Les connaissances sur les associations PDV-guêpes parasitoïdes montre qu'elles forment des symbioses uniques dans le sens où elles associent de manière étroite un organisme eucaryote à un virus. On a vu que cette association était essentielle au succès parasitaire de la guêpe. En effet, le virus en contrôlant et en manipulant les fonctions physiologiques de l'hôte, assure le développement de la guêpe et garantit sa propre transmission via le génome de la guêpe.
La particularité du génome viral réside dans sa taille et sa structure, mais aussi dans l'information codée par les gènes viraux. Le génome contenu dans les particules virales est presque totalement dépourvu de gène d'origine virale. Au contraire, il est constitué de gènes codant pour des facteurs de virulence potentiels, probablement acquis depuis le génome de la guêpe.

Comme cela a été illustré en introduction, les symbioses constituent de véritables moteurs de l'évolution dans le sens où elles permettent des évolutions plus rapides par rapport à aux mutations ou aux recombinaisons génétiques. De très nombreuses associations symbiotiques impliquant des micro-organismes bactériens ou mycéliens ont été sélectionnées au cours de l'évolution. Ces symbioses bien connues, interviennent dans des fonctions très différentes mais elles ont toutes contribuées à l'adaptation des organismes à leur milieu.
En revanche, les associations impliquant des virus sont très peu connues, et malgré certains exemples ont décrit la domestication de gènes d'origine virale, réutilisés pour des fonctions physiologiques (Mallet et al., 2004). Mais la domestication d'une machinerie virale complète dans le cas des PDV, représente un cas unique de symbiose virus-eukaryote à ce jour.
En quoi cette symbiose constitue-t-elle un avantage adaptatif ? Comment les fonctions transmises à l'hôte via le virus sont-elles impliquées dans le succès parasitaire de la guêpe ? Comment ces fonctions évoluent-elles au cours de l'histoire de la symbiose ?

Mon travail de thèse s'inscrit dans la compréhension du rôle de cette symbiose dans le succès parasitaire et l'évolution de l'hôte parasitoïde. En étudiant **l'évolution de facteurs de virulence** particuliers et en explorant les **fonctions potentiellement ciblées** chez l'hôte, j'ai voulu comprendre les enjeux évolutifs impliqués dans cette association et le devenir des gènes acquis au cours de l'évolution de l'association.

En regard des gènes codés par le virus, de leur organisation en familles multigéniques et de leur acquisition progressive au cours de l'association, le génome du polydnavirus apparaît comme une entité dynamique, capable de perdre ou d'intégrer de nouvelles fonctions.
Parmi les fonctions acquises par le virus, les PTP qui jouent un rôle clé dans les processus de régulation, sont apparues tôt au cours de l'évolution de l'association. Après son intégration dans le génome de la guêpe, le gène ancêtre des PTP a été hérité au cours de l'évolution et a connu une forte expansion pour former aujourd'hui la famille multigénique la plus large observée chez les PDV. En effet, l'expansion de cette famille constitue un cas exceptionnel dans l'histoire des gènes du polydnavirus. Actuellement, on trouve 13 copies de ce gène chez MdBV, 9 copies chez GiBV encore partiellement séquencé et 27 copies chez CcBV (Espagne et al., 2004; Gundersen-Rindal and Pedroni, 2006; Webb et al., 2006).

Les études menées sur le rôle des PTP au cours du parasitisme ont montré que ces gènes étaient exprimés au cours du parasitisme et que leur expression variait dans le temps et selon le tissu de l'hôte (Provost et al., 2004). Ces protéines régulent les voies de signalisation cellulaire par des réactions de déphosphorylation et jouent ainsi un rôle clé dans la régulation des voies de signalisation en particulier du système immunitaire (Ibrahim et al., 2007; Ibrahim and Kim, 2008; Pruijssers and Strand, 2007). Cependant, parmi les nombreux gènes de cette famille, certains ne présentent pas de site catalytiquement actif. Ces formes particulières pourraient réguler la réponse immunitaire, probablement en séquestrant les protéines phosphorylées, modulant ainsi l'activité phosphatase intracellulaire au niveau des hémocytes par exemple (Ibrahim and Kim, 2008).

Si de nombreux travaux ont été focalisés sur la fonction de ces gènes et sur les processus cellulaires potentiellement ciblés, aucune étude ne permet d'expliquer l'extraordinaire expansion de ce gène en famille multigénique. En effet, comprendre la fonction des PTP et leur rôle au cours de l'évolution de l'association implique aussi de comprendre quelles sont **les forces évolutives** et **les mécanismes moléculaires** qui ont façonné l'évolution de ce gène et probablement sa fonction.

Contrairement aux PTP, les cystatines sont apparues tardivement au cours de l'évolution de l'association et sont ainsi présentes chez un nombre restreint d'espèces. A ce jour une copie unique de cystatine a été isolée chez GiBV et 3 copies sont présentes sur le même cercle d'ADN chez CcBV. Les PTP et les cystatines sont des familles de gènes très différentes en regard de leur distribution dans les espèces de virus et de leur expansion en famille multigénique. On peut alors

supposer qu'elles n'ont pas été soumises aux mêmes pressions de sélection au cours de l'évolution de l'association. En effet la présence des PTP dans plusieurs génomes viraux suggère que ces protéines ciblent des fonctions communes et essentielles pour la défense de l'hôte. Au contraire, si les cystatines sont apparues plus tardivement nous pouvons supposer qu'elles codent pour des fonctions moins capitales mais qui confèrent une plus grande efficacité au parasitisme.

Les cystatines forment un complexe inhibiteur avec des protéases à cystéine de la famille C1, comme la papaine chez les plantes ou les cathepsines chez les animaux. Chez les plantes, les cystatines jouent un rôle important dans la protection contre les insectes phytophages qui libèrent des protéases afin de digérer les tissus de la plante (Arai et al., 2002). Chez les nématodes filaires, les cystatines constituent un facteur de virulence majeur qui inhibe la réponse immunitaire (Maizels et al., 2001; Schierack et al., 2003). Le rôle des cystatines virales dans l'interaction hôte-parasitoïde est encore inconnu. Toutefois, la purification d'une forme recombinante de la cystatine 1 chez CcBV montre que celle-ci est active contre des protéases à cystéine. De plus, nous savons que les cystatines sont exprimées très tôt au cours du parasitisme et en grande quantité. Cette expression est maintenue à un taux élevé 3 jours après le parasitisme. Ce schéma d'expression suggère que les cystatines doivent jouer un rôle important, probablement dès le debut de l'infestation (Espagne et al., 2004).

Pour comprendre le **rôle** des cystatines virales dans le parasitisme, il est nécessaire de comprendre en quoi les cystatines constituent un **facteur de virulence** potentiel et d'étudier la **fonction ciblée** par les cystatines virales chez l'hôte lépidoptère. De plus, en étudiant l'évolution de ce gène nouvellement acquis, il nous sera plus facile de comprendre son **rôle dans l'évolution de l'association et dans le succès parasitaire.**

Modèle expérimental: *Cotesia congregata* / *Manduca sexta*

Le parasitoïde*: Cotesia congregata*

Nous avons travaillé à partir de guêpes du genre *Cotesia,* guêpe de la famille des Braconidae et de la sous-famille des Microgastrinae. La sous-famille des Microgastrinae constitue l'une des sous-familles les plus diversifiées de la famille des Braconidae (Mason, 1981). Tous les Microgastrinae sont des koinobiontes endoparasitoïdes de lépidoptère (Shaw, 1991). Les guêpes du genre *Cotesia* présentent pour la plupart un spectre d'hôte étroit malgré certaines rares exceptions où le parasitoïde peut avoir un spectre d'hôte plus large que le cadre de la famille. Le genre *Cotesia* est un groupe très diversifié en termes de richesse spécifique et de spectre d'hôte (Mason, 1981). Le développement larvaire compte 3 stades dont les 2 premiers se déroulent à l'intérieur de l'hôte lépidoptère. Au 3eme stade, les larves percent la cuticule de l'hôte et tissent un cocon duquel émergeront les guêpes adultes.

L'hôte*: Manduca sexta*

L'étude des fonctions potentiellement ciblées par le virus a été réalisée chez le lépidoptère *Manduca sexta*. *M. sexta* appartient à la famille des Sphingidae et est distribuée sur tout le continent américain. Les larves se nourrissent de feuilles de plantes de la famille des Solanaceae. Le cycle larvaire présente 5 stades séparés chacun par des mues. Au terme du 5eme stade, la larve entre dans le stade pupal qui s'achève par la métamorphose en adulte.
La propagation de cet insecte ravageur est biologiquement contrôlée par les guêpes du genre *Cotesia.*

Encadré 2 Modèle expérimental

B-Objectifs

Ce travail de thèse porte sur les facteurs de virulences codés par le virus et plus particulièrement les PTP et les cystatines. Nous avons étudié les processus évolutifs qui ont façonné leur divergence depuis leur intégration et inféré leur rôle dans l'adaptation des guêpes. De plus, pour comprendre le rôle de ces facteurs, nous avons recherché leurs cibles chez l'hôte.

Mon travail de thèse s'articule ainsi en cinq grands chapitres :

- Le chapitre III présente l'étude des PTP dont l'acquisition s'est produite tôt au cours de l'histoire de l'association. Ce travail, rédigé sous la forme d'un article en préparation, vise à caractériser les **mécanismes moléculaires** et les **forces évolutives** qui ont conduit à l'expansion des **PTP** en **famille multigénique**.

- Dans le chapitre IV, en étudiant l'évolution d'un gène viral récemment acquis au cours de l'histoire de la symbiose, nous avons caractérisé les **marques moléculaires** portées par un **facteur de virulence** potentiel impliqué dans l'interaction avec l'hôte Lépidoptère. Cette étude réalisée sur les **cystatines virales** a fait l'objet d'un article accepté dans BMC Biology.

- Le chapitre V, porte sur les protéases à cystéine qui sont des cibles potentielles du parasitisme. La première partie de ce chapitre porte sur l'étude de la régulation de ces protéines chez l'hôte *M. sexta* au cours du parasitisme (manuscrit en préparation).
La deuxième partie de ce chapitre, présente les travaux réalisés pour isoler les **protéases à cystéine effectivement ciblées** au cours du parasitisme.

- Dans le chapitre VI, nous avons étudié en quoi **l'histoire évolutive de l'hôte** et **du symbiote viral** étaient **corrélées** et en quoi l'évolution de cette fonction au cours de l'histoire de la symbiose pouvait être impliquée dans les **processus de diversification** des guêpes parasitoïdes. Cette étude est en cours de préparation.

- Dans le chapitre VII, l'ensemble des résultats obtenus est résumé et discuté.

III-Large gene expansion in mutualistic polydnaviruses: Molecular and evolutionary mechanisms at the origin of PTPs

Serbielle Céline, Dupas Stéphane, Héricourt François, Huguet Elisabeth, Drezen Jean-Michel.

Abstract

Gene duplications have been proposed to be the main mechanism involved in genome evolution and in acquisition of new functions. They constitute therefore an important source of innovations and adaptations. Gene duplications are found to be particularly common in mutualistic viruses associated with parasitoid wasps. In these systems the virus is integrated into the wasp genome and virions injected in the parasitoids' hosts are essential for parasitism success. These viruses encode virulence factors which are involved in host immune suppression and developmental arrest. How did gene family expansion occur and what are the evolutionary forces inducing gene copy divergence? In order to understand gene duplication and divergence mechanisms which occurred during virus-wasp associations, we studied the protein tyrosine phosphatase (PTP) gene family which is the largest virus gene family described in these viruses.

Here, we show that viral PTPs expansion occurred through three main mechanisms; by duplication of large genomic segments (segmental duplication) and by tandem and dispersed gene duplications within the viral genome. These duplication events became sources of evolutionary innovations conferring to wasps adaptive properties. Indeed, PTP gene copy evolution was shown to undergo conservative evolution along with episodes of adaptive evolution which were correlated with duplication and wasp speciation processes. Altogether duplications and subsequent gene copy evolution likely contributed to the different patterns of PTP gene regulation and activities observed today.

Given the essential role played by the virus in wasp parasitism success and the extraordinary expansion of PTP genes, we can propose that PTP duplication contributed to wasp adaptation and diversification.

INTRODUCTION

Gene duplications have been recognized as an important source of evolutionary innovation and adaptation. The contribution of gene duplication to the evolution of new functional genes has been widely demonstrated in various organisms (Arguello et al., 2006; Katju and Lynch, 2003). Duplicated genes usually can be classified into tandem and dispersed duplicates, and duplicated copies supposedly evolve to improve the ancestral function. Two major questions are addressed when studying gene duplications: what are the molecular mechanisms underlying duplications and how are duplicated copies maintained during evolution? Two main hypotheses have been proposed to explain gene duplication evolution and acquisition of new function. Classical models propose that after gene duplication, one copy evolves under purifying selection and conserves the parental function whereas the extra copy is assumed to be neutral (Force et al., 1999; Hughes, 1994). In contrast, alternative models propose that duplications are adaptive and that duplicated copies diverge under positive selection for acquisition of novel function (Bergthorsson et al., 2007; Des Marais and Rausher, 2008).

Here we present a functional gene family encoded by a mutualistic virus involved in a host-parasitoid interaction. This family has been subjected to particularly strong expansion and constitutes therefore a remarkable model to study evolutionary processes involved in gene duplications.

The organization of genes into families constitutes a common characteristic in viruses of the polydnaviridae family associated with parasitoid wasps. The gene families encode putative virulence factors, some of which were proven to disrupt lepidopteran host physiology (Desjardins et al., 2007; Espagne et al., 2004; Lapointe et al., 2007; Webb et al., 2006). In wasp-PDV associations, PDVs persist as stably integrated proviruses in the genome of their associated wasp (Desjardins et al., 2007 ; Bezier 2008) and replicate in female ovaries only. Virus particles are injected into the lepidopteran host during wasp oviposition at the same time as wasp eggs. PDVs do not replicate in the parasitized host insect, but viral gene products suppress the host immune system and cause physiological alterations ensuring parasitoid development (Asgari et al., 1996; Beckage and Gelman, 2004; Tanaka et al., 2000). This unique example of mutualism between a virus and an eukaryotic organism, constitutes a real evolutionary success in regards to the tens of thousands of parasitoid species which carry PDVs. All wasps carrying PDV are found within the two separate lineages which constitute the Ichneumonoidea wasp superfamily: Ichnoviruses (IV) are associated with Ichneumonid wasps and Bracoviruses (BV) are found in Braconid wasps (Turnbull and Webb, 2002). Recently, PDVs associated with the Banchinae

wasps, previously belonging to the Ichnovirus genera, have been shown to be sufficiently distinct from both Ichnoviruses and Bracoviruses to justify the creation of a third PDV group (Lapointe et al., 2007). Each PDV genera presents distinct morphological and packaging characteristics (Lapointe et al., 2007; Webb et al., 2006) and the absence of PDVs in the basal Ichneumonoids suggests that IV, BV and Banchinae viruses arose independently in the three wasp lineages (Lapointe et al., 2007; Whitfield, 2002). To date, several PDV genomes have been sequenced and have all been shown to have large genomes segmented in multiple dsDNA circles. The other common and original feature of PDVs is that putative genes encoding virulence factors are organized in multigene families (Desjardins et al., 2007; Espagne et al., 2004; Lapointe et al., 2007; Webb et al., 2006). The diversification of virulence genes into families may reflect the adaptive pressures imposed on PDV genome evolution and underline their role in wasp parasitism.

The Braconid wasps carrying PDVs form a monophyletic group called the Microgastroid complex which was estimated, thanks to the calibration of the molecular clock by fossil records, to have arisen 103 Mya ago from an unique BV-braconid ancestral association (Murphy et al., 2008). Some genes encoding IκBs and protein tyrosine phophatases (PTP) respectively are common to most sequenced bracoviruses, suggesting they were acquired early in the course of the wasp-bracovirus evolution. IκB genes found in the three PDV lineages, encode proteins which are inhibitors of nuclear transcriptional factors involved in vertebrate and in *Drosophila* immune responses (De Gregorio et al., 2001; Falabella et al., 2007; Hoffmann, 2003; Thoetkiattikul et al., 2005). PTP genes are not found in IV and form a distinct clade in Banchinae, the lack of evidence of a common ancestor between PTPs from Banchinae and BV suggests that PTPs evolved separately in these two virus lineages (Lapointe et al., 2007). In all Bracovirus genomes described so far, PTPs belong to the largest gene family with 27 members in *Cotesia congregata* Bracovirus (CcBV), 13 members in *Microplitis demolitor* Bracovirus (MdBV) and at least 9 members in *Glyptapanteles indiensis* Bracovirus (GiBV) genome, which is partly sequenced (Desjardins et al., 2007; Espagne et al., 2004; Webb et al., 2006). PTP genes are known to play a key role in the control of signal transduction pathways by dephosphorylating tyrosine residues on regulatory proteins (Andersen et al., 2001). All PDV PTPs studied so far are expressed in virus infected hosts but only a subset of these genes encodes catalytically functional PTPs (Ibrahim and Kim, 2008; Provost et al., 2004; Pruijssers and Strand, 2007). The "inactive" PTPs have been suggested to play a role in trapping phosphorylated proteins to impair cellular PTP activity in a competitive way (Provost et al., 2004). Moreover PTP gene expression is regulated in a tissue specific and time dependant manner (Gundersen-Rindal and Pedroni, 2006; Ibrahim et al., 2007;

Provost et al., 2004; Pruijssers and Strand, 2007). Bracovirus PTPs appear therefore to be important virulence factors which have undergone a high expansion rate and a high functional divergence (Bézier et al., 2007). In this context, Bracovirus PTPs emerge as an interesting gene family model to study the mechanistics and the evolution of gene duplication. To date, complete or partial sequence data for Bracovirus genomes are available giving us a support for understanding the genomic organization and the transmission of duplications in a dynamic interaction between a parasitoid wasp, a virus and a lepidopteran host.

Studying PTP gene family evolution enabled us to determine the molecular and evolutionary mechanisms at the origin of this family, and to highlight viral genome plasticity and evolutionary forces at the basis of PTP diversity. We discuss the critical role of duplications and natural selection in regard to functional divergence and adaptation.

Materials and Methods

Wasp specimens

Fourteen PTP genes previously isolated from Bracoviruses associated with Braconid wasps have been studied: PTP P, Q, Y, K, L, C, α, S, M, E, X, H, R and Δ.
These PTP genes were isolated from nine *Cotesia* species were considered: *C. congregata* (laboratory reared, France, Drezen,J-M), *C. chilonis* (laboratory reared, USA, Wiedenmann,. R), *C. flavipes* (Field collected, Kenya, S. Dupas), *C. glomerata* (laboratory reared, Netherland, Vet,L), *C. melanoscela* (Field collected, France, C. Villemant), *C. marginiventris* (laboratory reared, USA, Joyce,A), *C. vestalis* (Field collected, Benin, Guilloux,T), *C. rubecula* (laboratory reared, Netherland, Smid,H), *C. sesamiae* (Field collected, Kenya, S. Dupas). All specimens were placed in 95% ethanol and preserved at –20°C until DNA was extracted.

DNA extraction, amplification and sequencing

DNA was extracted with the « chelex » method from 2 individuals for each species except for *C. rubecula* where 1 individual was used. Briefly, individuals were ground in a 5 % chelex 100 resin (Biorad) solution with proteinase K (0.12 mg/ml) and incubated at 56°C for 30 min, then incubated at 95°C for 15 min and supernatants were collected. The primers were designed

according to the sequences of the CcBV PTPs: each pair of primers flank motifs 1 to 10 that characterize PTPs and are specific for each PTP. Primer sequences are listed in table 1. PCR conditions varied in stringency depending on whether amplified PTP genes belonged to closely related species (55°C annealing temperature and 1,5 mM of MgCl2) or to distantly related species (annealing at 45°C and 3 mM of MgCl2) of CcBV. One µl of DNA was used for each PCR reaction. The standard PCR program was : 95°C for 2,5 min ; 30 cycles at 95°C for 30 sec, annealing for 45 sec, 72°C for 60 sec ; 72°C for 5 min. The amplimers were purified with the Qiaquick kit (Qiagen) and sequenced directly. For most PTP genes direct sequencing was possible and sequence profiles did not show multiple peaks suggesting allele mix. For PTP Cα and PTP EX, direct sequencing was not possible therefore cloning of PCR products (Qiagen cloning kit) was performed and 10 clones for each gene were sequenced. The sequencing reactions were performed with the BigDye Terminator Sequencing Kit (Perkin Elmer ABI) and analysed on an ABI PRISM 3100 Genetic Analyzer.

Table 1 : Primers used for PTP amplification in different *Cotesia* species

Name	Direction	Sequence
PTPR5	Forward	TGGATCGGACTATTGATGAGC
PTPR3	Reverse	ACAGCTTGTTGGATCGGAGT
PTPΔ5	Forward	GCAAACAAAATGGCACATGA
PTPΔ3	Reverse	TCGGACCGGACTTGTCTTTA
PTPH5	Forward	CCACATTTTTCAAAGTTGGTGA
PTPH3	Reverse	CTGAACAACAAATCCACGTCA
PTPS5	Forward	TGGCTACCAACCTCTCAATG
PTPS3	Reverse	TCCAGCGACAATAAATACGC
PTPM5	Forward	CCGATTTGTTTGCACCTTTT
PTPM3	Reverse	AAGTGCAACAAAACACTGTGC
PTPE5	Forward	TGAGCAAGTAGCCGAATCAAG
PTPE3	Reverse	CGATGACAGAATAATCGTTT
PTPX5	Forward	GAAGCAAGTAGCTGAATCTGA
PTPX3	Reverse	CGATTACAGAGAAATCGGAA
PTPL5	Forward	GAATGCAAAAACTCGCCATT
PTPL3	Reverse	TGCAGGCAATCGTATCTTTG
PTPK5	Forward	TTTTCTGGAGACCTGGGAAA
PTPK3	Reverse	GAACGCGTTAAATAGAAACGAA
PTPQ5	Forward	TGGGTTGTGGCAACTCTAAA
PTPQ3	Reverse	GATATGTCAATGGCGCAGAA
PTPP5	Forward	TTCAAAAACGCTAAGCCGTAA
PTPP3	Reverse	TGCCTTTCCTTTCTTAGATTCG
PTPY5	Forward	TGGGGAGTGGAAATTCTAAGTC
PTPY3	Reverse	TCAACATAACAAAGCAAAACTGC
PTPC5	Forward	CGAAGAGCTATCTGCCGTTG
PTPC3	Reverse	TGTTTTATCGAGTGAGTTCT
PTPα5	Forward	TGAGTACCGAATTCGAAGAGC
PTPα3	Reverse	TGTTTTATCAATTGAGTCCC

PTP duplication patterns

In order to understand molecular mechanisms which induced the high PTP gene numbers we searched for homologous PTP genes in CcBV, CvBV, GiBV and MdBV. Thirty-seven PTP protein sequences from CvBV, 9 protein sequences from GiBV, 13 protein sequences from MdBV and 27 PTP protein sequences from CcBV were blasted and genes were considered to be orthologous when identities were higher than 80%. PTP genes were mapped within virus genomes by using gene positions in Genbank.

Sequence analysis and phylogeny

We isolated 87 PTP genes from *Cotesia* bracoviruses and other orthologous PTP sequences from CcBV, CvBV, GiBV and MdBV PDV genome sequencing projects were joined to the analysis (Espagne et al., 2004; Gundersen-Rindal and Pedroni, 2006; Webb et al., 2006). (CcBV PTP K, L, Q, P, M: AJ632304; CcBV PTP S, E,C:AJ632313; CcBV PTP α, X, Y: AJ632319; CcBV PTPH: AJ632307, CcBV PTPR: AJ632310; CvBV PTP2: AY871265; CvBV PTP3: AY651829; CvBVPTP6: AY651829; CvBV PTP10: AY651828; CvBV PTP11:DQ075354; GiBV PTP4: AY871265; GiBVPTP3:AY871265)
Translated sequences were aligned manually in McClade version 4.03 (Maddison, 2001) based on the PTP conserved motifs (Andersen et al., 2001).

This sequence alignment was used to construct a tree in order to have an overview of PTP evolution in Microgastrinae. Using Modeltest version 2.2 (Posada and Crandall, 1998), the GTR+I+G model of sequence evolution was selected according to the likelihood ratio test (LRT) and the Akaike information criterion (AIC). Bayesian MCMC analyses were performed for the entire data set using MrBayes version 3.12 (Ronquist and Huelsenbeck, 2003). Two independent analysis were run simultaneously for each data set, each consisting of 1000000 generations, sampling every 1000 generations and using four chains and uniform priors. Maximum parsimony (MP) analyses were performed using a heuristic search with stepwise random addition sequence. Support values for internal nodes were estimated using a non-parametric bootstrap resampling procedure after 100 replicates.

PTP genes studied could be separated in 3 monophyletic subclades that shared conserved motifs. Each subclade was studied independently, the tree topologies were obtained using maximum parsimony (MP) in PAUP 4.0b10 (Swofford, 2002) and Bayesian inference in MrBayes 3.12 (Ronquist and Huelsenbeck, 2003). For MP analysis, we performed a heuristic search starting with stepwise addition trees replicated 10 times and using a simple input order of

sequences to get the initial tree. Robustness of MP topologies was assessed by bootstrap with 1000 replicates (full heuristic search) of 10 random stepwise additions. For Bayesian inference, the best substitution model was selected using Modeltest 3.7 (Posada and Crandall, 1998). When this model was not available in MrBayes, the closest generalisation of the selected model was used, which for all clades happened to be the GTR+I+G model. The data were partitioned by codon position. We performed a 1000000 generation run sampled every 1000 generations on 4 incrementally heated chains. The burnin period was estimated by plotting likelihood values against generation time and after 20000 generations all were stabilized.

Branch and site selection analyses

The bayesian consensus trees were chosen as a phylogenetic hypothesis for the estimation of nonsynonymous to synonymous substitution rate ratio (ω = dN/dS) models on each clade using PAML 3.14. Six different models of site- and/or branch-specific ω ratios (Yang and Nielsen, 2002; Zhang et al., 2005) were optimised using Bayesian methods in PAML 3.14 (Yang, 1997). The maximum likelihoods of each model tested were compared between all models by the Akaike information criterion (AIC) and between nested models by the Likelihood ratio test (LRT).

Site specific positive selection was tested by comparing the selective model M8 (beta + ω>1) to the non selective model M8a (beta + ω = 1) in a likelihood ratio test (LRT) (Swanson et al., 2003). Branch specific selection was tested by comparing models M0b (branch specific selection and no variation among sites) to M0 (no branch or site specific selection) using a LRT.
The sites under selection were determined on different clades or subclades. For each gene clade or gene subfamily clade, the best model selected by AIC was considered and the percentage attributed to each class of sites (conserved ω<0.2, nearly neutral 0.2<ω<1, neutral ω=1 and selected ω>1) was recorded. When branch specific selection was shown to better explain lineage evolution by AIC and LRT tests, the site attribution was performed separately on positively selected branches (ω>1 in model M0b) and non positively selected branches (ω<1 in model M0b) using the branch and site model (MA new) (Zhang et al., 2005). In this model, two classes of branches are considered; the foreground branches where positive selection is allowed ($0<\omega_0<1$; $\omega_1=1$; $\omega_2>1$) and the background branches where sites are conserved ($0<\omega_0<1$; $\omega_1=1$). In our case we assigned the branches selected in M0b to the foreground branches and the branches not selected in M0b to the background branches.

Results

The PTP genes studied form three different clades

PTP genes can be organized by clusters based on phylogeny and on conserved motifs. PTP phylogeny revealed three major clades well supported among the Bracoviral PTPs studied (Figure 4). The PQLKYCα clade is formed by P, Q, L, K, Y, C and α PTP genes in which all PTP genes are separated in distinct and well supported monophyletic groups of homologues belonging to different species. Three orthologous PTPs were found in GiBV; PTP2, 3 and 4. The RΔ clade formed by R and Δ PTP gene families is also monophyletic. And finally the MHSEX clade, where the H-S, M, and EX PTP genes formed distinct subclades. No MdBV PTPs were found to group with the PTPs studied suggesting that MdBV PTPs are too distantly related to CcBV and GiBV PTPs to be considered as orthologous genes.

Bracovirus PTP sequences carry the 10 conserved motifs that define the protein tyrosine phosphatase family (Andersen et al., 2001) (Figure 5). Two kinds of motifs can be described, the structural motifs which are involved in PTP secondary or tertiary structure (motifs 2, 3, 4, 5, 6, 7) and motifs which are involved in phosphotyrosine recognition and activity (motifs 1, 8, 9 and 10). Conservation of these motifs in bracovirus PTPs differ depending on PTP clades considered. In the PTP RΔ cluster all motifs are strongly conserved except for a few differences in motif 4 where the conserved residues Gln (Q)-Gly (G) are mutated to Glu(E)-Ala(A) in PTP Δ and to Gln(Q)-Ala(A) in PTP R.

In the MHSEX cluster most motifs are well conserved. Amino acid differences occurring within motifs are usually conserved among species within a same PTP gene. For example, in motif 4 the conserved residues Gln(Q)-Gly(G) are mutated to Glu(E)- Gln(Q) in PTPM or to Gln(Q)- Glu(E) in PTP HS and PTP EX, in motif 5 the conserved residues are mutated to Tyr(Y)-Trp(W) in PTP EX, and in motif 8 the conserved Pro(P) is mutated to Thr(T) or Ala(A) in PTP EX. In all gene families of the MHSEX cluster, the most radical difference occurs in motif 9 where the Cys(C) residue which confers the PTP activity is mutated to Gly (G) or Ser(S). Numerous differences are observed in the PQLKYCα cluster, but again these differences are mostly conserved within PTP gene families. Despite high divergence among these genes, motif 9 carrying the active site shows a high conservation. In conclusion, PTP domains are overall well conserved in Bracovirus PTPs and when differences are observed they are shared within the

same PTP gene cluster.

All together, conserved motifs and phylogenetic analyses allow us to separate the bracovirus PTP studied in three main clusters each constituted by PTP gene subclades.

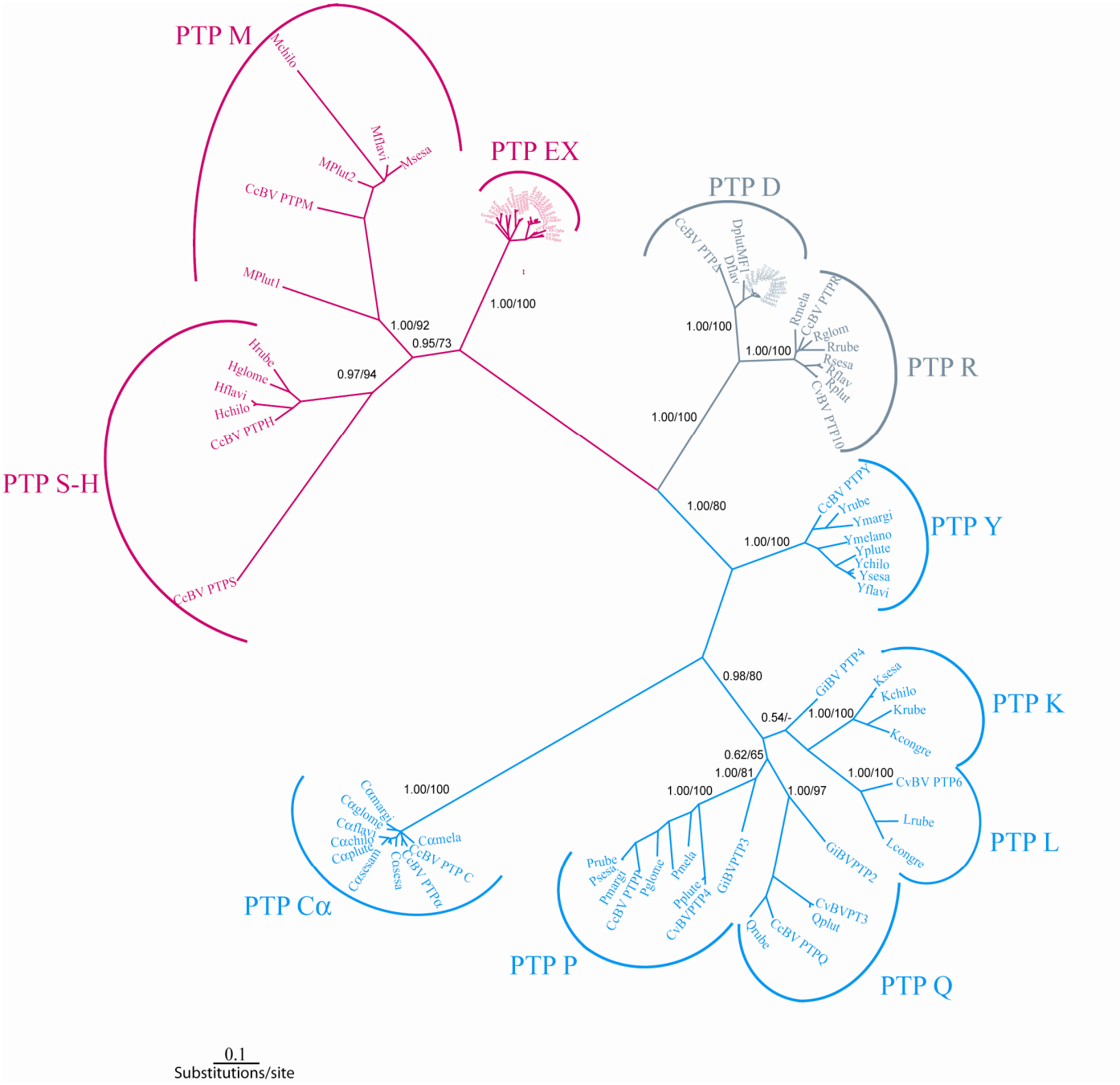
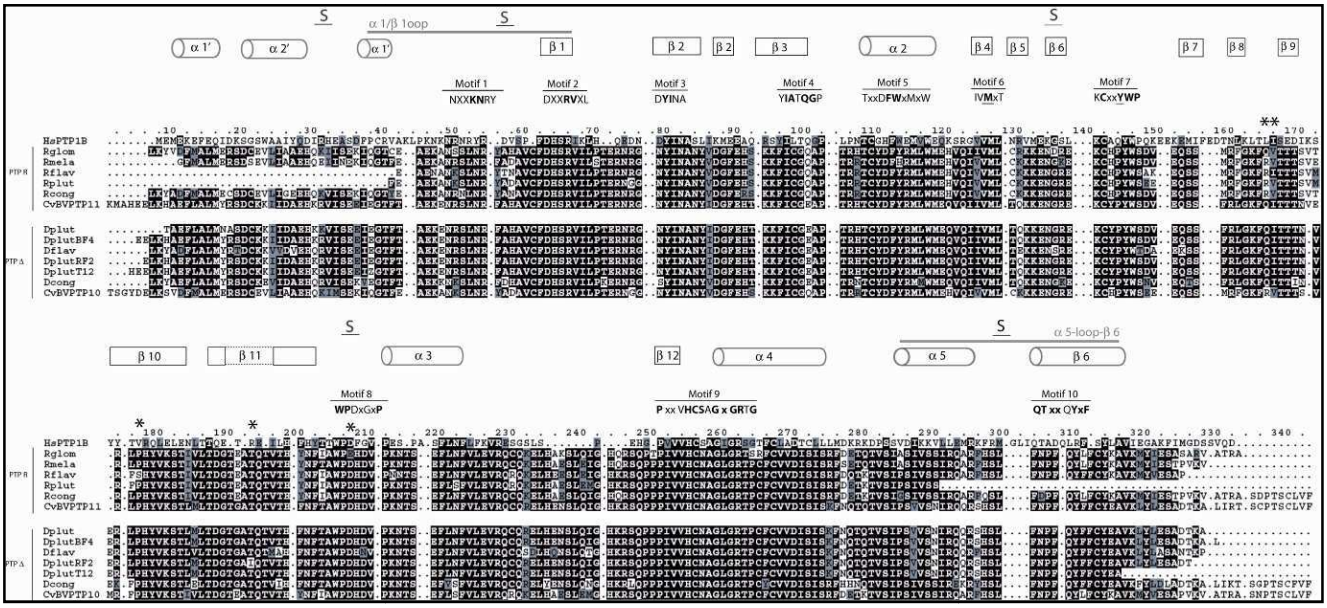


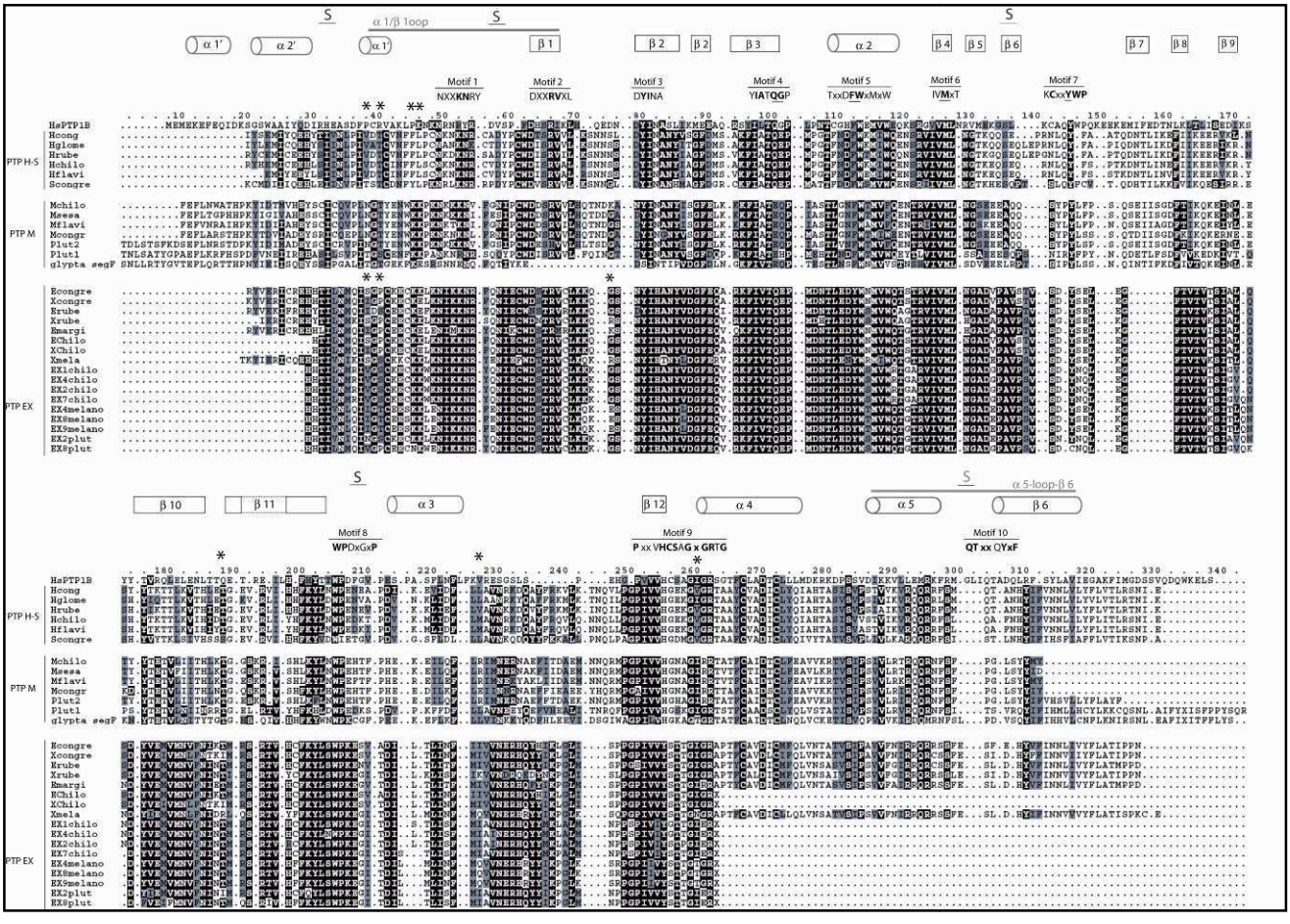
Figure 4 Unrooted PTP phylogenetic tree from Bayesian inferences under the GTR+I+G substitution model

Posterior probabilities and non parametric bootstraps are indicated between major clades. Bootstrap values inside groups of orthologs are not shown for readability. Sequence names are given in abbreviated form.

A)



B)



C)

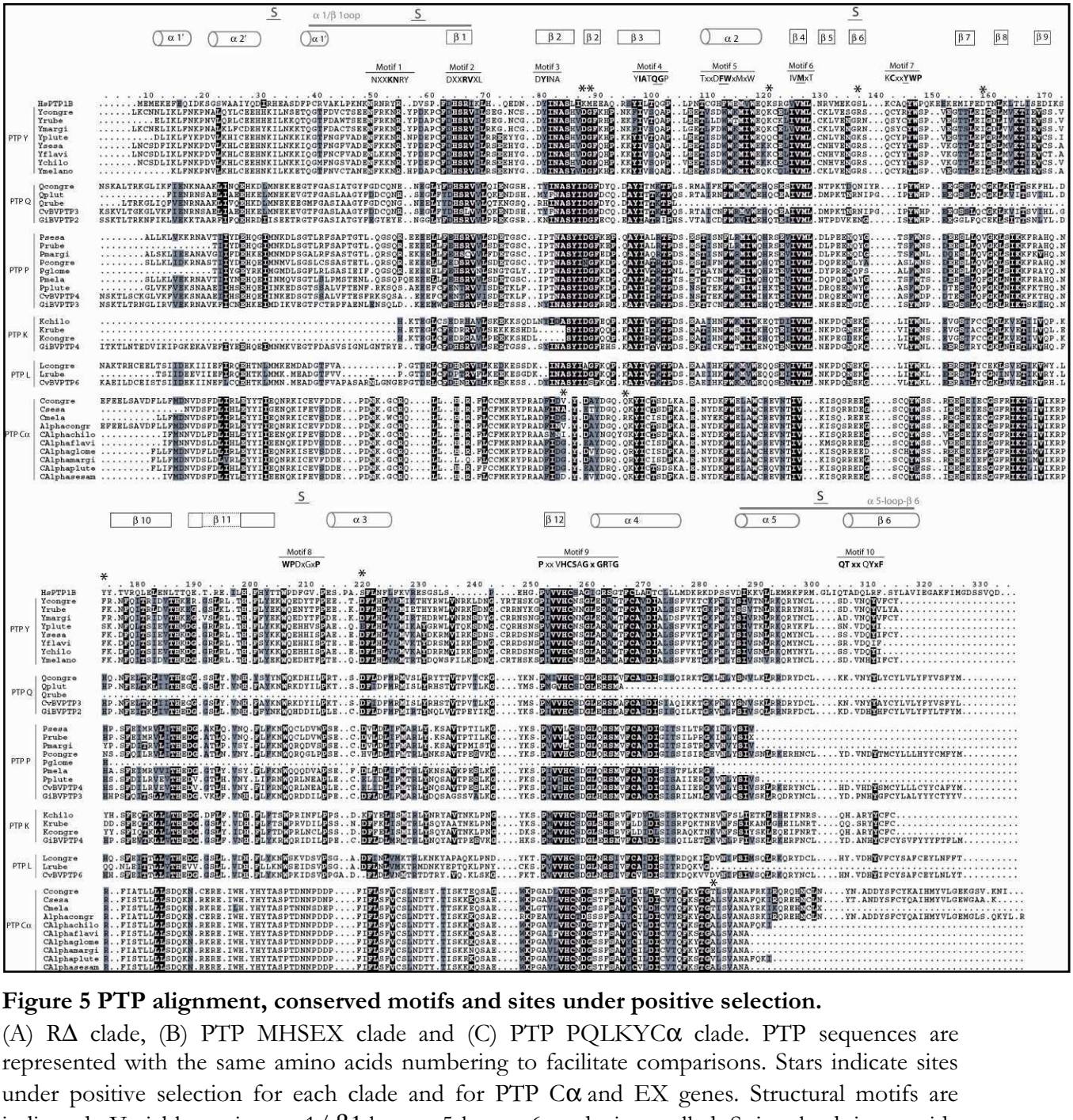


Figure 5 PTP alignment, conserved motifs and sites under positive selection.

(A) RΔ clade, (B) PTP MHSEX clade and (C) PTP PQLKYCα clade. PTP sequences are represented with the same amino acids numbering to facilitate comparisons. Stars indicate sites under positive selection for each clade and for PTP Cα and EX genes. Structural motifs are indicated. Variable regions α1/ β1 loop, α5-loop-α6 and sites called S involved in peptide binding are indicated.

Duplication patterns studied by comparison of three Bracoviruses species

Cross species BLAST analysis using PTPs from CcBV, CvBV, GiBV as queries allowed us to infer homologies between PTPs and to determine orthology between bracovirus genome segments (Figure 6).

27 genes present in CcBV, 23 CvBV PTP genes and 6 GiBV PTP genes were found to be orthologous and among them the synteny was conserved. Indeed the GiBV segment F, CvBV segment 28 and CcBV circle 1 present orthologous PTP families organized in the same order and are therefore considered as homologous segments. The same was shown between CvBV segment 36, 27, 30, 26 and 50 and CcBV circle 17, 10, 7, 2 and 4 respectively (Figure 6).

Moreover this analysis allowed us to infer molecular mechanisms which occurred during evolution resulting in PTP gene diversity observable today.

In CcBV circle 1 and its potential orthologs in other species, 6 orthologous PTP genes are found in the three species. The genes P Q and L group together in a monophyetic group (Figure 4). Each gene is found to group with an ortholog in GiBV segment F (genes 3, 2 and 4) or in CvBV segment 28 (gene 4, 3, 6), suggesting the duplications that led to circle 1 copy occurred before the divergence between *Glyptapanteles* and *Cotesia* bracoviruses. After the split between *Cotesia* and *Glyptapanteles* genus, GiBV PTP9 ancestral form was lost and the orthologous PTP 2 and D genes were acquired by the *Cotesia* bracoviruses. CcBV lost the orthologous genes of CvBV PTP 8 and 9 but acquired PTP K which is closely related to PTP L and thus most probably results from a recent tandem duplication event.

The CvBV segment 36 and 27 and the CcBV circle 17 and 10 were shown to be orthologous between the two species respectively, and within species each segment pair are paralogues that have evolved from duplication. Indeed, CvBV segment 36 and 27 correspond to the CcBV circle 17 and 10 respectively and moreover CvBV segment 36 is a near replicate of segment 27 and CcBV circle 10 is a near replicate of circle 27. These replicates were shown to be tandemly associated in the proviral form of CcBV (Bezier, unpublished). This pattern of circles duplicated in tandem was probably produced by segmental duplications. After this duplication the orthologous genes PTP 21 and Y and the orthologous genes PTP 34 and N were acquired in ancestral *Cotesia* bracovirus. Finally PTP 35 and PTP 19 were lost in CcBV circle 17 and circle 10 respectively.

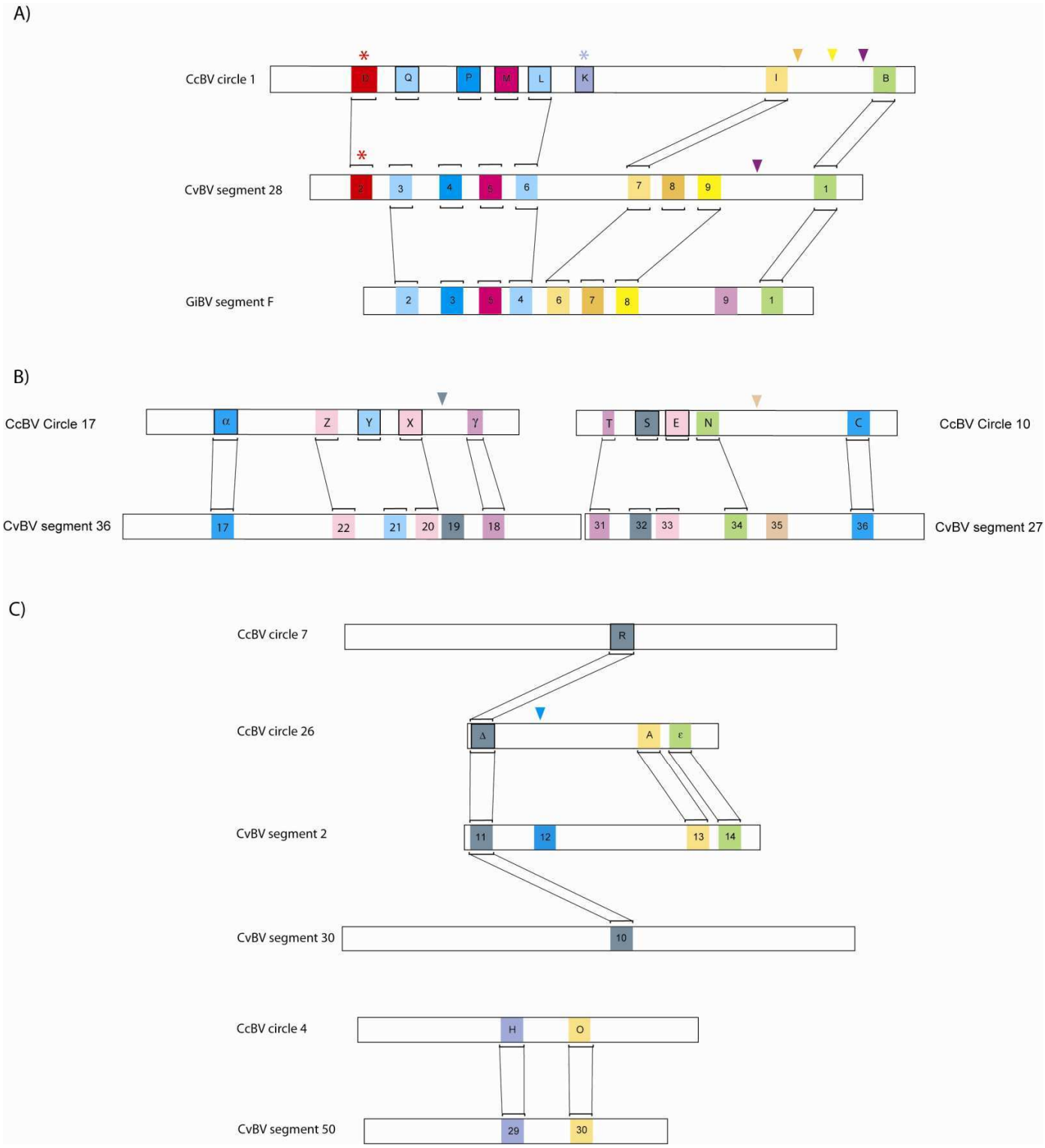


Figure 6 Orthologous genomic regions from CcBV, CvBV and GiBV.
PTP gene names correspond to those used in GenBank. (A) Orthologous genes between CcBV circle 1, CvBV segment 28 and GiBV segment F, (B) Orthologous genes between CcBV circle 17 and 10 and CvBV segment 36 and 27, (C) Orthologous genes between CcBV circle 7, 26 and 4 and CvBV segment 2, 30 and 50. Stars indicate gene acquisition and triangles gene loss. The same color is used between orthologous genes and between common events of gene acquisition or gene loss. Genes outlined are CcBV genes studied in PAML and phylogeny analyses.

Bracovirus genome comparisons allowed us to pin point two molecular mechanisms at the origin of PTP families. First, segmental and tandem gene duplications which either occurred in ancestral lineages or more recently in specific bracovirus lineages. Second, loss and acquisition of genes which occurred between genus but also within bracovirus species. Extension of the PTP gene family can also be explained through dispersed duplications. Indeed, according to PTP phylogeny PTP M, H, S, E, X and PTP R and Δ (and their homologs in CvBV) respectively form monophyletic clades which implies each set of genes shares a common ancestor. However, these PTP genes are found in different genome segments, for example R and Δ genes are found in circles 7 and 26 respectively and PTP M genes related to PTP E and X are found in circle 1 whereas PTP E and X genes are found on circles 10 and 17.

As a conclusion, duplications appear as a major molecular mechanism involved in PTP diversification including gene and segmental duplication as well as tandem and dispersed duplications. These duplications occurred at different times during PTP diversification after or before wasp divergence and were acompagnied by gene loss particularly in CcBV.

Episodes of positive selection during PTP evolution

To determine if selection pressures are acting on PTP gene evolution we measured branch specific selection in the 3 PTP clusters independently. Figure 7 shows phylogenetic relationships obtained from Bayesian analysis along with branch specific selection obtained in PAML and represented by branch width. MP (maximum parcimony) and BI (Bayesian inference) majority rule trees were generally congruent and when differences occurred, the bayesian tree was chosen to provide phylogenetic hypothesis for selective pressure inferences. Each of the 3 PTP gene family clusters are shown to diverge under varying selection pressures depending on the branch. Indeed the comparison of the M0 and M0b models by LRT indicates that a model in which branches evolve under different selective pressures (M0b) explains better the evolution of the PTP gene family rather than a model which does not (M0) (Table 2). These analyses enabled us to determine and to define the selective pressures which governed PTP gene evolution before and after duplication events.

For the PQLKYCα cluster, M0b model fit significantly better our data compared to M0 model ($p<0.05$). Branches supporting PTP genes of the PQ and KLCα clades were found to be highly selected ($\omega>>10$), suggesting that natural selection acted after a duplication event. In contrast, after this initial duplication, branches supporting the PTP K, L and Cα on one hand and Q and P on the other hand were not followed by episode of positive selection but evolved

rather under purifying and nearly neutral evolution ($0.2<\omega<0.7$). A second episode of positive selection was visible within PTP Cα, P, Q and Y clades, this is particularly true for PTP C and α genes which arose from a segmental duplication, for which more than 40% of branches show $\omega>1$ (Table 3).

PTP RΔ cluster evolution was also better explained by the M0b evolutionary model. The ancestral copy of the PTP R cluster diverged under strong positive selection ($\omega>10$) in contrast to the ancestral copy of PTP Δ cluster ($\omega<0.03$). Positive selection was also detected within PTP R and Δ genes where 16.9 % and 2% of branches were observed to be positively selected ($\omega>2$) respectively.

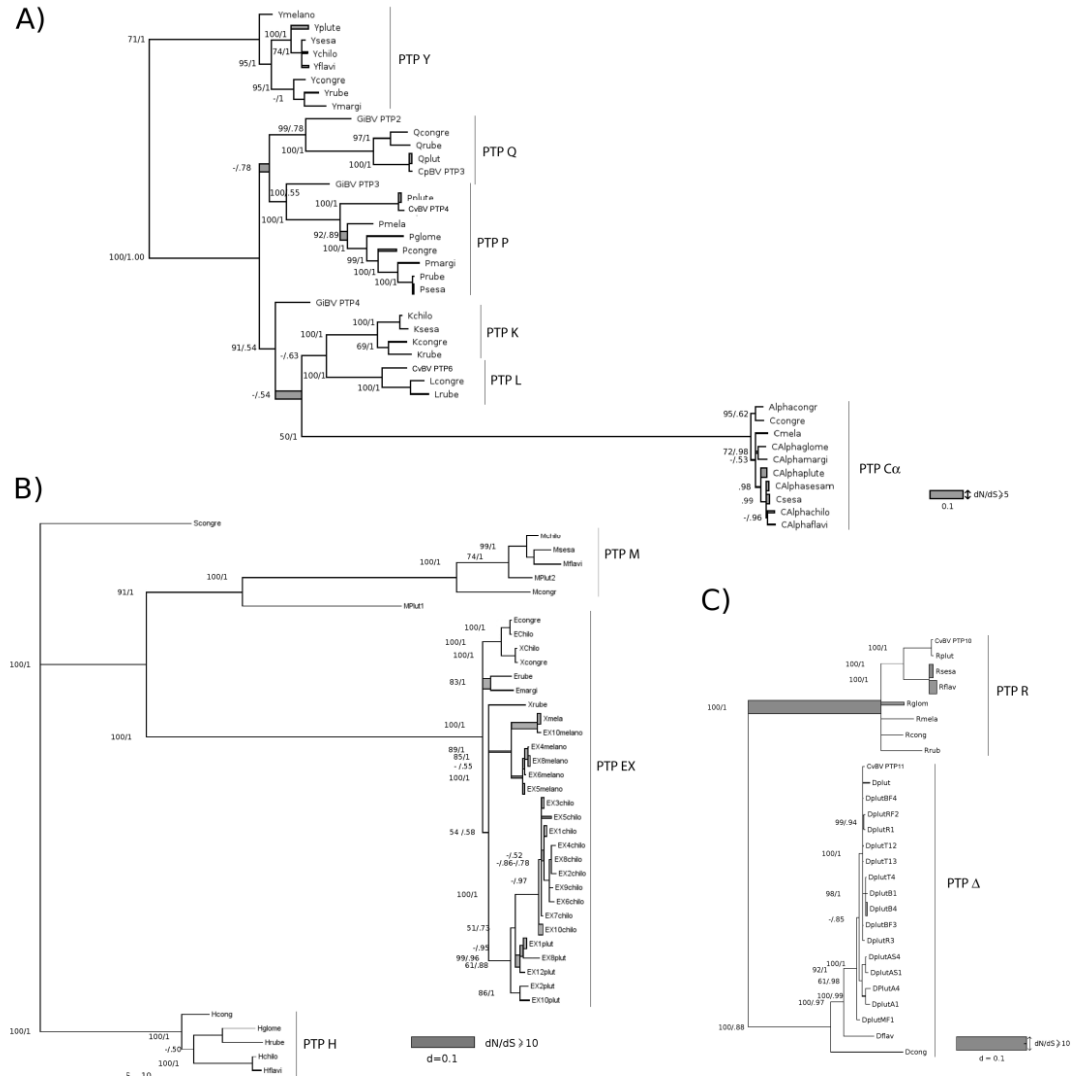


Figure 7 Unrooted phylogenetic trees from Bayesian inferences under the GTR+I+G substitution model of the three major PTP clades.
Posterior probabilities and non parametric bootstraps are indicated on the left of clades branchs. Branch widths are proportional to the ω ratio estimated under the branch-specific model in PAML (A) PTP PQLKYCα clade; (B) PTP MHSEX clade, (C) RΔ clade.

Finally, the MHSEX cluster also presented branches evolving under different selective pressures. In this case, branches supporting each PTP gene were not showed to have evolved under positive selection but were better explained by evolution under purifying selection ($\omega<0.5$). The M0b model is probably statistically significant thanks to the EX cluster, in which 26.5% of branches evolved under positive selection (Table3). In the EX cluster, numerous PTP forms are found and they appear to have diverged under positive selection between and within species ($1.7<\omega< 7$ and $\omega>10$). PTP E and X were shown to have arisen from a segmental duplication and positive selection would have acted to induce PTP E and X divergence.

To conclude, except for the MHSEX cluster where PTP clade ancestral lineages principally evolved under purifying selection, duplication events in other PTP clades were followed by episodes of positive selection. A second episode of positive selection acted more recently within PTP clades and is still active in PTP R and Δ and in PTP C and α, as well as in EX clade. Interestingly these three clades correspond to those in the data set analyzed that originated from whole segment duplication instead of tandem gene duplication. This suggests a particular role of segment duplication in evolutionary dynamic f PDVs.

Positive selection acted on specific residues

To determine whether the positive selection observed acted on specific amino acid residues, we measured site selection in the three PTP clusters. We tested site selection model (M8) and branch-site selection model (MA), as shown in table 2.

Evolutionary model comparisons showed that the clusters MHSEX and PKLKYCα are better explained by selective models. The model selected by AIC and LRT was M8 for MHSEX and PKLKYCα. In contrast, the RΔ cluster is better explained by the M8a model which is non selective. When site selection analysis was performed within Cα and EX PTP families, amino acids in position 83, 94 and 281 for PTP Cα and 38, 40, 75 for PTP EX were shown to be positively selected.

Using the M0/M0b test, we showed that some PTP lineages evolved under different selective pressures (previous section). When M0b proved to better explain PTP lineage evolution, we used the MA/MAnull model comparison to test wether particular sites evolved under positive selection. The MA model proved to better explain PQLKYCα, MHSEX and RΔ cluster evolution. Site analysis on the entire PQLKYCα clade revealed 7 amino acids under positive selection (at positions 87, 88, 120, 136, 158, 173 and 219) none of these sites are common to

positively selected sites identified specifically in the Cα clade. These results suggest that a first episode of positive selection induced divergence between PTP families and that a second episode acted on sites within PTP Cα. Within the MHSEX cluster, 9 sites were detected to be positively selected (positions 38, 40, 45, 46, 188, 227, 260) and 2 of these sites are common to positively selected sites identified in PTP EX (positions 38, 40, 75). These results suggest that PTP EX clade divergence was accompanied by changes of particular amino acids by positive selection, and a second episode of positive selection acted on particular amino acids within the PTP EX. Within the RΔ cluster, 5 sites are under positive selection (positions 165, 166, 177, 193, 207).

Table 2 Branch, Site and Branch-site model comparisons and position of positively selected sites.

Cluster	*AIC rank order*	*LRT*			*Sites selected (site number)*
		M8/M8a	*M0b/M0*	*MA/MAnull*	
PQLKYCα	**M8<M8a<MA<MAnull<M0b<M0**	**$P<10^{-4}$**	**P<0.05**	**P<0.0005**	**87, 88, 120, 136, 158, 173, 219**
P	M8a<M8< M0b<M0	NS	NS	-	
Q	M8a<M8<M0<M0b	NS	NS	-	
K	M8a<M8<M0b<M0	NS	NS	-	
L	M8<M8a<M0<M0b	NS	NS	-	
Y	M8a<M8<M0b<M0	NS	NS	-	
Cα	M8<M8a<M0B<M0	P<0.01	NS	-	83, 94, 281
MHSEX	**MA<M8<M8a<MAnull<M0b<M0**	**NS**	**$P<10^{-4}$**	**$P<10^{-5}$**	**38, 40, 45, 46, 188, 227, 260**
M	MA<MAnull<M8a<M8<M0b<M0	NS	$p<10^{-4}$	NS	
H	MA<M0b<M8a<MAnull<M8<M0	NS	P<0.05	NS	
EX	M8<M8a<M0<M0b	$P<10^{-5}$	NS	-	38, 40, 75
RΔ	**MA<MAnul<M8A<M8<M0b<M0**	**NS**	**$P<10^{-2}$**	**P<0.05**	**165, 166, 177, 193, 207**
R	MA<MAnull<M0b<M8a<M8<M0	NS	P<0.01	NS	
Δ	M0b<M0<M8a<M8	NS	NS	-	

Model comparisons were performed using Akaike Information Criterion (AIC) between all models and Likelihood Ratio Test (LRT) between nested models. Model descriptions: Branch selection models: M0=one class of ω ratio and M0b=tree branches have different ω ratio. Site selection models (ω ratio varies according to two classes); M8a:beta distribution of ω_0 and $\omega_1=1$ and M8: beta distribution of ω_0 and $\omega_1>1$. Branch-site selection models (ω ratio varies among sites in specific lineages): MA: two classes of sites; $0<\omega_0<1$ and $\omega_1=1$ and MA: three classes of sites; $0<\omega_0<1$, $\omega_1=1$ and $\omega_2>1$.

Table 3 Percentage of branch length for 4 classes of ω ratio for the different genes.

PTP genes	% of clade branch length with :				
	ω<0.2	0.2<ω<0.5	0.5<ω<1	1<ω<2	ω>2
P	31.9%	30.1%	30.3%	0.0%	7.7%
Q	0.0%	98.6%	0.0%	0.0%	1.4%
L	0.0%	78.5%	21.5%	0.0%	0.0%
K	2.5%	12.8%	84.6%	0.0%	0.0%
Y	5.0%	68.7%	12.1%	14.1%	0.0%
M	8.7%	45.5%	41.37%	4.4%	0.0%
H	0.0%	75.0%	4.14%	0.0%	0.0%
EX	1.7%	53.6%	18.15%	14.0%	12.5%
R	52.9%	30.3%	0.00%	0.0%	16.9%
Δ	18.7%	30.0%	49.19%	0.0%	2.0%
Cα	0.0%	11.6%	44.35%	22.5%	21.6%

The ω ratio for each branch was estimated using branch specific codon substitution models in PAML.

In conclusion, positive selection is shown to have acted between PTP families of a same cluster to fix particular amino acids but also between and within closely related PTP families (PTP Cα and PTP EX) inducing divergence between PTP lineages from different species or within a same species.

Positively selected sites are different according to the PTP cluster which means that each cluster diverged by selection acting on different amino acids. Most selected sites are in the vicinity of PTP conserved domains, and two positively selected sites are within motif 8 and 9. The amino acid in position 207 is positively selected in the RΔ cluster; the aspartic acid commonly found in other PTPs is mutated to glutamic acid in PTP R of *C. glomerata* bracovirus. Mutations in this site are expected to modify PTP activity efficiency (Andersen et al., 2001). In the MHSEX cluster, one positively selected site is found in motif 9 which carries the catalytic site. Interestingly, by co-cristallography and Cα-regiovariation in human PTP 1B, two regions (α1/ β1 loop and α5-loop-α6) and four specific areas located in proximity to the active site were shown to be involved substrate specificity. (Andersen et al., 2001). In MHSEX we found that positively selected residues number 38, 40, 45 and 46 fall in the α1/β1loop region and in Cα, residue 281 is found

near α5-loop-α6 (see figure 5). In these areas, the combination of residues is unique and could consequently represent a region determining protein specificity.

Discussion

The PTP gene family is known to be the most diversified family found in polydnaviruses associated with Braconid wasps (Bailey and Eichler, 2006; Desjardins et al., 2007; Espagne et al., 2004; Webb et al., 2006). Our study focused essentially on PTP genes from Bracoviruses associated with *Cotesia* genus wasps, for which 27 genes have been found in CcBV. Furthermore, PTP genes are annotated in the PDV genome offering interesting support to understand molecular duplication mechanisms which conduced to this large gene family. Therefore, bracovirus PTPs are a particularly interesting gene family to study molecular and evolutionary mechanisms involved in duplications and to understand the role of these duplications in a biological context.

PTP duplication mechanisms

Three major mechanisms are suggested to be involved in PTP diversification; first, large genomic regions have been duplicated (segmental duplication) and secondly individual genes are duplicated either in tandem or dispersed. Segmental duplications are clearly involved in PTP diversification. Indeed CcBV circle 10 and 17 were shown to be linked in the wasp genome (Bezier, unpublished) and harbour 5 homologous PTP genes suggesting they arose from a segmental duplication. These processes have previously been proposed to play a critical role in primate evolution in creating new genes and shaping human genetic variation (Bailey and Eichler, 2006). They seem particularly important in stimulating evolutionary changes since most of the recently selected Bracovirus PTP lineages were located in clades that emerged from segment duplication.

Tandem duplications are thought to be the major mechanism for the creation of new genes and have been documented in several organisms (Fan et al., 2008; Ganko et al., 2007; Hoffmann et al., 2008; Hooper and Berg, 2003b). In bracoviruses this pattern was observed for PTP K, L, P, Q genes found in CcBV circle 1. According to PTP phylogeny we can suggest that these genes were produced after two rounds of duplications which occurred at different periods.

First, the PTP P, Q and L genes emerged from an ancient duplication shared by *Glyptapanteles* and *Cotesia* bracoviruses whereas PTP K which is the recent PTP L duplicate is only shared by some *Cotesia* Bracoviruses. These results emphasize that PTP tandem duplications constitute dynamic processes which appear to be lineage specific. Tandem and segmental duplications are expected to produce genes or genomic regions closely associated in the genome. However, dispersed duplications which produced closely related genes were also shown to be involved in PTP diversification. How could these particular duplication patterns arise? It has been proposed that dispersed duplications arise from RNA mediated retrotransposition. This process is allegedly mediated by retrotransposons and produces intronless genes. Bracovirus PTPs studied to date are all characterized by their lack of introns (Espagne et al., 2004; Gundersen-Rindal and Pedroni, 2006; Webb et al., 2006) but other traces of RNA mediated retrotransposition have not been investigated to date. These patterns of duplications are important to consider in regard to gene function because tandem, segmental and dispersed duplications have different consequences on gene regulation. Indeed, genes originating from tandem or segmental duplications tend to maintain a similar function to their parental copy due to their sharing the same regulatory elements (Arisue et al., 2007; Darbo et al., 2008; Ponce and Hartl, 2006). In contrast, dispersed copies are expected to develop different functions since they are separated from their original regulatory elements (Wang et al., 2002). Interestingly, Weber and colleagues showed in *Chelonus inanitus* Bracovirus that genes found in the same PDV segment are similarly expressed suggesting that genome segmentation plays a role in gene regulation (Beck et al., 2007; Weber et al., 2007). Moreover, several studies suggest that DNA circles are not produced at the same rate in CcBV (Provost et al., 2004). Therefore, gene dispersion following gene duplications would tend to change gene replication rate and gene regulation pattern. The different duplication processes which led to the expansion of the Bracovirus PTPs probably played an important role in PTP function diversification and consequently in wasp parasitism success. Moreover, Bracovirus PTP gene expansion was shown to be accompanied by gene loss, implying that PTP gene evolution was conducted by the "Birth and Death" model described by Nei and colleagues (Nei and Rooney, 2005). According to this model, genes arise continually by duplication and are lost by deletion or by mutational events. Therefore, some PTP ancestral lineages were lost while others were created by duplications and were transmitted in particular lineages. As it has been shown in primates or in *Drosophila*, gene expansion and contraction could explain important adaptive traits illuminating the physiological adaptations of their host species (Babushok et al., 2007; McBride, 2007). By studying bracovirus PTP genes, we showed that genome reorganisation occurred at a very fine evolutionary scale with gene acquisition and loss occurring between species and

pseudogenization was also observed for some PTPs in particular bracovirus species (data not shown). Braconid wasps associated with polydnaviruses have been shown to be a highly diversified group with a very narrow host range (Smith et al., 2008) and virus genome plasticity could be viewed as a powerful mechanism allowing wasp adaptive radiation. Indeed PTP gene expansion and more generally virus genome expansion may be a source of evolutionary innovations offering wasps dynamic adaptive properties.

How did PTP family divergence occur?

PTP genes underwent several duplication events which appear to be lineage specific. Gene duplication has been considered as the most important mechanism in creating new genes. It is therefore essential to understand evolutionary forces which could explain how duplicated copies evolve new functions.

Our analysis emphasizes that PTP gene families did not evolve under the same selective pressures and PTP evolution underwent two episodes of positive selection. The first episode of positive selection occurred after a duplication event and before the speciation processes. The second episode occurred between and within PTP families after speciation processes. Understanding evolutionary processes underlying divergence of duplicated copies is of major interest to determine how genes can be innovated and how new functions could appear. For classical models, duplications are selectively neutral and maintenance of duplicated copies depends on a beneficial mutation which appears randomly and increases in frequency in the population (Hugues, 1994; Ohno, 1970). In contrast both in the "escape from the adaptive conflict" (EAC) (Des Marais and Rausher, 2008) and in "the innovation, amplification, divergence" (IAD) (Bergthorsson et al., 2007) models, duplications are immediately advantageous allowing the increase of protein amount.

The fact that positive selection is shown to be involved in bracovirus PTP copy divergence is concordant with the second class of models. How could PTP function be improved and innovated after duplications? The EAC model assumes that a novel function arises in the ancestral gene and after duplication each copy is selected to improve ancestral or novel function. For the IAD model, the novel function existed as minor activity in the ancestral gene and its duplication offered multiple targets to improve this function. The particular characteristics of polydnavirus life cycle may favour the EAC model. The virus is transmitted by the wasp but expressed by another organism. Therefore the wasp does not pay directly the cost of gene expression. This low cost of functional gene duplication may favour acquisition of new function.

During the period of gene change, the gene may escape from the adaptive conflict, and not be lost due to the limited cost payed by the wasp. Some Bracovirus PTP genes play an important role in host immune alteration particularly by modulating PTP cell activity in hemocytes (Pruijssers and Strand, 2007). Interestingly, one particular class of PTP have been identified for not carrying PTP activity, they were proved to reduce PTP cells activity probably through competition with host PTPs (Ibrahim and Kim, 2008; Pruijssers and Strand, 2007). Furthermore, it has been shown that PTPs are differentially expressed in the course of parasitism suggesting they performed different functions (Gundersen-Rindal and Pedroni, 2006; Ibrahim et al., 2007; Pruijssers and Strand, 2007). Interestingly, in a baculovirus infecting lepidopteran, PTP play a major role in host behaviour manipulation. They were shown to enhance locomotory activity and thus increase baculovirus transmission (Kamita et al., 2004). Nevertheless we are still unable to determine the precise role of each bracovirus PTP and correlate PTP evolution with functional innovations. For that we should link the different functions of PTPs with mutational events.

Our analyses identified particular amino acids targeted in the different PTP families to evolve under positive selection. We showed that selected amino acids were different in the three PTP clusters and have been fixed within PTP families. One of these mutations was shown to occur in a conserved motif known to be important for PTP efficiency (Andersen et al., 2001). Furthermore, some amino acids were found in regions shown to be involved in PTP specificity. PTPs are known to be highly specific proteins (Andersen 2001), these residues may play a role in PTP specificity and they could be therefore related to a particular function.

The challenge will now be to determine how PTP function has been innovated through duplications by studying the role and regulation of PTPs in relation with mutational events occurring after duplications.

Conclusion

The virus-wasp association is the only mutualism involving a virus known so far. Viruses confer to wasps new essential functions ensuring wasp parasitism success. Our study reflects the dynamic evolution of virus genes undergoing multiple gene duplication events and several episodes of natural selection. In other viruses, genomes are normally limited to a few genes, in our system gene expansion is a common process observed for most virulence factors and may be related to the essential role played by the virus in wasp parasitism success. Gene duplications are

likely to have offered new sources of innovation allowing wasps to colonize new host environments.

Acknowledgment

Cindy Ménoret and Carole Labrousse are gratefully acknowledged for taking care of the insects. We are grateful to Jérôme Lesobre for help in processing sequence reactions.

This work was funded by an EC grant (QLK3-CT-2001-01586 "Bioinsecticides from insect parasitoids") and by the ANR grant "EVPARASITOID". CS was supported by a PhD grant from the French ministry de l'Enseignement Supérieur.

IV-Viral cystatin evolution and 3D structure modelling: a case of directional selection acting on a viral protein involved in a host-parasitoid interaction.

Céline Serbielle, Shafinaz Chowdhury, Samuel Pichon, Stéphane Dupas, Jérôme Lesobre, Enrico O. Purisima, Jean-Michel Drezen and Elisabeth Huguet

Abstract

In pathogens, certain genes encoding proteins that directly interact with host defences coevolve with their host and are subject to positive selection. In the lepidopteran host-wasp parasitoid system, one of the most original strategies developed by the wasps to defeat host defences is the injection of a symbiotic polydnavirus at the same time as the wasp eggs. The virus is essential for wasp parasitism success since viral gene expression alters the host immune system and development. As a wasp mutualist symbiont, the virus is expected to exhibit a reduction in genome complexity and evolve under wasp phyletic constraints. However as a lepidopteran host pathogenic symbiont, the virus is likely undergoing strong selective pressures for the acquisition of new functions by gene acquisition or duplication.

To understand constraints imposed by this particular system on virus evolution, we studied a polydnavirus gene family encoding cyteine protease inhibitors of the cystatin superfamily.

We show that *cystatins* are the first bracovirus genes proven to be subject to strong positive selection within a host-parasitoid system. A generated 3-dimensional model of *Cotesia congregata* bracovirus cystatin 1 provides a powerful framework to position positively selected residues and reveal that they are concentrated in the vicinity of actives sites which directly interact with cysteine proteases. In addition, phylogenetic analyses reveal two different *cystatin* forms which evolved under different selective constraints and are characterized by independent adaptive duplication events.

Positive selection acts to maintain *cystatin* gene duplications and induces directional divergence presumably to ensure the presence of efficient and adapted cystatin forms. Directional selection has acted on key cystatin active sites, suggesting that cystatins coevolve with their host target. We can strongly suggest that cystatins constitute major virulence factors, as was already proposed in previous functional studies.

Introduction

In a host-parasite interaction the associated partners can influence each others evolution (Woolhouse et al., 2002). Molecular signatures of these complex evolutionary processes can be detected in the genomes of both organisms involved in such associations. Indeed, genes encoding pathogenicity factors directly involved in counteracting host defences or vice-versa are expected to be subject to positive selection, driven by an arms race between the two partners. Such coevolutionary processes have been well described in certain plant-pathogen interactions, where the host resistance genes and corresponding avirulence genes in the pathogen show evidence of positive selection (Dodds et al., 2006). In the *Xanthomonas*-pepper interaction, the Hrp pilus, a filamentous structure allowing bacteria to directly inject toxins into plant cells, also evolves under positive selection, thereby avoiding the plant defence surveillance system (Weber and Koebnik, 2006). Positive selection has also been detected in insect-pathogen interactions. For example, in *Drosophila*, RNA interference molecules involved in anti-viral defence are among the fastest evolving genes in this insect. This rapid evolution is due to strong positive selection, illustrating that the host pathogen arms race between RNA viruses and host antiviral RNAi genes is very active and significant in shaping RNAi function (Obbard et al., 2006).

We are interested in characterising the evolutionary processes underlying the insect host-parasite interactions between lepidopteran hosts and parasitoid wasps. In these systems, the endoparasitoid wasp larvae develop inside the lepidopteran host despite the hostile environment this habitat represents. One of the most original strategies developed by these wasps to defeat these defences is the injection of a symbiotic polydnavirus (PDV) at the same time as the wasp eggs (Beckage, 1998; Beckage and Alleyne, 1997; Stoltz et al., 1984). PDVs are divided in two genera, ichnoviruses and bracoviruses, which are associated with tens of thousands of endoparasitoid wasps belonging to two different families, Ichneumonidae and Braconidae, respectively (Fleming and Krell, 1993). PDVs are found in these wasps as proviruses which are transmitted vertically from one wasp generation to the next (Belle et al., 2002; Desjardins et al., 2007; Fleming and Summers, 1991; Gruber et al., 1996; Xu and Stoltz, 1991). Proviruses are excised from the wasp genome in the female ovaries and after replication, are injected in the host caterpillar as multiple double stranded DNA circles packaged in capsids. The virus does not replicate in the host caterpillar, but viral gene expression and protein production are essential for host immune and developmental alterations leading to successful development of the wasp larvae.

In this biological system, the virus plays key roles both in the mutualistic association with the wasp and in the parasitic association between the wasp and the caterpillar. PDVs are therefore likely to display molecular signatures which reflect constraints imposed both by the wasp and the host caterpillar. So far however, reports have principally concentrated on the influence of wasp evolution on viral genomes. Braconid wasps carrying PDV form a monophyletic lineage, suggesting an unique event of association between the wasp ancestor and the virus ancestor and a vertical transmission of the virus along wasp lineages (Whitfield, 2002). Accordingly, a phylogenetic study of *Cotesia* spp. and their associated viruses has shown a codivergence between the two mutualists (Whitfield and Asgari, 2003). Finally, recent data on the genome sequence of several PDVs has revealed that these viruses harbour a large number of eukaryotic genes likely picked up from the wasp genomes. These genes form multigene families that are good candidates to be involved in alteration of host caterpillar physiology (Espagne et al., 2005; Espagne et al., 2004; Provost et al., 2004; Strand and Pech, 1995b; Webb et al., 2006). Surprisingly, very few studies have focused on the potential influence of the host caterpillar on viral gene evolution despite the strong selective pressure this habitat represents. In this paper, we report the molecular evolution of a viral gene family considering both wasp evolution and the selective pressure imposed by the caterpillar hosts.

Our model system is the interaction between the braconid wasp *Cotesia congregata* and its lepidopteran host, the tobacco hornworm, *Manduca sexta*. The PDV associated with *C. congregata* (CcBracovirus, CcBV) has been sequenced, revealing the presence of numerous genes possibly involved in host deregulation (Espagne et al., 2004). Among these viral genes, one gene family encoding cystatins constitutes an interesting candidate system to study the influence of the host-parasitoid association at the viral molecular level. Cystatins are tightly binding reversible inhibitors of papain-like cysteine proteases, and are widespread in plants and animals (Rawlings et al., 2004). They are characterized by three conserved domains forming the site of interaction with C1 cysteine proteases: an N-terminal glycine, a glutamine-X-valine-X-glycine motif and a C-terminal proline-tryptophane amino acid pair (Bode et al., 1988; Stubbs et al., 1990). Cystatins and their target proteases have often been shown to be involved in host-parasite interactions with cystatins either playing the role of defence molecules or virulence factors. For example, in parasitic nematodes, cystatins are thought to play a key role in controlling the host immune response (Dainichi et al., 2001; Maizels et al., 2001; Schierack et al., 2003). Remarkably, plant cystatins acting as defence proteins have been shown to evolve under strong positive selection in response to cysteine proteases released by phytophagous insects. In this system, it has been

suggested that plant cystatins and insect cysteine proteases are involved in a coevolutionary process (Kiggundu et al., 2006).

CcBV *cystatins* constitute the first description of *cystatin* genes in a virus and are organized in a multigene family, composed of three genes present on the same circle (Espagne et al., 2005; Espagne et al., 2004). To date, there is no evidence of *cystatin* genes in *Microplitis demolitor* bracovirus (MdBV) which has been fully sequenced (Webb et al., 2006) and they have only been identified in one other polydnavirus (GiBV) from the braconid wasp *Glyptapanteles indiensis* (Desjardins et al., 2007). Both genomic and physiological features of cystatins suggest that these viral proteins could play an important role in the host-parasite association. First, the genomic organisation in a multigene family could be indicative of selective pressures acting on these genes. Indeed, Francino (2005) (Francino, 2005) suggested that gene duplications that can lead to an increase in protein dosage are favored by selective pressures. Secondly, *cystatin* genes are expressed rapidly and at an extremely high level during parasitism. This early and prolonged expression could be indicative of a role of cystatins in the early steps of host physiological disruption, as well as in the maintenance of this perturbed state. Finally a recombinant viral cystatin (Cystatin 1) was shown to be a functional and specific cysteine protease inhibitor (Espagne et al., 2005).

In this study we checked for molecular signatures associated with positive selection that may act on the viral *cystatin* gene family. We demonstrate strong and lineage specific adaptive evolution acting on these genes. Using homology modelling and molecular dynamics simulation techniques we obtained the three dimension (3D) structure of CcBV cystatin 1. The predicted model of the 3D structure of CcBV cystatin provides a framework to position the positively selected residues, and reveals that these are situated in key sites which are important for the interaction with target proteases. This particular selection, which is probably imposed by host defences, emphasizes the potential role of cystatins as pathogenic factors and suggests that cystatins coevolve with host cysteine proteases.

Methods

Wasp specimens

Cystatin genes were isolated from nine viruses associated with the following *Cotesia* species: *C. congregata, C. flavipes, C. chilonis, C. melanoscela, C. vestalis, C. rubecula, C. sesamiae, C. kariyai and C. glomerata* (Table 4). These species provide a good representation of *Cotesia* species diversity based on the *Cotesia* phylogeny (Michel-Salzat and Whitfield, 2004).

Table 4 Wasp samples and primers used for cystatin gene amplification

Wasp species	Location	Collections	Primers	Species Abbreviations
C. congregata	Lab reared	Drezen, J.M (Fr)	Cyst15/Cyst93	*CcBV*
C. chilonis	Lab reared	Wiedenmann,. R (USA)	Cyst15/Cyst93	*CchBV*
C. flavipes	Kenya	Dupas, S (Fr)	Cyst15/Cyst93	*CfBV*
C. glomerata	Lab reared	Vet, L (NL)	Cyst15/Cyst93	CgBV
C. kariyai	Japan	Tanaka, T (J)	Cyst15/Cyst103	*CkBV*
C. melanoscela	France	Villemant, C (Fr)	Cyst15/Cyst93	*CmBV*
C. vestalis	Benin	Guilloux, T (Fr)	Cyst15/Cyst93	*CvBV*
C. rubecula	Lab reared	Smid, H (NL)	Cyst15/Cyst103	*CrBV*
C. sesamiae	Kenya	Dupas, S (Fr)	Cyst15/Cyst103	*CsBV*

DNA extraction, amplification, cloning and sequencing

DNA extractions were performed using the Chelex method from a whole individual wasp. Wasp tissues were disrupted in a 5% Chelex solution including proteinase K (0.12 mg/ml). Three primers for *cystatin* gene amplification were designed based on an alignment of the three *cystatin* genes from *C. congregata* bracovirus [EMBL: AJ632321] and one *cystatin* gene from *Glyptapanteles indiensis* bracovirus [genbank: AC191960]; one forward primer Cyst15 5'-ATGGGCAAGGAATATCGAGTG-3' and two reverse primers Cyst93 5'-GTAAGGACAGTTTTTATCTAG-3', Cyst103 5'-GTAAGGACGACTTTTATCTAG-3'. The amplified product is composed of 279 nucleotides and encodes a 93 amino acid sequence containing the first two conserved domains of cystatins. PCR amplification was performed in a 50 µl volume containing 1X Taq buffer, 3mM of $MgCl_2$, 2.5mM of dNTP, 0.3 µl Taq polymerase (Goldstar, Eurogentec) and 50 pmol of each primer. Goldstar polymerase displays a very good fidelity of one error every 5.10^{-5} bases. PCR conditions consisted of an initial denaturation step at 94°C for 2 min followed by 30 cycles of a denaturation step at 94°C for 45 s, annealing step at 45°C for 1 min and polymerization step at 72°C for 45 s and final elongation at 72°C for 10 min.

PCR products were cloned into the pDrive-cloning vector (Qiagen cloning kit). For each species, 12 positive clones were sequenced in order to isolate all the *cystatin* gene copies and to obtain a minimum of two identical clones per sequence. Only CcBV21L and CcBV21I correspond to unique sequences. However excluding these sequences from the data set does not change the results of the analysis on PSS. Cloned inserts were sequenced in both directions using the Big Dye^{R} Terminator v3.1 cycle sequencing kit and the sequenced products were analysed using a capillary DNA sequencer (ABI PRISM 3100).

Sequence analysis and phylogeny

Cystatin sequence obtained and sequences already available from viral genome sequencing (CcBVcyst1, CcBVcyst2 and CcBVcyst3) were aligned using ClustalW implemented in Bioedit version 5.06 (Hall, 1999). We estimated the intraspecies and interspecies *cystatin* gene divergence using MEGA ver3.1 (Kumar et al., 2004). Divergence was calculated by a pairwise distance under the Kimura 2-parameter substitution model.

Recombination can mislead phylogenetic estimation and positive selection analysis. In order to avoid this bias we tested the *cystatin* gene family for recombination using a Genetic Algorithm Recombination Detection (GARD) implemented in Hyphy (Pond et al., 2005).

The program MrModeltest ver2.2 (Posada and Crandall, 1998) was used to determine the appropriate model of DNA substitution by the hierachical likelihood ratio test (hLRTs). Phylogenetic trees were obtained by Maximum Likelihood (ML) using PHYML program (Guindon and Gascuel, 2003) and by Bayesian inference in Mr Bayes 3.12 (Ronquist and Huelsenbeck, 2003). Modeltest chose the Kimura 80 model with a gamma distribution of parameter shape α=0.7875, a transition/transversion ratio of 1.12 and a proportion of invariables sites equal to 0. These parameter estimations were used as initial parameter values for ML and Bayesian inference. The topology and branch length estimation by ML was repeated 1000 times and for Bayesian analysis we performed 1000000 generations until the standard deviation was below 0.01.

Positive selection among sites

All the analyses on the rate of protein evolution among taxa and tests of positive selection were conducted using the codeml program in the PAML package v3.14 (Yang, 1997). Pairwise estimates of the number of non synonymous substitutions per synonymous site (d_N) and the

number of synonymous sites (d_S) were calculated using maximum likelihood (Goldman and Yang, 1994).

To test for evidence of positively selected sites, we performed different models allowing evolutionary rates ($\omega=d_N/d_S$) to vary across codon sites (Models M0, M3, M8A and M8) (Yang, 2000). M0 (one ratio model) assumes that all branches in the phylogeny and all sites have the same ω. The model M3 classifies sites in the sequence into three discrete classes with ω estimated from the data (Yang, 2000). M8A assumes a β-distribution of d_N/d_S ratio constrained to lie between 0 and 1.0 and adds to the β-distribution a point mass at $\omega = 1$ (Swanson et al., 2003) whereas the selection model M8 permits one additional d_N/d_S ratio to be above 1. Nested models (i.e., M0 vs M3 and M8A vs M8) (non positively selected vs positively selected models) were compared using the likelihood ratio test: 2X the log-likelihood difference between the two models can be compared to a χ^2 distribution, with the number of degrees of freedom equal to the difference between the two models. Codon sites under positive selection were identified using the Bayes Empirical Bayes (BEB) calculation of posterior probability for site classes (Yang et al., 2005) that analyses the sites under positive selection identified by the selective models. The numbers of substitutions between *cystatin* genes were counted using "Codeml" program in the PAML package (Yang and Swanson, 2002), with the F1X4 model of codon frequencies. Four sequences containing a stop codon were eliminated from the analysis. Each analysis was repeated ten times with different initial ω values to avoid problems of multiple local optima.

Positive selection among lineages

To test for evidence of positive selection among sites but also among lineages we performed a branch-site analysis using the codeml program in the PAML package v3.14. In this analysis, the branches under positive selection are called "foreground" branches and all other branches are called "background" branches. Sites changing in the foreground lineage are permitted to have $\omega>1$. Yang and Nielsen (2002) (Yang and Nielsen, 2002) implemented two versions of branch-site models called MA and MB. In MA, ω_0 is estimated from the data under the constraint $0<\omega_0<1$; hence positive selection is permitted only in the foreground branch. This model is compared with model M1a. In MB ω_0 and ω_1 are free parameters. Thus some sites evolve by positive selection across the entire phylogeny, whereas other sites evolve by positive selection in the foreground branch only. MB is compared with M3. Parameters used to perform this analysis are the same as those used in the site analysis.

The branch-site analysis was used to gain information on the possibility of different evolutionary constraints in different lineages. A problem with this method is that it assumes an *a priori* hypothesis. Indeed we have to specify foreground and background lineages with no knowledge on lineage history or on the type of substitutions that occur.
To determine the precise lineage analysis we used a local codon model implemented in HyPhy (Pond et al., 2005) able to estimate non synonymous and synonymous substitutions per site for each branch. This analysis informs us about the kind of substitutions that occurred during *cystatin* lineages divergence.

In addition we used a naïve approach to detect branches specifically under positive selection in the tree. The basic principle of this method is to assign each branch of a phylogenetic tree to a particular ω class. Different models assigning branches into different ω classes were tested and compared using the Akaike information criterion (AIC_c). To search the space of possible models HyPhy employs a genetic algorithm (Ga) that measures the fitness of each model by its AIC_c score. Ga-branch analysis enables the assignment of lineages in a phylogeny to a fixed number of different classes of ω, thus allowing variable selection pressure without *a priori* specification of particular lineages. Ga-branch analysis as most molecular evolution programs is computationally challenging and imposed that the number of sample sequences be reduced to 25. We therefore removed from our sample all nearly identical sequences and pseudogenes. The evolutionary codon model used for this analysis was determined from the AnalyzecodonData implemented in the Hyphy package.

Structure prediction and model building

The available structure of chicken egg white cystatin (pdb code 1cew) and human cystatin D (pdb code 1roa) were used as templates (Alvarez-Fernandez et al., 2005; Bode et al., 1988). The sequence of mature CcBV cystatin1 shares 28% and 24% identity with chicken egg white cystatin [Swissprot: P81061] and human cystatin D [Swissprot: P28325], respectively. Despite the relatively modest level of sequence identity, a reasonable alignment could be made. In particular, the "wedge" region containing the conserved QxVxG motif could be readily aligned. CcBV cystatin 1 corresponds to a type 2 cystatin, which has two conserved disulfide bridges. For the inter-beta-strand disulfide bond, the sequence alignment and template structure place the Cys residues within disulfide bonding distance. The second pair of Cys residues in the initial model were too distant to form a disulfide bond, and had to be brought closer together through energy refinement . The homology modelling was carried out using the program COMPOSER in

SYBYL 7.3 (Tripos Inc, St Louis, MO) on residues 6 to 108 of the mature protein. Three structurally conserved regions (SCRs) were used to build an initial model of CcBV cystatin 1, with three deletions and no insertion relative to the template.

Molecular model refinement and molecular dynamics (MD) simulations

Structural refinement of the complex was done by stepwise energy minimization in Sybyl using the AMBER all atom force field (Cornel et al., 1995) to a gradient of 0.05 kcal/mol/Å. First, only the side chains of the SCRs were energy-minimized, followed by energy minimization of the entire structure. The energy-minimized model was then used as the starting point for molecular dynamics (MD) simulations using the AMBER ff03 force field in the AMBER 9 suite of programs (Case et al., 2005). The protein was solvated in a truncated octahedron TIP3P water box (Jorgensen et al., 1983). The distance between the wall of the box and the closest atom of the solute was 12.0 Å, and the closest distance between the solute and solvent atoms was 0.8 Å. Counterions (Cl^-) were added to maintain electroneutrality of the system. The solvated system was energy-minimized with harmonic restraints of 10 kcal mol^{-1} $Å^{-2}$ on all solute atoms, followed by heating from 100 K to 300 K over 25 ps in the canonical ensemble (NVT). Then, the solvent density was adjusted by running a 25 ps isothermal isobaric ensemble (NPT) simulation under 1 atm pressure. The harmonic restraints were then gradually reduced to zero with four rounds of 25 ps NPT simulations. After an additional 25 ps simulation, a 10 ns production NPT run was carried out with snapshots collected every 1 ps. For all simulations, a 2 fs time-step and 9 Å non-bonded cutoff were used. The Particle Mesh Ewald (PME) method was used to treat long-range electrostatic interactions (Darden et al., 1993), and bond lengths involving bonds to hydrogen atoms were constrained by SHAKE (Ryckaert et al., 1977).

Results

Cystatin genes from polydnaviruses associated with *Cotesia* species exhibit weak genetic divergence

To study the molecular evolution of viral *cystatin* genes, we isolated 48 sequences from polydnaviruses associated with nine *Cotesia* species, revealing that several *cystatin* forms exist in a same species. Accession numbers are provided in Additional File 1. The divergence of the third domain prevented amplification of this region, therefore the *cystatin* sequences isolated contain

only the first two interactive sites. It is extremely unlikely that endogenous wasp cystatins could be amplified by this approach given that polydnavirus cystatins show a low level of relatedness to insect cystatins, and are no more related to insect cystatins than to mammalian inhibitors (Espagne et al., 2005).

Four alleles isolated from *C. glomerata* correspond to a pseudogene with a stop codon situated in the same position for all sequences obtained. Genetic divergence estimated by pairwise distance, which gives the mean number of substitutions per site, ranges from 0.007 to 0.31; these weak values suggest that *cystatin* genes are very similar. Finally, GARD detected no evidence of recombination, allowing us to estimate phylogenies and test for positive selection on *cystatin* genes.

Cystatin **phylogeny shows two main** ***cystatin*** **forms**

Cystatin phylogeny was studied using Bayesian inference and maximum likelihood analysis. Both methods gave the same tree topology. The best tree obtained by maximum likelihood is presented in Figure 8 with bootstrap scores and posterior probabilities. The tree presented was unrooted because there is no suitable outgroup for this study. Phylogenetic analysis revealed the presence of two major cystatin forms supported by high bootstraps and posterior probabilities. The form A cystatins are constituted by CkBV, CmBV, CgBV, CvBV sequences and CfBVD and F, CsBV1, 2 and 3. The form B cystatin are constituted by CcBV, CchBV, CrBV sequences and CfBV 5, 7, 8 and 9, CsBV4 and 6. The form A, in which each clades is supported by high scores, matches the wasp phylogeny (Michel-Salzat and Whitfield, 2004). Indeed, in this case, *cystatin* sequences from a same species group together in the same way as in the wasp phylogeny (Michel-Salzat and Whitfield, 2004). In form B the organisation is different and does not match wasp phylogeny (Michel-Salzat and Whitfield, 2004). Indeed, we do not find a preferential association between sequences from the same wasp species, and the internal branches of this clade are not well supported. The second important difference concerns the branch length: form A *cystatins* exhibits higher overall branch length than form B, suggesting different rates of evolution for these two cystatin forms. This phylogeny strongly suggests the existence of two main ancestral *cystatin* gene forms which have evolved under different constraints to give form A and B. Indeed in form A *cystatins*, long branch lengths are exhibited and follow wasp speciation, as opposed to the form B *cystatins*, which exhibit shorter branch lengths and seem to evolve independently of the wasp speciation process.

Among *cystatin* sequences isolated from a same species some are likely to correspond to allelic forms like *CgBV cystatin* sequences whereas others seem to be different *cystatin* copies such as *CsBV*1, *CsBV*2 *and CsBV*3 (form A). *Cystatin* copies obtained from the CcBV genome sequencing project (*CcBV*cyst1, *CcBV*cyst2 and *CcBV*cyst3) are found in form B and therefore do not seem to have any orthologous sequences in *Cotesia melanoscela*, *Cotesia glomerata* and *Cotesia kariyai* bracoviruses. In form B *cystatins*, these three *cystatin* copies are not grouped together, indicating that duplications occurred before or at the same time as wasp speciation. On the contrary, *cystatin* copies or *cystatin* alleles in form A are grouped by wasp species, suggesting that duplications occurred after wasp speciation.

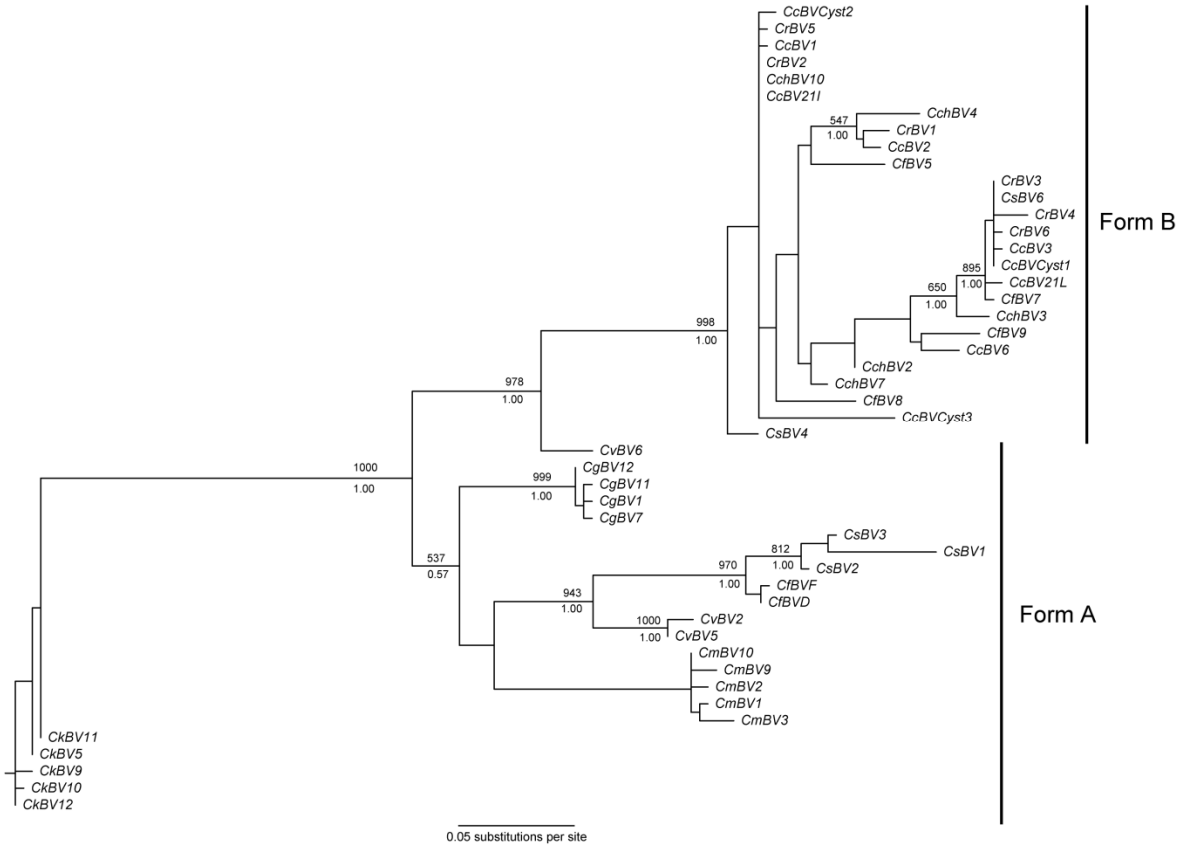


Figure 8 Cystatin gene tree obtained by maximum likelihood
Node supports are shown by bootstraps and by posterior probabilities from Bayesian inferences respectively above and below each branch. Bootstrap scores or posterior probabilities lower than 50% are not represented. Sequences were obtained from bracoviruses of *Cotesia congregata* (CcBV), *Cotesia flavipes* (CfBV), *Cotesia chilonis* (CchBV), *Cotesia melanoscela* (CmBV), *Cotesia vestalis* (CvBV), *Cotesia rubecula* (CrBV), *Cotesia sesamiae* (CsBV), *Cotesia karyai* (CkBV) *and Cotesia glomerata* (CgBV). Cystatin sequences from *CcBV* genome are noted *CcBV*cyst1, *CcBV*cyst2 and *CcBV*cyst3.

***Cystatin* genes evolve under positive selection**

In order to analyse protein evolution and test for positive selection in *cystatins*, ML of different substitution models were determined and compared using chi-squared statistics. Model M0 assumes that all sites have the same ω value whereas M3 distributes amino acids into three classes allowing sites to evolve under different evolutionary constraints. M8A model constrains amino acids to have ω values equal or under 1 whereas the M8 model adds a supplementary class of ω allowing sites to evolve under positive selection. LRTs indicated that selected models M3 and M8 fit the data better than M0 and M8A respectively with *P* values < 0.001 (Table 5). These results suggest firstly that all amino acids are not constrained by the same selective pressures and secondly that cystatin sequences, with an average ω value of 1.2 over all sites and branches, evolve under positive selection. A class-specific site selection analysis was performed to determine the heterogeneity of selection regimes relative to the amino acid position. This analysis indicates that more than 30% of all amino acids are under strong diversifying selection (Table 5).

Table 5 Positive selection analysis among sites and lineages of viral cystatins from *Cotesia* spp. parasitoid wasps

Site analysis						
Models	χ^2 value	df	P value for best model	Gobal ω	ω>1, parameters	PSS
M0 vs M3	125.00	6	< 0.001	1.31	ω=2.65, p= 0.310	26, 5*, 11**
M8A vs M8	19.57	1	< 0.001	1.21	ω=2.85, p= 0.289	26, 7*, 5**
Branch site analysis						
Models	χ^2 value	df	P value for best model		ω>1, parameters (foreground lineage)	
M1a versus MA	126.81	2	< 0.01		ω=11.93, p=0.232	
M3 versus MB	73.20	3	< 0.01		ω=12.04, p=0.233	

Notes: PSS is the number of positive selected sites; * corresponds to a posterior probability>95% of having ω>1 and ** corresponds to a posterior probability >99% of having ω>1.

Modelling by molecular dynamics simulations reveals that the overall folding of known cystatin structures are preserved in CcBV cystatin 1

We wanted to determine whether PDV cystatins adopt a similar 3D structure to chicken cystatin and human cystatins for which the 3D structures have been resolved by crystallography (Alvarez-Fernandez et al., 2005; Bode et al., 1988; Janowski et al., 2001; Stubbs et al., 1990). This constitutes an important prerequisite to be able to interpret the potential consequences of the position of the positively selected sites with respect to the function and the evolution of function of PDV cystatins.

In a previous study, a multiple sequence alignment of CcBV cystatin 1 was performed with insect, chicken, mouse and human cystatins (Espagne et al 2005). Although there is only a modest level of sequence identity among CcBV cystatin 1, human cystatin D and chicken egg white cystatin, a reasonable alignment could be found that permitted a homology model to be built. A 10 ns molecular dynamics simulation was carried out to check the stability of the modelled structure. The energy of the system levelled off after about 800 ps, indicating that an equilibrium state had been reached (data not shown). The overall structure was stable during the simulation. Visual inspection of the trajectory showed that the global fold remained essentially intact. The PROCHECK program (Laskowski et al., 1993) did not flag any conformational problems with the structure. Figure 9A shows a superposition of three average structures during three different time frames in the trajectory. We see that the structures of L1, L2 and L4 are very stable during the simulation. L3 shows somewhat greater structural variability.

The modelled structure preserves the overall fold of solved cystatin structures – a five stranded anti-parallel β-sheet wrapped around a five-turn α-helix (Figure 9). However, □1 maintains its beta strand conformation for only part of the MD simulation. The protease binding site shows a wedge shaped area formed by N-terminal residues (Glycine 6), the first hairpin loop L1 (QxVxG motif positions 50 to 54) and the second hairpin loop L2 (PW). The two conserved type 2 cystatin disulfide bonds are also preserved in this 3-D model of CcBV cystatin 1. Importantly, the 3D model shows that the three conserved domains in CcBV cystatin 1 form the typical tripartite 'wedge' which was shown in the crystal structure of human cystatin B in complex with papain to slot into the protease's active site (Stubbs et al., 1990). These domains therefore display a correct conformation in CcBV cystatin 1, consistent with previous data showing that cystatin 1 is a functional cysteine protease inhibitor (Espagne et al., 2004).

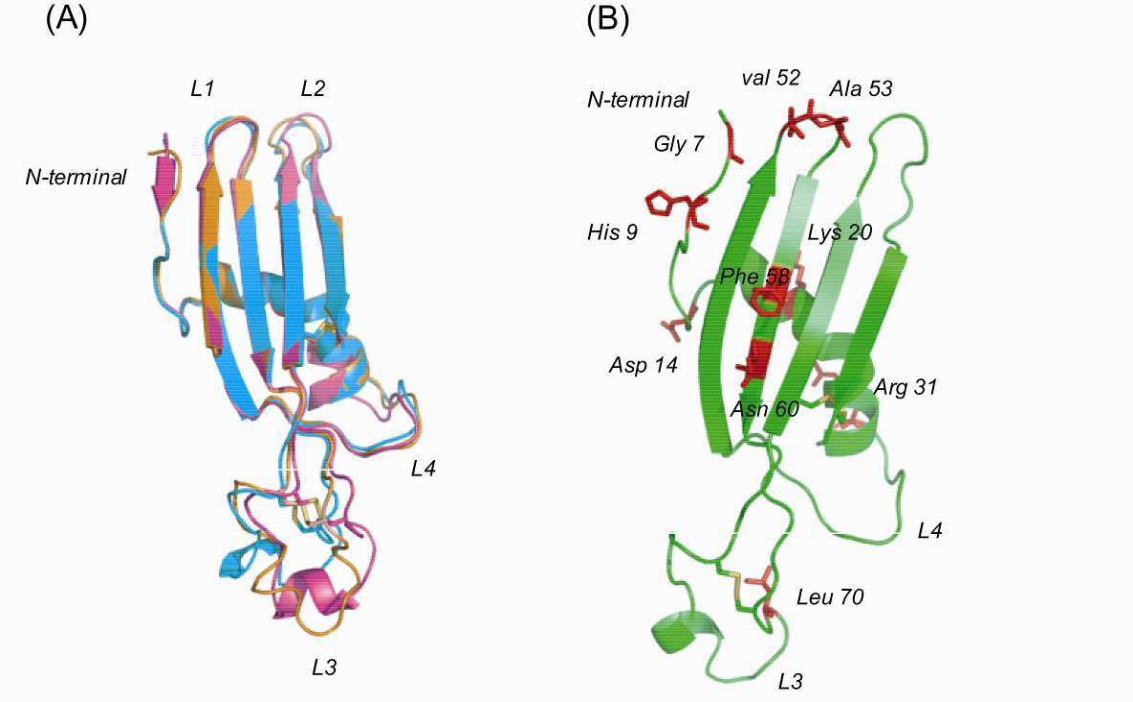


Figure 9 Molecular model of CcBV cystatin 1 obtained from the average structure
(A) The superimposed MD average structure of CcBV cystatin 1 orange (1-5ns), cyan (5-7ns) and purple (8-10ns) of 10ns MD simulation trajectory. (B) Positively selected residues (probability 95%) are represented as a red colour capped stick model on the secondary structure (green) of the final model of CcBV cystatin 1 average structure (1-10ns). Glycine in N-terminal and Valine and Alanine in the L1 are important for C1 protease binding. CcBV mature cystatin 1 amino acid numbering is used.

Most positively selected sites are situated in the vicinity of the cystatin active sites

Sites showing a significant probability (p>95%) of being positively selected in viral cystatins were mapped onto the primary sequence (Figure 10) and on the structural model of CcBV cystatin 1 (Figure 9B). Out of the 12 positively selected sites identified in the mature protein, four are situated in the N-terminal segment containing the conserved Glycine 6 residue (residues Lysine 5, Glycine 7, Histidine 9 and Aspartic acid 14) and two residues are within the first hairpin loop L1 containing the QxVxG motif (residues Valine 52, Alanine 53) (Figure 9B, Figure 10). Lysine 20 and Arginine 31 are located in the α-helix and Phenylalanine 58 and Asparagine 60 at the β3 sheet. Leucine 70 is located at loop 3 between β3 and β4 near the first disulfide bond.

Analysis of the viral cystatin protein alignment among the different wasp species revealed that out of the 12 positively selected sites, 8 (corresponding to Lysine 5, Glycine 7, Histidine 9, Aspartic Acid 14, Lysine 20, Arginine 31, Asparagine 60 and Leucine 70) undergo radical changes in biochemical properties which could induce changes in protein conformation and specificity

(see Additional File 2). For example, Lysine 5, which is a polar and hydrophilic amino acid, can be replaced in other viral cystatin lineages by a leucine which is a hydrophobic residue.

Two amino acids under strong positive selection are also found in the signal peptide. These residues are located in the central, commonly hydrophobic part of the signal peptide, and they do not undergo changes in hydrophobicity. Although selection on signal peptides has rarely been analysed it has already been described in virulence proteins (Liu et al., 2005) and it is thought that variations in the signal peptide could affect exportation of proteins (Byun-McKay and Geeta, 2007; Fujiwara and Asogawa, 2001). In our biological system, viral cystatins are secreted by the host secretory system, therefore we could speculate that the modification of the signal peptide composition could ensure more efficient secretion.

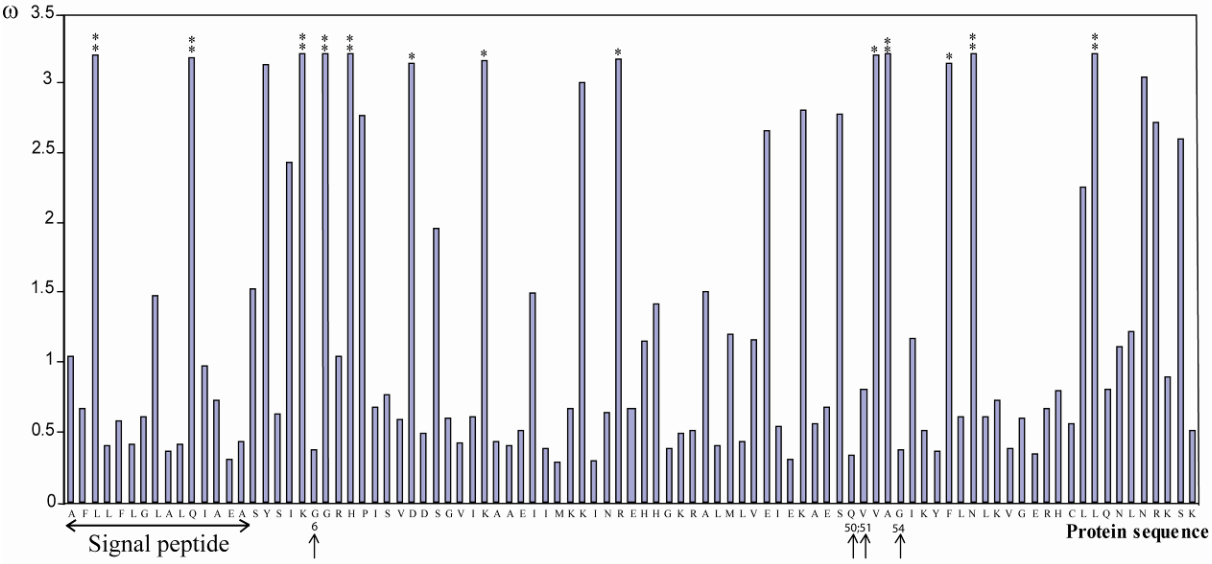


Figure 10 Graphic representation of variable selective pressures (ω) along the protein sequence.

The * indicates the posterior probability >95% of having ω>1 and ** indicates the posterior probability >99% of having ω>1. Conserved amino acids implicated in the interaction with target proteases are indicated by arrows and are numbered according to the mature protein.

Two main *cystatin* lineages show different evolutionary histories

To test for evidence of positive selection among lineages we performed a branch site analysis. In MA and MB models we assigned a $\omega \leq 1$ (ω_0) for form A *cystatins* (background branches) which is congruent with the wasp tree and should evolve under purifying selection and a $\omega > 1$ (ω_1) for form B *cystatins* (foreground branches) which is not congruent with wasp phylogeny and therefore should evolve under positive selection. LRTs indicate that MA and MB fit the data better than models M1a and M3 respectively, with p values <0.01. Furthermore, these analyses suggest that in foreground lineages about 23% of sites evolve under strong positive selection with ω values around 12 (Table 5). Branch-site analysis results therefore suggest that form A *cystatins* are mainly undergoing purifying selection, whereas form B *cystatins* are mainly evolving under positive selection.

Because this PAML analysis did not allow us to determine the nature of selective pressures acting on each branch, we constructed trees in which branch length represents the expected number of substitutions per codon. The tree in Figure 11A is based on nonsynonymous substitutions, whereas the tree in Figure 11B represents the expected number of synonymous substitutions in *cystatins*. These representations clearly showed a difference in the type of substitutions occurring in the two *cystatin* forms and suggested that divergence between *cystatin* sequences from the form A are particularly due to synonymous substitutions which occur principally in the internal branches, whereas divergence in form B is principally explained by nonsynonymous substitutions. A similar analysis conducted with a nuclear wasp gene (COI) did not reveal differences in synonymous and non synonymous substitutions between wasp species (data not shown) suggesting that the different evolutionary patterns observed above are specific to viral cystatins.

Figure 11 Cystatin sequence tree under local substitution model

(A) Tree scaled on expected number of nonsynonymous substitutions per site and (B) Tree scaled on expected number of synonymous substitutions per site.

To gain further insight into the nature of selective pressures acting on each branch we performed a Ga-branch analysis that confirmed that all lineages are not constrained by the same evolutionary forces. Ga-branch analysis selected a model with two classes of ω. In total, 49 % of branches are assigned to a ω of 0.6 and 51% to a ω class of 5.7 (Figure 12). Both types of branches are present in form A and B, however their position in the tree differs. In form A, positive selection occurs in terminal branches between intraspecies *cystatin* copies. This analysis emphasizes that divergence between *cystatin* copies from the same wasp species occurred by positive selection. Internal branches in form A *cystatins* are characterized by purifying selection, indicating that *cystatin* genes evolved under conservative selection during wasp speciation. A different pattern is observed in form B *cystatins*, where positive selection occurred preferentially in internal branches of the tree. Indeed positive selection occurred in the original branch and in almost all internal branches of this clade, thereby diluting the effect of wasp speciation on *cystatin* divergence. In conclusion, PDV *cystatin* divergence has been driven by positive selection, which has acted at different levels either before, during or after the wasp speciation process.

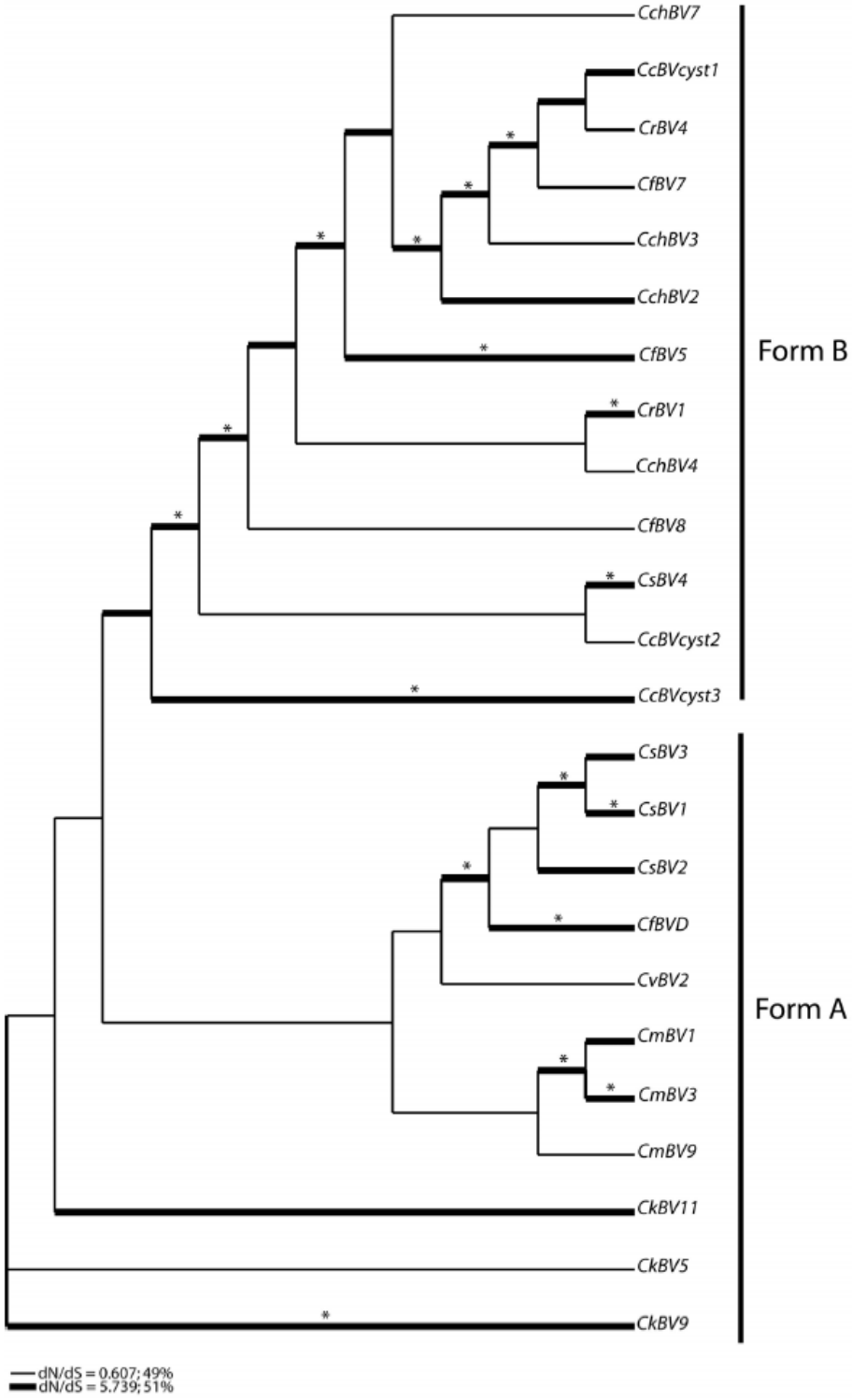


Figure 12 Branches under positive selection estimated according to the Ga-branch analysis.

Percentages for branch classes in the legend reflect the proportion of total tree length (measured in expected substitutions per site per time unit) and evolving under the corresponding value of dN/dS. The * indicates the posterior probabilities >95% of having ω>1.

Discussion

Cystatin genes constitute a young multigene family compared to the other *Cotesia* bracovirus genes

Cystatin genes appear to be unique compared to the other gene families found in the viruses associated with *Cotesia* genus. First *cystatin* divergence, which gives the mean number of substitutions per site, is very weak ranging from 0.007 to 0.31, whereas divergence between CcBV copies of other viral genes like protein tyrosine phosphatases (PTP) or IκB-like proteins range from 0.56 to 0.832 (Bézier et al., 2007). In contrast to PTP or IκB-like proteins, which are both widely distributed in the Bracoviruses carrying PDV (Webb et al., 2006), *cystatin* genes are so far restricted to *Glyptapanteles* and *Cotesia* (Desjardins et al., 2007; Espagne et al., 2004). Furthermore *G. indiensis cystatin* is found in a single copy, whereas three copies are found in *C. congregata* (Desjardins et al., 2007; Espagne et al., 2004), suggesting that the *C. congregata cystatin* gene family resulted from a recent duplication event. The weak divergence between *cystatin* lineages as well as their narrow phylogenetic distribution constitute evidence of the recent acquisition of *cystatin* genes by the bracovirus.

As a consequence, studying *cystatin* gene evolution might allow us to understand the preliminary evolutionary processes involved in the diversification of a young multigene family. The recent events of acquisition and duplication of *cystatin* genes might explain the lack of divergence between *cystatin* copies and our inability to distinguish orthologous and paralogous relationships between copies. For this reason in our analysis, all *cystatin* copies that might include orthologs and paralogs were analysed together.

Are cystatin genes codivergent with wasp species?

PDVs are integrated into wasp chromosomal DNA as a provirus which is inherited exclusively in a Mendelian fashion (Stoltz, 1990). There is no evidence that PDVs can be transferred horizontally between parasitoids and PDVs do not replicate in the host caterpillar. In view of this particular virus life-cycle we can hypothesize that PDV gene evolution is in part determined by evolutionary constraints acting on wasps, such as a phyletic constraints. Nevertheless, viral genes, which are likely to be involved in parasitism success, also have to adapt to caterpillar defences.

A study comparing wasp phylogeny of seven *Cotesia* species based on mitochondrial DNA and viral evolution using the CrV1 gene, has shown a perfect congruence between wasp and viral phylogenies (Whitfield and Asgari, 2003). In our study the *cystatin* gene tree also shows perfect codivergence between wasp and viral genes for some *cystatin* gene lineages. The evolution of these *cystatin* forms appears therefore to be constrained by wasp phylogeny and the molecular constraints acting on the wasp genome. However, in contrast to the results obtained using CrV1, not all *cystatin* lineages follow wasp evolution; instead, some *cystatin* genes are submitted to other constraints since their phylogeny does not match wasp phylogeny.

Cystatins are under strong selective pressure acting on key sites

The study of selective pressures acting on *cystatin* genes confirms that *cystatin* genes are not simply constrained by wasp evolutionary history. Indeed, we showed that *cystatin* gene evolution is driven by a strong positive selection. The global ω value of 1.2 obtained through analysis of viral *cystatins* is similar to the value obtained with plant cystatins (Kiggundu et al., 2006). Plant cystatins are involved in a plant-phytophagous interaction, but in that case cystatins play a role in defence against digestive cysteine proteases of herbivorous insects. Plant cystatins and their targets are thought to be involved in a coevolutionary process. Other examples of positive selection are also available with pathogen molecules. A previous study performed on an Ichnovirus protein involved in host immune inhibition has shown that positive selection was only detected at particular protein sites (Dupas et al., 2003b). Our study constitutes the first example of a major impact of positive selection in the evolution of a bracovirus protein.

The identification of the position of positively selected sites in PDV cystatins in the primary sequence and in the 3D-model revealed that 70 % of sites are situated within or proximal to the N-terminal segment harbouring the conserved Glycine and the first hairpin loop containing the QxVxG motif. These two domains, together with the C-terminal PW sequence, make up the « wedge » in the cystatin1 model, shown by crystallography in cystatin B and chicken cystatin to interact directly with the active-site cleft of target C1 proteases (Bode et al., 1988; Stubbs et al., 1990). These results suggest that diversifying selection could be acting on viral cystatins to modify the inhibitor's sites of interaction with host target proteases, which could translate into an increased or reduced affinity towards these enzymes. Interestingly, modifications in inhibitor affinity have been reported in engineered cystatin proteins carrying deletions or mutations in the N-terminal segment or the first hairpin loop (Abrahamson et al., 2003; Hall et al., 1995; Kiggundu et al., 2006; Machleidt et al., 1989). In chicken cystatin, the removal of the

residues preceding the conserved Glycine leads to a 5000 fold decrease in affinity towards papain (Machleidt et al., 1989). Furthermore, a site-directed mutagenesis approach used to pin-point which residues contribute the most to target enzyme affinity in human cystatin C revealed that the -1 residue (with respect to Glycine) is responsible for the major part of this affinity (Hall et al., 1995). In PDV cystatins it is noteworthy to stress that the equivalent site (corresponding to Lysine 5 preceding the conserved Glycine 6 in CcBV cystatin 1) is under positive selection. This suggests that PDV cystatins have evolved under diversifying selection possibly to produce inhibitors of varying affinity for caterpillar proteases, just as cystatin C laboratory engineered mutants have been developed that have discriminating affinities for mammalian cysteine proteases (Mason et al., 1998). We can predict that the other sites under positive selection in the N-terminal region of viral cystatins are also likely to influence the interaction with proteases. Indeed, comparison of positions of positively selected sites of PDV cystatins and plant cystatins revealed that 2 of these sites are in equivalent positions with respect to the conserved Glycine residue in both sets of inhibitors (positions -1, +3). Furthermore, in plant cystatins, independent mutations in these sites lead to variations in inhibitory activity towards papain and cathepsin B (Kiggundu et al., 2006) (Goulet et al., 2008).

Two positively selected sites have also been identified in the first hairpin loop of PDV cystatins including the central valine of the QxVxG motif. These sites, corresponding to Valine 52 and Alanine 53 in cystatin 1, are inside this region with one affecting the central Valine. However this central site is not absolutely conserved in all cystatins. In the chicken egg white cystatin the hairpin loop motif is QLVSG and an increase in binding affinity to cysteine proteinases was obtained when this motif was mutagenized to QVVAG (Auerswald et al., 1995) indicating that variation in central residues of this loop affects binding with target proteases.

In summary, the majority of positively selected sites identified in PDV cystatins are located in the vicinity of the two inhibitory sites analysed in this study. Furthermore, these sites affected by positive selection have been shown experimentally in other cystatins to be important for affinity with target proteases. Taken together these results suggest that positive selection is acting presumably to modulate viral cystatin affinity for caterpillar protease targets.

It will now be interesting to determine what could be the role of the positively selected sites which are more distant from the cystatin inhibitor sites (Lysine 20, Arginine 31, Phenylalanine 58, Asparagine 60 and Leucine 70 in cystatin1). Phenylalanine 58 and Asparagine 60 may still be influencing the L1 loop at position 50-54. Leucine 70 is located near the disulphide bond and variations in this position may affect the structure of the protein. These sites could also be unmasking new sites of interactions with proteases, indeed in chicken cystatin it

was suggested that others regions or sites of the protein could be important for the strong interaction with the cysteine protease cathepsin L (Auerswald et al., 1995).

Scenario for cystatin gene evolution

The strong selective pressure observed emphasizes the important role of cystatins in the host-parasitoid interaction. These results suggest that cystatins have to continuously evolve in order to adapt to their target in the host caterpillar. Given the potential pathogenic role of viral cystatins and also the probable involvement of cysteine proteases in insect immunity (Saito et al., 1992), these results can be interpreted by integrating cystatins in a coevolutionary context. Nevertheless, this diversifying evolutionary pattern could also be explained by wasp host switches and the subsequent necessity for cystatins to evolve rapidly to respond to new biochemical targets.

Our analysis reveals the existence of two viral cystatin forms which display different evolutionary patterns in regard to wasp evolution. In more classical non-obligate mutualist associations, horizontal gene transfer can explain incongruences between host and symbiont phylogenies. However, in this case, virus and wasp have a long and stable relationship since more than 100 MYA (Murphy et al., 2008) and artificial infection of wasps by PDV is not possible. Therefore we propose and our analyses strongly suggest that adaptive constraints have contributed to the different evolutionary patterns observed in the two *cystatin* forms.

Moreover, for both of these two forms duplication events occurred independently in the different *Cotesia* bracoviruses studied and are fixed by positive selection which is also responsible of the ensuing divergence of *cystatin* copies.

Interestingly, Francino (2005) (Francino, 2005) proposed in the "radiation adaptive model" that duplications are fixed to their selective advantage and that gene copies evolve under natural selection before new functions appear. This mode of evolution, particularly for functional genes, could be a response to specific environmental pressures such as new biochemical niches. Therefore, the particular evolution of the *cystatin* gene family could be a response to particular cystatin targets in a specific host-parasitoid system.

Conclusions

Unravelling the molecular evolution of proteins can lead to a better understanding of their function. For the first time in a host-parasite interaction system, we show that viral *cystatins*

are subject to strong positive selection. 3D Modelling of a viral cystatin revealed that most of the positively selected residues are in the vicinity of the inhibitory active sites, suggesting that adaptive selection acted to improve the inhibitory activity of viral cystatins. Furthermore two different *cystatin* forms have been identified, each of them evolving under different selective constraints probably imposed by different host cysteine proteases.

In order to better explain *cystatin* gene family evolution, we have now to consider the host range of each wasp species studied. For this purpose, studying the Melitaeini-*Cotesia* system appears clearly adapted since their ecology in terms of host range is well characterized (Kankare and Shaw, 2004). Such a study would precise the potential coevolutionary processes involved between viral *cystatins* and host cysteine proteases.

Authors' contributions

CS participated in the data acquisition and performed the sequence alignment, the phylogenetic and selection analysis and wrote the main draft of the manuscript. SC performed the 3-D structure modelling and participated in drafting the manuscript. SP was involved in the main acquisition of data and participated in the critical revision of the manuscript. SD participated in selection analysis and in intellectual revision of the manuscript. JL participated in experimental discussions and sequence acquisitions. EP was involved in the cystatin 3-D structure modelisation and in intellectual discussions. JMD was involved in obtaining funding, in project conception and in critical revision of the manuscript. EH was responsible of the project conception and of the experiments and was actively involved in drafting the manuscript. All authors have given their final approval of the version to be published.

Acknowledgments

We are extremely grateful to Cindy Ménoret and Carole Labrousse for insect rearing. We thank Franck Dedeine for critical reading of the manuscript and Gilles Lalmanach for fruitful discussions. We thank the three anonymous reviewers for corrections and helping to improve the manuscript. We sincerely thank Louise E.M Vet, Robert N. Wiedenmann, Claire Villemant, Thomas Guilloux, Tomoyuki Tanaka and Hans M. Smid for providing specimens for this study. CS was supported by a PhD grant from the French ministry de l'Enseignement Supérieur. Financial support was provided by the ANR grant EVPARASITOID.