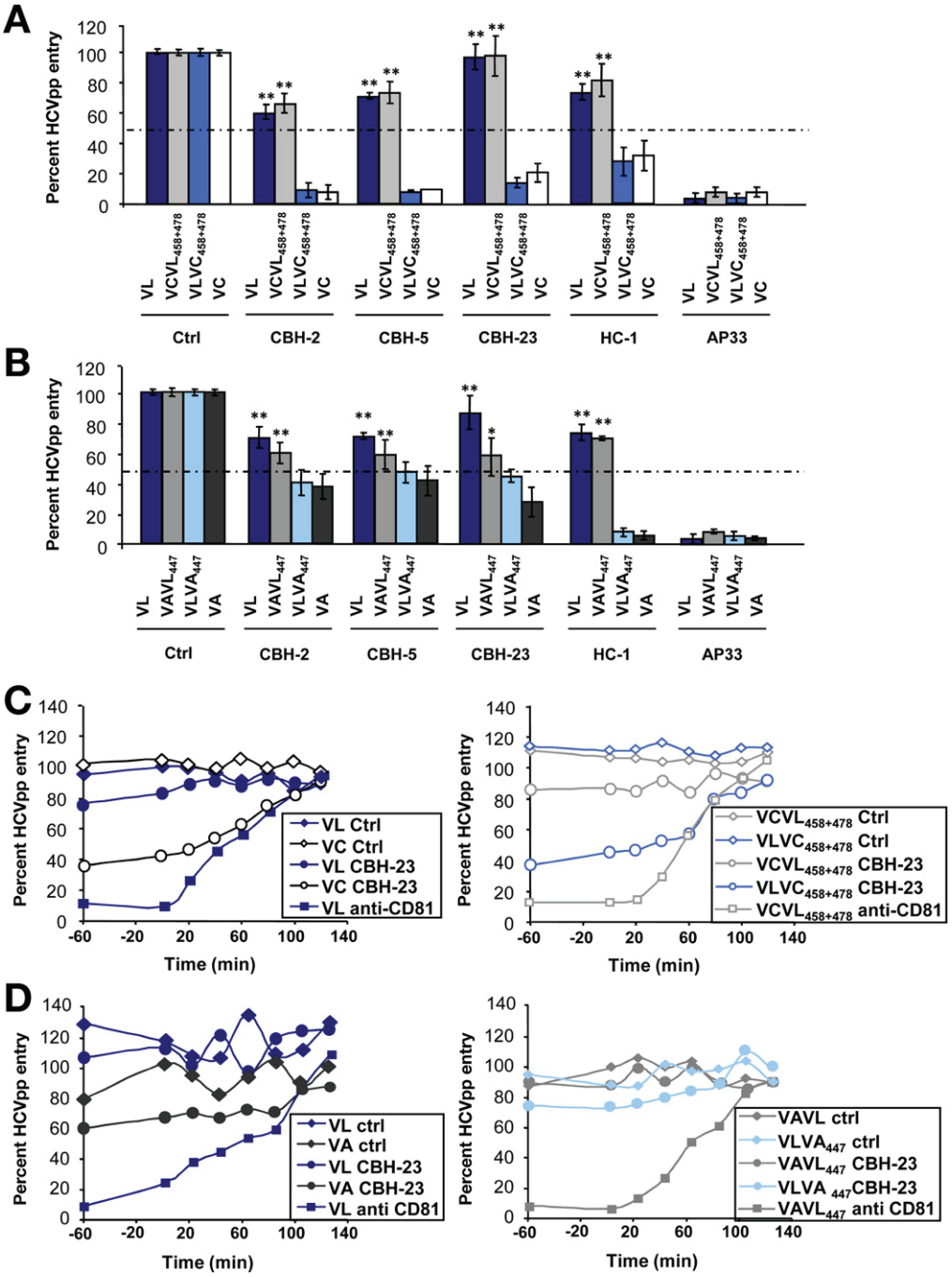


Figure 5. Mechanisms of viral evasion from neutralizing antibodies. (*A* and *B*) Escape from neutralization by HMAbs directed against conformational and linear epitopes. HCVpp produced from isolates shown in Figure 1 were incubated with HMAbs (Supplementary Table 1) or control Ab (10 μg/mL) for 1 hour at 37°C before incubation with Huh7.5.1 cells. Results are expressed as the percentage of viral entry relative to HCVpp incubated with control mAb. Means ± standard deviation from at least 4 experiments performed in triplicate are shown. Significant differences in HCVpp entry between variants are indicated (**$P < .001$). (*C* and *D*) Escape from neutralization of anti-E2 antibody CBH-23 in kinetic assays. Kinetics were performed as described in Figure 3 (HMAb, 10 μg/mL; JS-81, 5 μg/mL). One representative experiment of 4 is shown.

Antibody-mediated neutralization occurs at binding and postbinding steps during viral entry.[16] To map the entry step involved in viral evasion from neutralizing antibodies by VL, we investigated the neutralization kinetics of parental and chimeric variants.[16,21,23] The anti-E2 HMAb CBH-23 inhibited viral entry of VC and $VLVC_{458+478}$ at postbinding steps during time points closely related to HCV-CD81 interaction (Figure 5*C*). Partial inhibition at postbinding steps by CBH-23 also was observed for VA and $VLVA_{447}$ (Figure 5*D*). The VL variant escaped antibody-mediated neutralization at the same steps.

Interestingly, purified HCVpp expressing envelope glycoproteins of the escape variant bound similarly to neutralizing anti-E2 antibody CBH-23 as the envelope glycoproteins of nonselected variants or variants containing mutations of the identified escape residue (Supplementary Figure 4). Thus, it is likely that viral evasion is not caused by decreased antibody binding to circulating virions but rather occurs during postbinding steps of viral entry in which E2-host entry factor interactions result in conformational changes of the envelope and failure of antibodies to inhibit entry. Taken together, these data indicate that positions 447, 458, and 478 mediate viral evasion from neutralizing antibodies at postbinding steps and time points closely related to HCV-CD81 interaction.

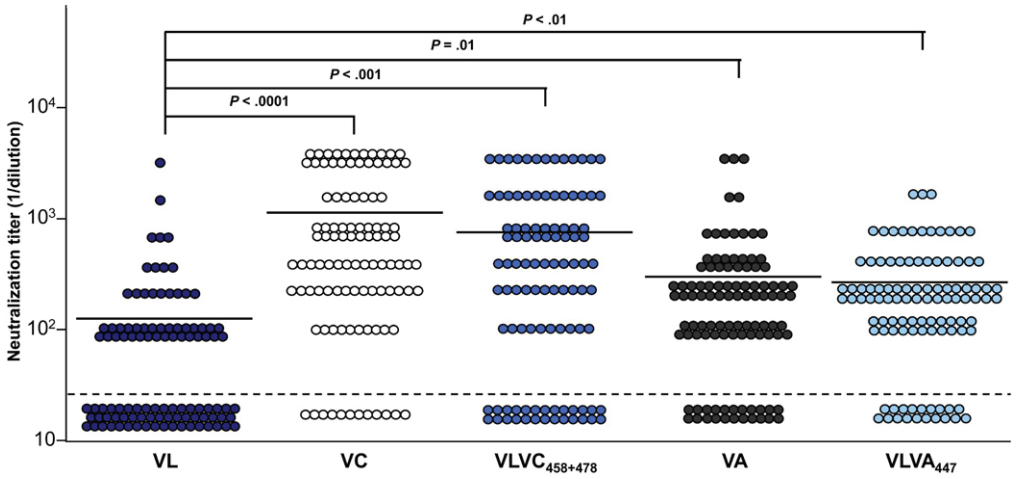


Figure 6. The HCV VL strain is poorly neutralized by antibodies present in sera from a large panel of nonrelated patients with chronic HCV infection. Parental HCVpp (VL, VC, and VA) and chimeric HCVpp ($VLVC_{458+478}$ and $VLVA_{447}$) strains, adjusted for p24 antigen expression, were preincubated for 1 hour with serial dilutions of anti-HCV–positive sera from randomly selected patients with chronic hepatitis C before incubation with Huh7.5.1 target cells. Patient number, sex, HCV genotype, and viral load are indicated in Supplementary Tables 2 and 3. Neutralization was determined as in Figure 4. Mean neutralization titers are marked by lines. Means from at least 3 independent experiments performed in triplicate are shown. Significant differences in neutralization are indicated.

Positions 447, 458, and 478 Mediate Escape From Antiviral Antibodies in Nonrelated Patients With Chronic HCV Infection

To investigate whether these mutations result not only in escape from antibodies from the same patient but also confer resistance to antiviral antibodies of nonrelated HCV-infected patients, we studied the neutralization of the parental variants by a large panel of sera randomly selected from chronically infected patients (n = 102). While VL was not neutralized by 53 of 102 patient sera (mean neutralizing titer, 1:144), VC was neutralized significantly by 90 of 102 patient sera (mean neutralizing titer, 1:1088; $P < .001$) (Figure 6 and Supplementary Tables 2 and 3). Similar results were obtained for VA (neutralization by 80 of 102 patient sera; mean neutralizing titer, 1:322; $P = .01$). Functional analysis of HCVpp expressing chimeric envelope glycoproteins showed that neutralization of VC and VA was mediated predominantly by the identified mutations in residues 447, 458, and 478 (Figure 6).

Confirmation of Differential Cell Entry Factor Use and Viral Evasion Using Chimeric HCVcc

Finally, we confirmed the functional impact of the 3 residues on virus–host interactions using the HCVcc system. To address this issue we constructed chimeric JFH-1–based HCVcc expressing the VL wild-type envelope or VL-containing VC- and VA-specific functional residues. Viruses containing patient-derived envelopes showed similar levels of replication and envelope production (data not shown). Phenotypic analyses of infection and neutralization of chimeric HCVcc confirmed the relevance of the identified residues for enhanced entry, differential CD81 use, and viral evasion (Figure 7). While the escape variant VL was poorly neutralized, the identified mutations at positions 447, 458, and 478 restored its sensitivity to conformational HMAb CBH-23 (Figure 7*C*) as well as to heterologous sera from chronically infected patients (Figure 7*D*). These data confirm the functional relevance of the obtained results in the HCVcc system expressing authentic patient-derived envelopes.

Discussion

By using acute infection of the liver graft as an in vivo model, we identified a novel clinically and therapeutically important mechanism of viral evasion, where coevolution simultaneously occurs between cellular entry factor use and escape from neutralization.

Several host selection forces operate concomitantly during HCV infection. These include proviral host factors resulting in selection of most infectious viruses best adapted to host factors and antiviral host immune responses leading to escape from immune responses. Antibody-mediated selective pressure is thought to be an important driver of viral evolution.[8,11] The immune response may fail to resolve HCV infection because neutralizing antibody-mediated response lags behind the rapidly and continuously evolving HCV glycoprotein sequences.[11] However, continuous generation of escape mutations during chronic HCV infection also may compromise virus infectivity: indeed, it has been reported that structural changes in E2 leading to complete escape from neutralizing antibodies simultaneously compromised viral fitness by reducing CD81 binding.[9] Moreover, escape from T-cell responses has been associated with impaired viral replication.[26,27] We show that clinically occurring mutations simultaneously lead to enhanced viral infectivity by optimizing host factor use and escape from host immune responses. Because this mechanism was uncovered in patient strains isolated during acute liver graft infection it is likely that the novel and unique mecha-

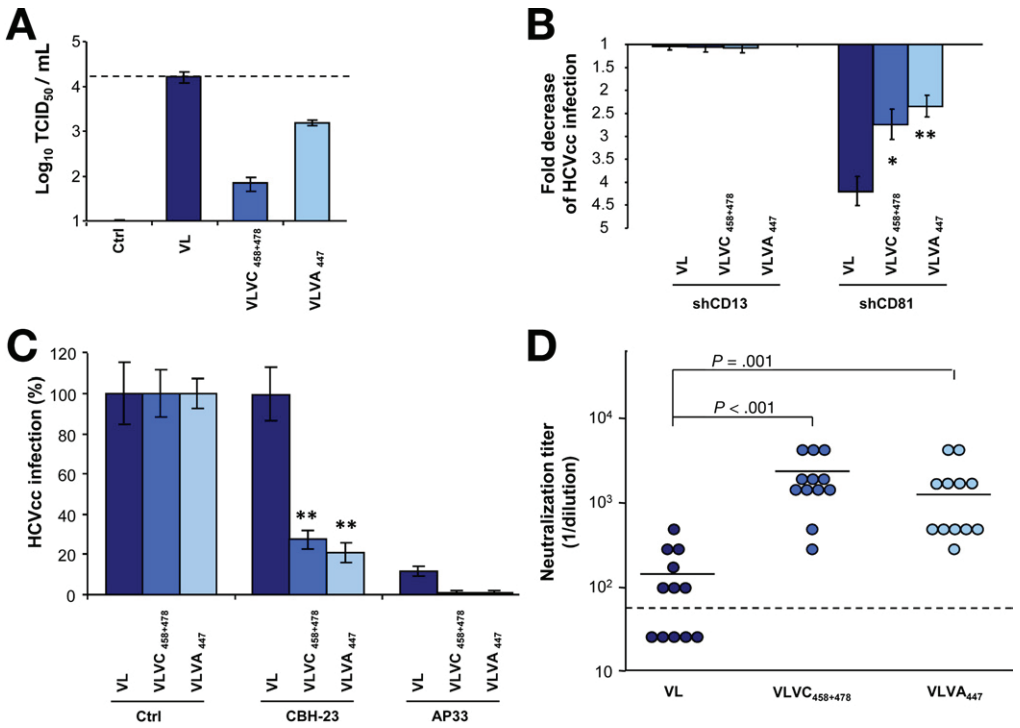


Figure 7. Viral entry and escape from neutralization of chimeric HCVcc expressing patient-derived viral envelopes. (*A*) Infectivity of HCVcc expressing envelopes of variant VL and functional residues of VA and VC is indicated by $TCID_{50}$. Means ± standard deviation from 1 representative experiment are shown. (*B*) Relative infectivity of chimeric HCVcc expressing patient-derived viral envelopes in Huh7.5 cells with silenced CD81 or CD13 expression. Means ± standard deviation from 3 independent experiments performed in triplicate are shown. (*C*) Escape from neutralization by HMAb CBH-23. Neutralization was performed as described in Figure 5. Results are expressed as the percentage of viral infectivity relative to HCVcc incubated with control mAb. Means ± standard deviation (SD) from at least 3 experiments performed in triplicate are shown. (*D*) Inhibition of HCVcc infection by anti-HCV–positive sera described in Supplementary Table 3. Neutralization was performed as described in Figure 6. Means from 1 representative experiment performed in triplicate are shown. Significant differences in HCVcc infection between wild-type and chimeric variants are indicated (*$P \leq .05$; **$P < .001$).

nism of co-evolution between host factor use and viral evasion ensures optimal initiation, dissemination, and maintenance of viral infection in the early phase of liver graft infection. In addition, because the VL strain escapes autologous antibodies from the transplant patient (Fig. 4) and resists monoclonal and polyclonal antibodies of heterologous patients (Figures 5, 6, and 7, and Supplementary Tables 1 and 2), and given the high prevalence of the identified mutations in a large genomic database of viral isolates (Supplementary Figure 5 and Supplementary Results section), the co-evolution of receptor use and escape from neutralizing antibodies also may play an important role for viral evasion in chronic HCV infection in general.

Our mechanistic studies show that the identified viral evasion factors are part of a conformational neutralizing epitope modulating E2-CD81 interactions at postbinding entry steps.[28,29] It is noteworthy that the same mutations also were responsible for immune escape of VL. Neutralization studies using HMAbs directed against discontinuous envelope glycoprotein regions termed *domain B* and *domain C*[30,31] show that the 3 positions are part of an epitope that plays a key role for neutralization and viral evasion. Because the mutations are outside the known contact residues within the epitopes of the HMAbs CBH-2, CBH-5, CBH-23, and HC-1[9,17] (Supplementary Table 1), and complementary to previously identified regions associated with escape from neutralizing monoclonal antibodies,[25] positions 447, 458, and 478 either modulate the interaction of the majority of antibodies directed against domain B and C epitopes or are part of a novel E2 epitope mediating evasion from host neutralizing antibodies.

Based on previous functional observations and structural predictions, Krey et al[29] proposed a model for a potential tertiary organization of E2. In this model, E2 comprises 3 subdomains with the CD81 binding regions located within domain I (W420, A440LFY, Y527, W529, G530, and D535) and potential CD81 binding sites overlapping with domain III (Y613RLWHY).[28,29,32,33] In this model, positions 447, 458, and 478 are located outside but in close proximity of the previously suggested CD81 binding domains. Moreover, position 447 is located immediately downstream of a conserved motif between HVR1 and HVR2, which has been shown to play an important role in CD81 recognition as well as pre- or post-CD81–dependent stages of viral entry.[32] Position 478 is located within HVR2, which modulates, by a complex interplay with HVR1, binding of E2 glycoprotein to CD81.[34]

Because mutations F447L, S458G, and R478C (1) modulate CD81 dependency of HCV entry (Figures 2 and 3), (2) alter the interaction with cell surface CD81 (Supplementary 2), (3) mediate viral evasion from antibodies at postbinding steps closely related to HCV-CD81 interactions (Figure 5), and (4) are located within E2 loops of the predicted E2 secondary structure and tertiary organization,[29] positions 447, 458, and 478 may be part of 2 loops belonging to a larger cluster of closely related surface-exposed E2 loops. These loops most likely are involved in E2-CD81 binding either directly or indirectly as a key point for structural rearrangement during viral entry.[34,35]

The polar S and R residues present in the escape variant can form nonbonded interactions with other residues by hydrogen bonds and salt bridge, respectively. These interactions could increase the stability of the interacting E2-CD81 interface, allowing efficient entry of the VL escape variant through E2-CD81-CLDN1 co-receptor complexes, which are key determinants for viral entry.[13,23,36] Furthermore, the E2 cluster of loops containing the mutations bears linear epitopes but also defines at least one conformational epitope that is a target of neutralizing antibodies. According to residue physical-chemical properties, the VL variant S458 and R478 residues enhance the hydrophilicity of the loops they belong to and may promote the surface exposure of the loops. This change could modulate E2-CD81 interactions further and impair the binding of neutralizing antibodies by blocking access to their target epitopes. The F to L substitution present in the VA strain most likely does not profoundly alter the tertiary or quaternary structure of E2. This is suggested by the fact that this position is located in a loop as predicted by the proposed E2 model.[29] Thus, it is conceivable that this mutation, which increases E2 hydrophobicity, may reduce accessibility of the loop and its interactions with CD81 or CD81-CLDN1 co-receptor complexes. Alternatively, allosteric mechanisms may play a role in the observed virus-antibody-host interactions.

Taken together, our data identified key determinants of immune evasion in vivo. Mutations conferring neutralization escape altered CD81 receptor use and enhanced cell entry. Moreover, our data suggest that mutations in HVR1, which may modulate entry and neutralization by altering SR-BI dependency (Figures 1, 2, and 4, and data not shown), may contribute to the high entry and escape phenotype of the escape variant. Furthermore, interfering non-neutralizing antibodies may constitute another mechanism of escape (data not shown).

Although proof-of-concept studies in animal models have shown a potential role for HMAbs in prevention of HCV infection,[37,38] the partial or complete escape of the VL variant from autologous and heterologous serum-derived antibodies as well as many broadly cross-neutralizing HMAbs (Figure 5; Supplementary Table 1) shows the ability of the virus to evade cross-neutralizing anti-envelope mAbs. By identifying viral and host factors mediating immune evasion in the HCV-infected patient, our results may open new perspectives for the development of broadly cross-neutralizing anti-envelope or antibodies overcoming viral escape.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at doi:http://dx.doi.org/10.1053/j.gastro.

References

1. Alter H. Viral hepatitis. Hepatology 2006;43:S230–S234.
2. Hofmann WP, Zeuzem S. A new standard of care for the treatment of chronic HCV infection. Nat Rev Gastroenterol Hepatol 2011;8:257–264.
3. Watt K, Veldt B, Charlton M. A practical guide to the management of HCV infection following liver transplantation. Am J Transplant 2009;9:1707–1713.
4. Zeisel MB, Cosset FL, Baumert TF. Host neutralizing responses and pathogenesis of hepatitis C virus infection. Hepatology 2008;48:299–307.
5. Zeisel MB, Fofana I, Fafi-Kremer S, et al. HCV entry into hepatocytes: mechanism and potential therapeutic implications. J Hepatol 2011;54:566–576.
6. Lavillette D, Morice Y, Germanidis G, et al. Human serum facilitates hepatitis C virus infection, and neutralizing responses inversely correlate with viral replication kinetics at the acute phase of hepatitis C virus infection. J Virol 2005;79:6023–6034.
7. Dowd KA, Netski DM, Wang XH, et al. Selection pressure from neutralizing antibodies drives sequence evolution during acute infection with hepatitis C virus. Gastroenterology 2009;136:2377–2386.
8. Fafi-Kremer S, Fofana I, Soulier E, et al. Enhanced viral entry and escape from antibody-mediated neutralization are key determinants for hepatitis C virus re-infection in liver transplantation. J Exp Med 2010;207:2019–2031.
9. Keck ZY, Li SH, Xia J, et al. Mutations in hepatitis C virus E2 located outside the CD81 binding sites lead to escape from broadly neutralizing antibodies but compromise virus infectivity. J Virol 2009;83:6149–6160.
10. Pestka JM, Zeisel MB, Blaser E, et al. Rapid induction of virus-neutralizing antibodies and viral clearance in a single-source outbreak of hepatitis C. Proc Natl Acad Sci U S A 2007;104:6025–6030.
11. von Hahn T, Yoon JC, Alter H, et al. Hepatitis C virus continuously escapes from neutralizing antibody and T-cell responses during chronic infection in vivo. Gastroenterology 2007;132:667–678.
12. Osburn WO, Fisher BE, Dowd KA, et al. Spontaneous control of primary hepatitis C virus infection and immunity against persistent reinfection. Gastroenterology 2010;138:315–324.
13. Lupberger J, Zeisel MB, Xiao F, et al. EGFR and EphA2 are host factors for hepatitis C virus entry and targets for antiviral therapy. Nat Med 2011;17:589–595.
14. Stamataki Z, Grove J, Balfe P, et al. Hepatitis C virus entry and neutralization. Clin Liver Dis 2008;12:693–712.
15. Kuiken C, Combet C, Bukh J, et al. A comprehensive system for consistent numbering of HCV sequences, proteins and epitopes. Hepatology 2006;44:1355–1361.
16. Haberstroh A, Schnober EK, Zeisel MB, et al. Neutralizing host responses in hepatitis C virus infection target viral entry at postbinding steps and membrane fusion. Gastroenterology 2008;135:1719–1728.
17. Hadlock KG, Lanford RE, Perkins S, et al. Human monoclonal antibodies that inhibit binding of hepatitis C virus E2 protein to CD81 and recognize conserved conformational epitopes. J Virol 2000;74:10407–10416.

18. Haid S, Windisch MP, Bartenschlager R, et al. Mouse-specific residues of claudin-1 limit hepatitis C virus genotype 2a infection in a human hepatocyte cell line. J Virol 2009;84:964–975.
19. Koutsoudakis G, Herrmann E, Kallis S, et al. The level of CD81 cell surface expression is a key determinant for productive entry of hepatitis C virus into host cells. J Virol 2007;81:588–598.
20. Bartosch B, Dubuisson J, Cosset FL. Infectious hepatitis C virus pseudo-particles containing functional E1-E2 envelope protein complexes. J Exp Med 2003;197:633–642.
21. Zeisel MB, Koutsoudakis G, Schnober EK, et al. Scavenger receptor BI is a key host factor for hepatitis C virus infection required for an entry step closely linked to CD81. Hepatology 2007;46:1722–1731.
22. Fofana I, Krieger SE, Grunert F, et al. Monoclonal anti-claudin 1 antibodies for prevention of hepatitis C virus infection. Gastroenterology 2010;139:953–964, 964.e1–4.
23. Krieger SE, Zeisel MB, Davis C, et al. Inhibition of hepatitis C virus infection by anti-claudin-1 antibodies is mediated by neutralization of E2-CD81-claudin-1 associations. Hepatology 2010;51:1144–1157.
24. Koutsoudakis G, Kaul A, Steinmann E, et al. Characterization of the early steps of hepatitis C virus infection by using luciferase reporter viruses. J Virol 2006;80:5308–5320.
25. Owsianka A, Tarr AW, Juttla VS, et al. Monoclonal antibody AP33 defines a broadly neutralizing epitope on the hepatitis C virus E2 envelope glycoprotein. J Virol 2005;79:11095–11104.
26. Dazert E, Neumann-Haefelin C, Bressanelli S, et al. Loss of viral fitness and cross-recognition by CD8+ T cells limit HCV escape from a protective HLA-B27-restricted human immune response. J Clin Invest 2009;119:376–386.
27. Uebelhoer L, Han JH, Callendret B, et al. Stable cytotoxic T cell escape mutation in hepatitis C virus is linked to maintenance of viral fitness. PLoS Pathog 2008;4:e1000143.
28. Owsianka AM, Timms JM, Tarr AW, et al. Identification of conserved residues in the E2 envelope glycoprotein of the hepatitis C virus that are critical for CD81 binding. J Virol 2006;80:8695–8704.
29. Krey T, d'Alayer J, Kikuti CM, et al. The disulfide bonds in glycoprotein E2 of hepatitis C virus reveal the tertiary organization of the molecule. PLoS Pathog 2010;6:e1000762.
30. Keck ZY, Op De Beeck A, Hadlock KG, et al. Hepatitis C virus E2 has three immunogenic domains containing conformational epitopes with distinct properties and biological functions. J Virol 2004;78:9224–9232.
31. Helle F, Goffard A, Morel V, et al. The neutralizing activity of anti-hepatitis C virus antibodies is modulated by specific glycans on the E2 envelope protein. J Virol 2007;81:8101–8111.
32. Drummer HE, Boo I, Maerz AL, et al. A conserved gly436-trp-leu-ala-gly-leu-phe-tyr motif in hepatitis C virus glycoprotein E2 is a determinant of CD81 binding and viral entry. J Virol 2006;80:7844–7853.
33. Boo I, Tewierek K, Douam F, et al. Distinct roles in folding, CD81 receptor binding and viral entry for conserved histidines of HCV glycoprotein E1 and E2. Biochem J 2012;443:85–94.
34. Roccasecca R, Ansuini H, Vitelli A, et al. Binding of the hepatitis C virus E2 glycoprotein to CD81 is strain specific and is modulated by a complex interplay between hypervariable regions 1 and 2. J Virol 2003;77:1856–1867.
35. McCaffrey K, Boo I, Poumbourios P, et al. Expression and characterization of a minimal hepatitis C virus glycoprotein E2 core domain that retains CD81 binding. J Virol 2007;81:9584–9590.
36. Harris HJ, Davis C, Mullins JG, et al. Claudin association with CD81 defines hepatitis C virus entry. J Biol Chem 2010;285:21092–21102.
37. Law M, Maruyama T, Lewis J, et al. Broadly neutralizing antibodies protect against hepatitis C virus quasispecies challenge. Nat Med 2008;14:25–27.
38. Vanwolleghem T, Bukh J, Meuleman P, et al. Polyclonal immunoglobulins from a chronic hepatitis C virus patient protect human liver-chimeric mice from infection with a homologous hepatitis C virus strain. Hepatology 2008;47:1846–1855.
39. Rychlowska M, Owsianka AM, Foung SK, et al. Comprehensive linker-scanning mutagenesis of the hepatitis C virus E1 and E2 envelope glycoproteins reveals new structure-function relationships. J Gen Virol 2011;92:2249–2261.

Received July 1, 2011. Accepted April 6, 2012.

Reprint requests

Address requests for reprints to: Thomas F. Baumert, MD, Inserm Unit 748, Hôpitaux Universitaires de Strasbourg, 3 Rue Koeberlé, F-67000 Strasbourg, France. e-mail: Thomas.Baumert@unistra.fr; fax: (33) 3-68-85-37-50.

Acknowledgments

The authors thank F. Chisari (The Scripps Research Institute, La Jolla, CA) for the gift of Huh7.5.1 cells, J. A. McKeating (University of Birmingham, Birmingham, UK), C. Rice (Rockefeller University, New York, NY), C. Schuster (Inserm U748, Strasbourg, France), M. Heim (University of Basel, Basel, Switzerland), F. Wong-Staal (Itherx, San Diego, CA), J. Dubuisson (Inserm U1019, Lille, France), and F. Rey (Institut Pasteur, Paris, France) for helpful discussions. The authors acknowledge the excellent technical assistance of Michèle Bastien-Valle (Inserm U748, Strasbourg, France).

I.F., S.F.-K., and P.C. contributed equally to this article.

Conflicts of interest

This author discloses the following: Thomas Pietschmann is a member of the advisory board of Biotest AG and has received consulting fees. The remaining authors disclose no conflicts.

Funding

This work was supported by Inserm, the European Union (ERC-2008-AdG-233130-HEPCENT and Interreg IV FEDER-Hepato-Regio-Net 2009), the Agence Nationale de la Recherche chair of excellence program (ANR-05-CEXC-008), Agence Nationale de Recherches sur le Sida et les Hépatites Virales (2007/306, 2008/354, 2009/183, 2011/132), the Région d'Alsace (2007/09), the Else Kröner-Fresenius Stiftung (EKFS P17//07//A83/06), the Ligue Contre le Cancer (CA 06/12/08), INCA (2009-143), Cancéropôle du Grand-Est (30/03/09), the Finovi Foundation, the Infrastrutures en Biologie Santé et Agronomie, the Société Française d'Exportation des Ressources Educatives program of Higher Education Commission of Pakistan, and Public Heatlh Service grants HL079381 and AI081903.

Supplementary Materials and Methods

Analysis of HCVpp Envelope Glycoprotein Expression

Expression of HCV glycoproteins was characterized in HEK 293T producer cells and HCVpp purified through a 20% sucrose cushion ultracentrifugation as described.[1] Immunoblots of HCV glycoproteins were performed using anti-E1 11B7 and anti-E2 AP33 mAbs as described.[2]

Cellular Binding of Envelope Glycoproteins

Envelope glycoprotein-expressing HEK 293T cells were lysed in phosphate-buffered saline by 4 freezing and thawing cycles. Cell debris and nuclei were removed by low-speed centrifugation and supernatants containing native intracellular E1E2 complexes were used for binding studies. Huh7.5.1, shCD81-, or shCD13-Huh7.5 cells (2×10^5 cells per well) were seeded in 96-well plates. After incubation with lysates containing patient-derived E1E2 proteins, Huh7.5.1 target cells were first incubated with mAb AP33 (10 μg/mL) and then with phycoerythrin-conjugated anti-mouse Ab (5 μg/mL, BD Biosciences). Bound E2 was analyzed by flow cytometry as described.[3]

Construction of Plasmids for Production of Chimeric HCVcc Expressing Patient-Derived Envelopes

Genotype 1 JFH-based HCVcc chimeras expressing the structural proteins of patient-derived viruses were produced as previously described for Con1/C3-JFH1-V2440L.[4,5] Briefly, the complementary DNA region encoding for the HCV core to the first transmembrane domain of NS2 (C3 junction site) from variant VL was inserted into pFK-Con1/C3-JFH1-V2440L using fusion polymerase chain reaction with Pfu DNA polymerase (Agilent Technologies, Massy, France) and standard cloning procedures using appropriate restriction sites including BsmI and AvrII. The obtained construct was designated *VL/JFH1*. The VL/JFH1 encoding sequence was used as a template to insert individual and combined mutations using the QuikChange II XL site-directed mutagenesis kit (Agilent Technologies) as described previously.[1]

Galanthus nivalis *Capture Enzyme-Linked Immunosorbent Assay*

Binding of HMAb CBH-23 to viral envelopes was analyzed using an enzyme-linked immunosorbent assay with HCVpp as a capture antigen as described.[6] HCVpp expressing the E1E2 glycoproteins of HCV variants or control pseudoparticles with absent HCV envelope glycoprotein expression were partially purified and enriched through ultracentrifugation as described.[1] Purified particles were quantified as described previously.[1] Partially purified HCVpp or control pseudoparticles were captured onto *Galanthus nivalis* (GNA)-coated microtiter plates as described.[6] Soluble E2 (derived from strain HCV-H77 and expressed in 293T cells as described previously[3]) was used as a positive control for antibody binding. Neutralizing human anti-E2 antibody CBH-23 (25 μg/mL diluted in phosphate-buffered saline) then was added to captured HCVpp or soluble E2 (1 h at room temperature). After washing and removal of nonbound antibody, mAb binding to HCV envelopes was detected using horseradish-peroxidase anti-human IgG (GE Healthcare, Orsay, France) at a concentration of 1/3000 for 1 hour at room temperature, followed by incubation with 1-step Turbo TMB–enzyme-linked immunosorbent assay (Thermo Fisher Scientific, Illkirch, France) for color development. Absorbance was measured at 450 nm using a microplate reader Softmax program (Molecular Devices, Sunnyvale, CA).

Bioinformatics

Multiple sequence alignment of complete E2 proteins was performed using the European HCV databases (http://euhcvdb.ibcp.fr).[7] Two amino acid repertoires were computed with all E2 sequences of provisional/confirmed genotype 1b using the *ComputeRepertoire* tool as part of the euHCVdb *Extract* tool (http://euhcvdb.ibcp.fr).

Results

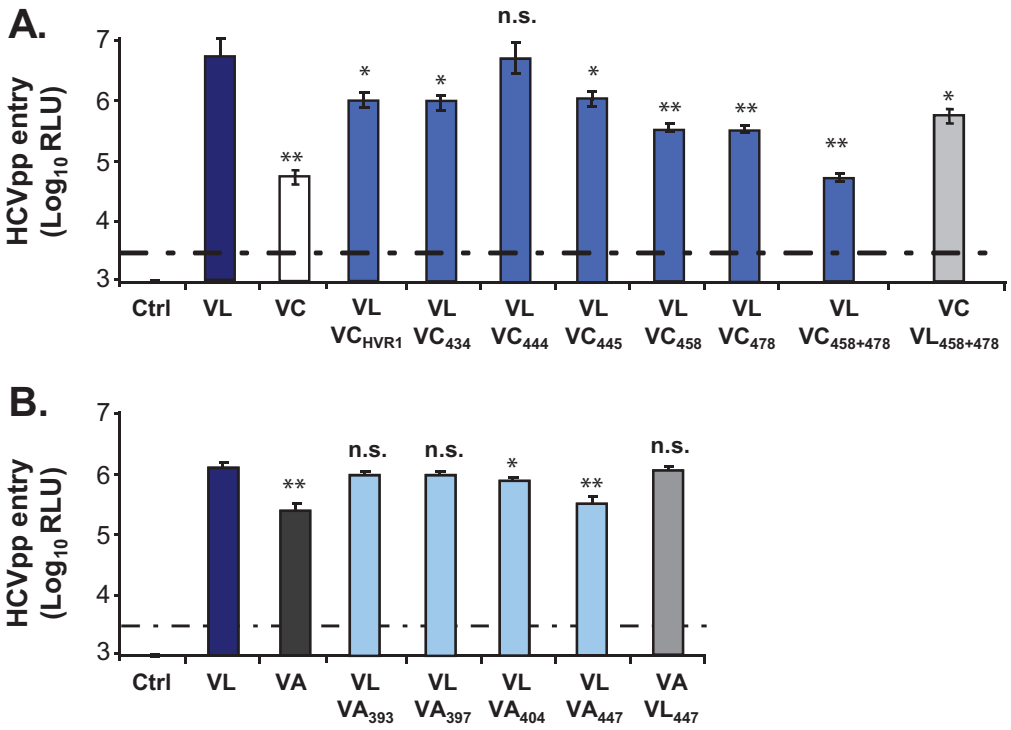
Prevalence of the Identified Mutations in a Large Genomic Database of Viral Isolates

Bioinformatic sequence analysis of a large panel of 2074 HCV strains within the European HCV database further supports the potential relevance of the identified positions for pathogenesis of HCV infection in general.[7] Residues F, S, and R are observed much more frequently at positions 447, 458, and 478 than L, G, and C. F and S are the most predominant residues at positions 447 and 458 in the large majority of 1b strains, respectively (F447 all, 98.4%; 1b, 96.2%; S458 all, 94%; 1b, 90.3%; Supplementary Figure 5). The position 478 is variable but R (all, 2.4%; 1b, 10.8%) is more frequent than C (all, 0.2%; 1b, 0.9%) (Supplementary Figure 5). The high prevalence of identified residues supports their functional relevance for virus survival and selection because more structurally and functionally relevant residues will be observed more frequently. These data suggest that the epitope containing the identified residues at positions 447, 458, and 478 is responsible not only for viral evasion from autologous antiviral antibodies during LT but also may contribute to viral evasion in chronic HCV infection in general.

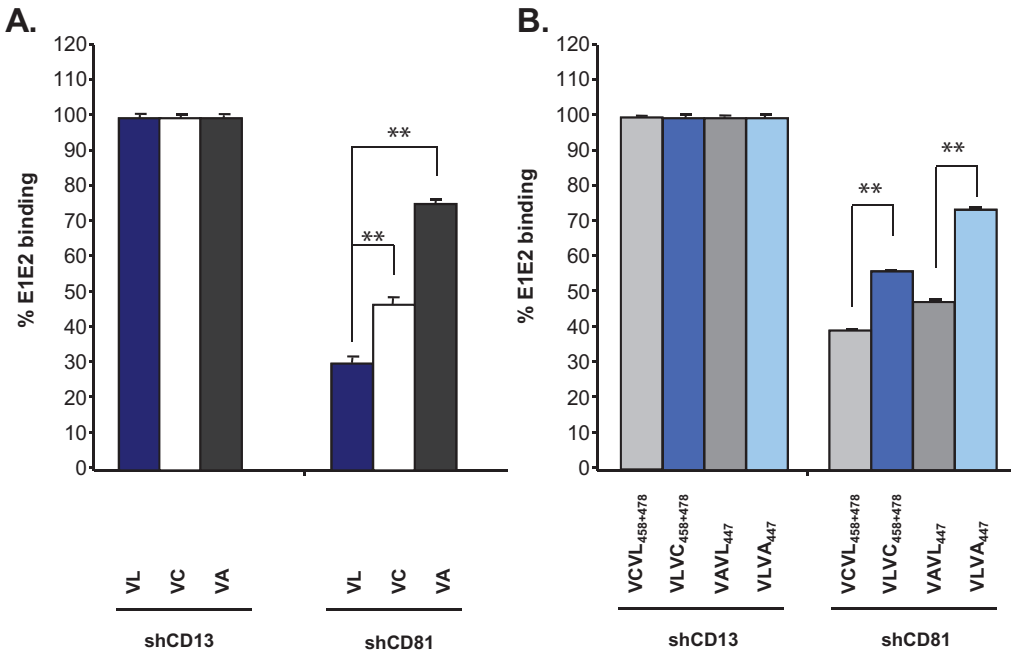
References

1. Fafi-Kremer S, Fofana I, Soulier E, et al. Enhanced viral entry and escape from antibody-mediated neutralization are key determinants for hepatitis C virus re-infection in liver transplantation. J Exp Med 2010;207:2019–2031.
2. Pestka JM, Zeisel MB, Blaser E, et al. Rapid induction of virus-neutralizing antibodies and viral clearance in a single-source outbreak of hepatitis C. Proc Natl Acad Sci U S A 2007;104:6025–6030.

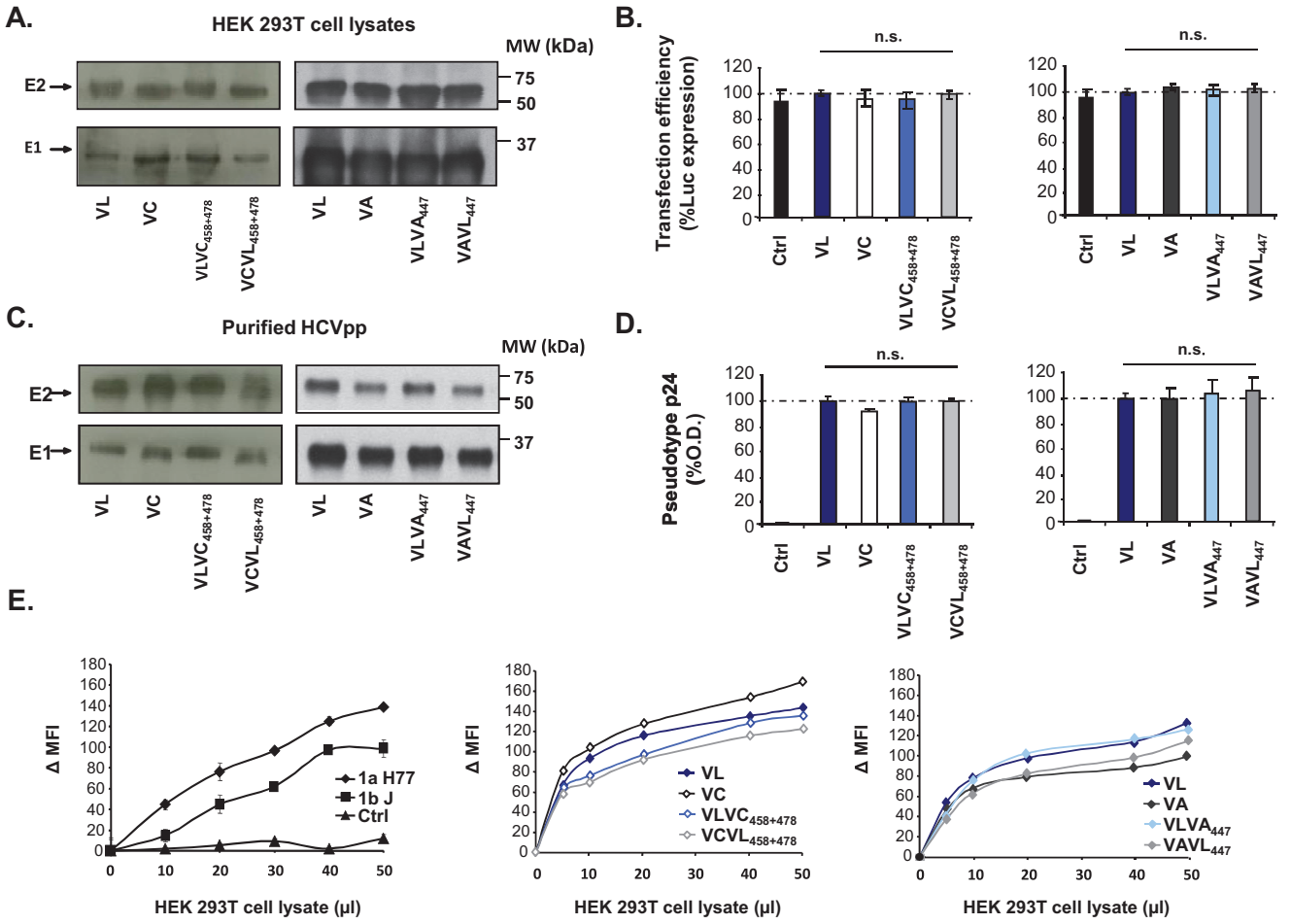
3. Krieger SE, Zeisel MB, Davis C, et al. Inhibition of hepatitis C virus infection by anti-claudin-1 antibodies is mediated by neutralization of E2-CD81-claudin-1 associations. Hepatology 2010; 51:1144–1157.
4. Pietschmann T, Kaul A, Koutsoudakis G, et al. Construction and characterization of infectious intragenotypic and intergenotypic hepatitis C virus chimeras. Proc Natl Acad Sci U S A 2006;103: 7408–7413.
5. Kaul A, Woerz I, Meuleman P, et al. Cell culture adaptation of hepatitis C virus and in vivo viability of an adapted variant. J Virol 2007;81:13168–13179.
6. Wang Y, Keck ZY, Saha A, et al. Affinity maturation to improve human monoclonal antibody neutralization potency and breadth against hepatitis C virus. J Biol Chem 2011;286:44218–44233.
7. Combet C, Garnier N, Charavay C, et al. euHCVdb: the European hepatitis C virus database. Nucleic Acids Res 2007;35:D363–D366.
8. Owsianka A, Tarr AW, Juttla VS, et al. Monoclonal antibody AP33 defines a broadly neutralizing epitope on the hepatitis C virus E2 envelope glycoprotein. J Virol 2005;79:11095–11104.
9. Haberstroh A, Schnober EK, Zeisel, et al. Neutralizing host responses in hepatitis C virus infection target viral entry at post-binding steps and membrane fusion. Gastroenterology 2008;135:1719–1728.
10. Hadlock KG, Lanford RE, Perkins S, et al. Human monoclonal antibodies that inhibit binding of hepatitis C virus E2 protein to CD81 and recognize conserved conformational epitopes. J Virol 2000;74:10407–10416.
11. Keck ZY, Li SH, Xia J, et al. Mutations in hepatitis C virus E2 located outside the CD81 binding sites lead to escape from broadly neutralizing antibodies but compromise virus infectivity. J Virol 2009;83:6149–6160.
12. Dimitrova M, Affolter C, Meyer F, et al. Sustained delivery of siRNAs targeting viral infection by cell-degradable multilayered polyelectrolyte films. Proc Natl Acad Sci U S A 2008;105:16320–16325.
13. Lupberger J, Zeisel MB, Xiao F, et al. EGFR and EphA2 are hepatitis C virus host entry factors and targets for antiviral therapy. Nat Med 2011;17:589–595.
14. Bartosch B, Vitelli A, Granier C, et al. Cell entry of hepatitis C virus requires a set of co-receptors that include the CD81 tetraspanin and the SR-B1 scavenger receptor. J Biol Chem 2003;278: 41624–41630.



Supplementary Figure 1. Actual viral infectivity of HCVpp derived from variants VL, VC, and VA shown as relative light units (RLU) of luciferase reporter gene expression. (*A* and *B*) Comparative analysis of viral entry of HCVpp shown in Figure 1. Results are expressed in RLUs plotted in a logarithmic scale. The threshold for a detectable infection in this system is indicated by *dashed lines*. The detection limit for positive luciferase reporter protein expression was 3×10^3 RLU/assay, corresponding to the mean ± 3 standard deviations of background levels (ie, luciferase activity of naive noninfected cells or cells infected with pseudotypes without HCV envelopes).[1,12,13] Background levels of the assay were determined in each experiment. Means ± standard deviation from at least 4 independent experiments performed in triplicate are shown. Significant differences in HCVpp entry VC, VA, and VL wild-type and mutant variants are indicated (*$P \leq .05$; **$P < .001$). Ctrl, control; HVR, hypervariable region; V, viral variant.



Supplementary Figure 2. Positions 447, 458, and 478 modulate binding of envelope glycoproteins to CD81 expressed at the cell surface. Binding of native E1E2 complexes expressed from patient-derived complementary DNAs to Huh7.5 cells with silenced CD81 expression (described in Figure 3) was detected by flow cytometry. Results are expressed as the percentage of E1E2 binding compared with shCD13-Huh7.5 control cells. Means ± standard deviation from 3 independent experiments performed in triplicate are shown. Significant differences in binding between variants are indicated (**$P < .001$).

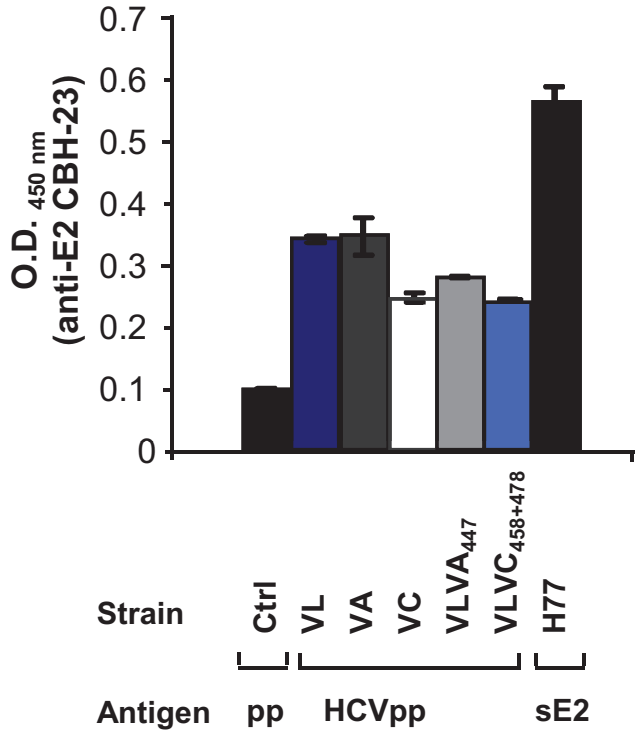


Supplementary Figure 3. Differences in viral entry are not caused by impaired HCVpp production. (*A*) Analysis of envelope glycoprotein expression. Protein expression was analyzed by immunoblotting as described in the Materials and Methods section. Molecular markers (in kilodaltons) are indicated on the *right*. (*B*) Transfection efficiency during HCVpp production. Transfection efficiency was analyzed for each variant and quantified by determining luciferase expression in HEK 293T producer cells expressed as a normalized percentage compared with control transfected cells. (*C*) Envelope glycoprotein expression in HCVpp. HCVpp were purified as described previously[1,2] and subjected to immunoblot as described in panel *A*. (*D*) Lentiviral p24 antigen expression was analyzed by enzyme-linked immunosorbent assay (ELISA) and is indicated as optical density (OD) values at 450 nm. (*E*) Cellular binding of E2 derived from patient-derived or H77 and HCV-J strains. Binding of native E1E2 complexes to Huh7.5.1 cells was detected as described in Supplementary Materials and Methods. Results are expressed as delta mean fluorescence intensity (ΔMFI) ± standard deviation. One representative experiment of 3 is shown. Da, dalton; MW, molecular weight.

Supplementary Table 1. Neutralization of Patient-Derived and Chimeric HCVpp by Monoclonal Anti-Envelope Antibodies

Antibody	Reference	Epitope, amino acid	HCVpp entry, %						
			VL	VC	$VCVL_{458+478}$	$VLVC_{458+478}$	VA	$VAVL_{447}$	$VLVA_{447}$
AP33	8	412–423	6 ± 3	12 ± 1	3 ± 1	11 ± 5	2 ± 1	5 ± 1	3 ± 1
IGH461	9	436–448	58 ± 4	56 ± 8	51 ± 7	53 ± 3	55 ± 2	56 ± 6	52 ± 7
16A6	9	523–530	76 ± 10	74 ± 8	83 ± 9	82 ± 2	73 ± 9	74 ± 4	81 ± 9
CBH-2	10	Domain B, conformational 431, 523–540	60 ± 5	8 ± 5	65 ± 6	9 ± 5	39 ± 8	61 ± 4	39 ± 10
CBH-5	10	Domain B, conformational 523–540	71 ± 2	10 ± 4	73 ± 7	8 ± 1	36 ± 5	59 ± 7	47 ± 8
CBH-23	Keck and Foung, unpublished data	Domain C, conformational	97 ± 9	21 ± 6	98 ± 13	14 ± 3	32 ± 7	53 ± 12	44 ± 3
HC-1	11	Domain B, conformational 523–540	73 ± 5	31 ± 9	81 ± 10	27 ± 9	2 ± 1	2 ± 1	77 ± 1

NOTE. HCVpp produced from isolates shown in Figure 1 were incubated with mAbs (10 μg/mL) for 1 hour at 37°C. HCVpp-antibody complexes then were added to Huh7.5.1 cells. Viral epitopes targeted by the respective antibody, percentage of HCV entry in the presence of antibody (strains VL, VC, $VCVL_{458+478}$, $VLVC_{458+478}$, VA, $VAVL_{447}$, and $VLVA_{447}$), and source or reference of antibody are shown. Means ± standard deviation from at least 3 experiments, each performed in triplicate, are shown.
V, viral variant.



Supplementary Figure 4. Binding of neutralizing anti-E2 HMAb CBH-23 to patient-derived envelope glycoproteins expressed on HCVpp as capture antigens in an enzyme-linked immunosorbent assay (ELISA). HCVpp expressing envelope glycoproteins of variants VL, VA, VC, $VLVA_{447}$, and $VLVC_{458+478}$ were used as capture antigens on GNA-coated ELISA plates. Control (Ctrl) pseudoparticles with absent HCV envelope glycoprotein expression and recombinant soluble E2 (sE2 derived from strain H77)[14] served as negative and positive controls, respectively. Anti-E2 CBH-23 reactivity was detected as described in the Supplementary Materials and Methods section and is indicated as optical density (OD) values at 450 nm. Means ± standard deviation from 1 representative experiment are shown.

Supplementary Table 2. Characteristics of Patients and Viruses Used for Neutralization Studies

					HCVpp neutralization titer, 1/dilution		
Patient number	Age, y	Sex	Genotype	Viral load, IU/mL	VL	VC	VA
1	65	M	1b	2.29×10^5	100	100	100
2	27	F	1b	9.7×10^4	100	3200	200
3	31	F	1b	1.53×10^5	400	3200	400
4	47	M	3a	1.02×10^6	20	20	100
5	58	M	1b	1.15×10^6	100	3200	200
6	72	M	1b	1.50×10^6	20	200	100
7	51	M	4	4.38×10^6	20	20	20
8	69	F	1b	9.7×10^5	20	400	100
9	36	F	1	1.29×10^5	800	1600	100
10	46	M	1a	1.05×10^6	100	800	100
11	55	M	1a	1.54×10^6	400	3200	200
12	56	M	4c/4d	2.41×10^4	20	800	200
13	56	F	4a	1.09×10^6	100	400	400
14	59	F	1b	3.54×10^5	200	800	200
15	62	M	1a	3.37×10^6	20	20	20
16	50	M	4a	1.48×10^6	20	200	20
17	46	M	4a	4×10^5	20	200	100
18	70	F	1b	1.3×10^6	100	800	20
19	77	F	1b	6.2×10^4	20	100	100
20	61	F	1b	2.58×10^4	200	800	200
21	46	F	1b	2.11×10^5	100	400	800
22	36	M	1a	2.04×10^6	20	200	400
23	52	F	4a	9.12×10^5	20	3200	400
24	54	M	1a	9.77×10^5	100	800	200
25	54	M	1b	1.12×10^6	20	100	200
26	54	F	1a	3.38×10^6	20	400	20
27	47	M	3a	6.16×10^5	100	3200	3200
28	43	M	1a	5.75×10^6	20	800	200
29	51	M	4a	1.44×10^6	100	400	400
30	54	M	2c	4.67×10^5	100	100	3200
31	51	M	1a	6.16×10^6	100	400	100
32	39	M	4a	1.12×10^6	20	200	800
33	62	F	4f	2.88×10^6	20	800	20
34	46	M	4k	3.54×10^5	20	20	100
35	42	M	1a	9.54×10^5	400	800	400
36	54	M	2c	4.67×10^5	200	3200	100
37	34	M	3a	3.23×10^6	20	20	100
38	47	M	3a	7.94×10^4	20	400	20
39	30	F	1b	1.00×10^6	20	200	400
40	47	F	1b	2.29×10^6	100	400	200
41	52	M	1a	1.73×10^6	200	3200	400
42	34	M	1b	1.45×10^6	3200	3200	200
43	46	M	1a	4.34×10^6	200	800	400
44	66	F	1b	3.89×10^5	200	1600	200
45	29	F	1a	1.08×10^5	400	400	200
46	45	M	3a	2.78×10^5	20	200	200
47	65	F	4f	1.46×10^6	20	3200	20
48	55	M	1a	8.81×10^6	20	800	100
49	53	M	1a	1.15×10^6	100	100	100
50	40	M	3a	2.46×10^6	100	3200	200
51	48	F	1a	1.00×10^5	20	800	20
52	37	M	1a	5.08×10^6	20	400	200
53	47	M	3a	6.8×10^6	100	1600	400
54	37	M	1a	1.84×10^6	800	800	200
55	65	F	1b	2.18×10^6	100	100	800
56	45	F	1a	3.93×10^6	1600	1600	400
57	49	M	4a	2.06×10^6	800	3200	200
58	30	M	1b	7.21×10^5	100	800	200
59	31	M	3a	6.66×10^6	100	200	200
60	37	M	1a	6.70×10^6	20	100	100

Supplementary Table 2. Characteristics of Patients and Viruses Used for Neutralization Studies

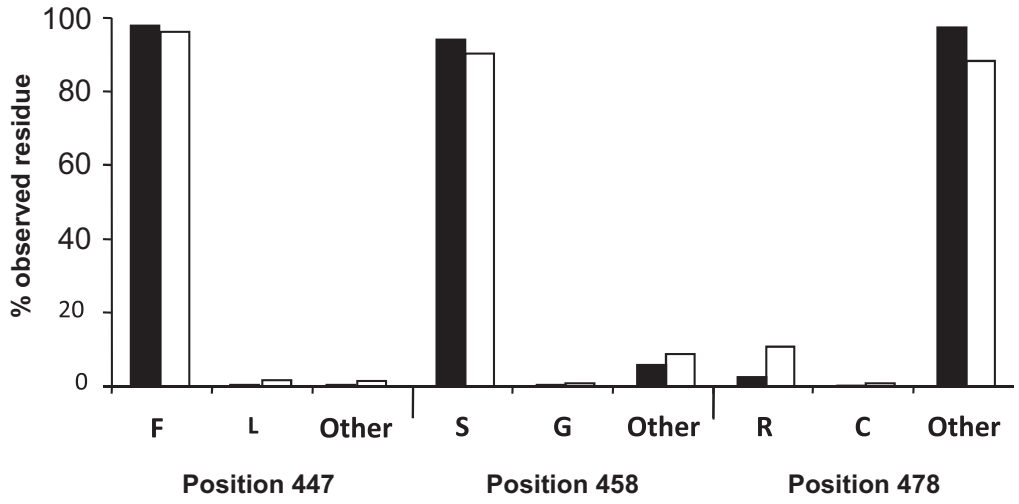
					HCVpp neutralization titer, 1/dilution		
Patient number	Age, y	Sex	Genotype	Viral load, IU/mL	VL	VC	VA
61	49	M	1a	3.16×10^5	20	800	20
62	43	M	1	6.83×10^5	20	20	20
63	69	M	1b	4.7×10^5	20	20	200
64	48	M	1a	3.28×10^6	20	3200	100
65	46	M	3a	8.55×10^5	20	800	100
66	51	M	1b	1.07×10^6	20	200	1600
67	43	M	1b	4.27×10^5	20	100	800
68	36	M	3a	1.14×10^6	20	800	20
69	53	F	1b	3.06×10^5	20	400	20
70	24	F	3a	1.29×10^6	20	20	20
71	63	M	1b	3.01×10^6	100	200	100
72	44	M	1	1.10×10^5	20	3200	200
73	28	M	3a	1.85×10^6	20	3200	20
74	54	M	1b	1.29×10^5	20	3200	20
75	17	F	1b	2.41×10^5	20	20	200
76	40	M	3a	1.26×10^6	20	20	100
77	35	M	1b	8.89×10^5	20	20	800
78	36	F	6a	1.4×10^7	20	100	400
79	70	F	1b	1.13×10^5	100	100	400
80	62	M	1a	2.68×10^6	100	200	20
81	70	M	1b	2.85×10^5	20	200	3200
82	63	M	1b	1.95×10^5	200	400	400
83	33	M	1a	1.76×10^6	100	200	800
84	35	M	1a	2.78×10^6	20	20	200
85	60	F	1	6.39×10^5	20	200	100
86	57	M	3a	1.22×10^6	200	3200	400
87	60	M	1	3.6×10^6	100	3200	20
88	49	M	4	2.24×10^6	20	1600	20
89	37	M	4	9.35×10^5	100	800	100
90	55	M	1a	3.77×10^6	20	3200	100
91	47	M	1a	2.36×10^6	20	1600	20
92	72	M	3a	3.83×10^5	20	400	20
93	79	M	1b	2.81×10^5	100	1600	100
94	58	F	1b	6.58×10^5	100	3200	200
95	50	M	3a	6.07×10^5	20	3200	100
96	67	F	1b	4.13×10^5	100	800	20
97	49	M	3a	5.22×10^5	200	400	200
98	53	F	1b	2.31×10^6	20	400	1600
99	37	M	1a	1.87×10^5	100	3200	200
100	54	F	4a	9.23×10^5	20	200	100
101	39	M	1a	1.76×10^5	100	800	200
102	51	F	2b	1.10×10^6	100	3200	800

NOTE. HCVpp were incubated with anti-HCV–positive sera from 102 patients with chronic HCV infection (ClinicalTrial.gov identifier NCT00638144). Patient number, age, sex, viral genotype, and load in serum are indicated. HCVpp-antibody complexes were added to Huh7.5.1 cells and infection was analyzed as described in Figure 4. Calculation of neutralization and determination of background and thresholds for neutralization were performed as described in Figure 6. Neutralization titers obtained by end point dilution are indicated for each variant. Means from at least 3 independent experiments, each performed in triplicate, are shown.
V, viral variant.

Supplementary Table 3. HCVcc Neutralization Titers

Patient number	HCVcc neutralization titer, 1/dilution		
	VL	$VLVC_{458+478}$	$VLVA_{447}$
11	400	1600	800
28	20	1600	800
33	20	400	400
35	400	1600	1600
36	200	1600	3200
45	800	1600	800
65	20	1600	1600
66	20	3200	800
68	20	1600	1600
94	100	3200	800
98	100	800	3200
99	100	3200	1600

NOTE. Results were confirmed using chimeric HCVcc expressing the HCV envelope glycoproteins depicted in Figure 7 and using 12 representative sera from patients. Neutralization assays were performed using a similar protocol as described in Supplementary Tables 2 and 3. Means from at least 3 independent experiments, each performed in triplicate, are shown.
V, viral variant.



Supplementary Figure 5. Distribution of residues at positions 447, 458, and 478 of HCV E2 sequences in the European HCV databases. Distribution of residues at positions 447, 458, and 478 for HCV complete E2 sequences from all subtypes (black) and from subtype 1b only (white) within the European Hepatitis C Virus databases[7] (available: http://euhcvdb.ibcp.fr). F and S are the predominant residue at positions 447 and 458 (F447, 98.4%; 1b, 96.2%; S458 all, 94%; 1b, 90.3%). The position 478 is variable (it belongs to HVR2) but R (all, 2.4%; 1b, 10.8%) is more frequent than C (all, 0.2%; 1b, 0.9%).

Nous avons développé la base HBVdb (http://hbvdb.ibcp.fr) pour permettre aux chercheurs d'étudier les caractéristiques génétiques et la variabilité des séquences du virus de l'hépatite B (VHB), ainsi que la résistance virale aux traitements. HBVdb contient une collection de séquences annotées automatiquement sur la base de génomes de référence annotés manuellement, ce qui assure une nomenclature normalisée pour toutes les entrées de la base. HBVdb est accessible via un site Web dédié avec des outils d'analyses génériques et spécialisés (annotation, génotypage, détection de profils de résistance), et des jeux de données pré-calculés.

La polymérase du VHB est la principale cible des traitements anti-VHB. Les analogues de nucléos(t)ides (NA) inhibent l'activité de la transcriptase inverse (RT), mais il existe des mutations de résistance aux NA. Cependant, un autre domaine enzymatique pourrait être une cible potentielle : la RNase H, liée au domaine RT, permettant la dégradation de l'ARN durant la transcription inverse. Pour pallier l'absence d'une structure expérimentale résolue, et grâce à l'analyse de séquences à partir de HBVdb, nous avons construit le modèle par homologie de la RNase H, qui a permis de définir les caractéristiques de cette RNase H de type 1. Enfin pour vérifier des hypothèses émises à partir de ce modèle, et pour le placer dans son contexte, nous avons construit un modèle plus étendu de la polymérase du VHB, qui comprend la les domaines RT et RNase H, et contribue à répondre à la question sur l'existence d'un domaine de connexion les reliant. Nous avons utilisé notre modèle pour analyser les interactions entre le site catalytique de la RT et le ténofovir.

Mots clés : Virus de l'hépatite B, base de données, bioinformatique, annotation, génotype, résistance, polymérase, modélisation moléculaire par homologie.

Development of HBVdb, a knowledge database for Hepatitis B Virus, for the study of drug resistance. Integration of sequence analysis tools and application to the polymerase molecular modeling

We developed HBVdb (http://hbvdb.ibcp.fr) to allow researchers to investigate the genetic characteristics and variability of the HBV sequences and viral resistance to treatment. HBVdb contains a collection of computer-annotated sequences based on manually annotated reference genomes. The automatic annotation procedure ensures standardized nomenclature for all HBV entries across the database. HBVdb is accessible through a dedicated website integrating generic and specialized analysis tools (annotation, genotyping, resistance profile detection), and pre-computed datasets.

The HBV polymerase is the main target of anti-HBV drugs, nucleos(t)ides analogues (NA), which inhibit the activity of reverse transcriptase (RT), but NA resistance mutations appeared. Nevertheless, another enzymatic domain could be a potential drug target: RNase H domain, linked to RT, and involved in degradation of the RNA during the reverse transcription. To overcome the lack of experimental solved structure, thanks to sequences analysis from HBVdb, we built an homology model of RNase H, which helped to define the features of this type 1 RNase H. Finally, to confirm assumptions from this model and to put it in a more global context, we built an extensive HBV polymerase model, which includes the RT and RNase H domains, and helps to answer the question about the existence of connection domain linking them. We performed analyses on this model, regarding the interactions between the RT catalytic site and the Tenofovir, mapping known resistance mutations and the most variables positions of the HBV polymerase.

Keywords : Hepatitis B Virus, database, bioinformatics, annotation, genotype, resistance, polymerase, homology molecular modeling.

Equipe BioInformatique : Structure et Interactions (BISI)
BMSSI - UMR 5086 CNRS - Université Lyon 1 - http://www.ibcp.fr
IBCP - 7, passage du Vercors - 69367 Lyon - FRANCE