In silico Analysis of Canine and Feline Parvovirus Evolution

Jennifer Hetzl

"Diplomarbeit"

submitted to the
Faculty of Natural Sciences and Mathematics
University of Vienna

in partial fulfilment of the requirements
for the degree

Magistra rerum naturalium
(Mag. rer. nat.)

Ich habe zur Kenntnis genommen, dass ich zur Drucklegung meiner Arbeit unter der Bezeichnung "Diplomarbeit" nur mit Bewilligung der Prüfungskommission berechtigt bin.

Ich erkläre weiters an Eides statt, dass ich meine Diplomarbeit nach den Grundsätzen für wissenschaftliche Abhandlungen selbständig ausgeführt habe und alle verwendeten Hilfsmittel, insbesondere die zugrunde gelegte Literatur, genannt habe.

Vienna, December 2004

To my family
for the patience
and to Regina
for the spirit

Abstract. Since the first appearance of the Canine Parvovirus (CPV) about 40 years ago, scientists from different disciplines have been seeking to answer the following questions: How and when has it emerged? Which other viral species has it emerged from? This single-stranded DNA virus is not only suitable for *in vivo* and *in vitro* experiments, but is also especially appropriate for *in silico* studies. In this thesis, we present the phylogenetic analyses of more than 100 CPV and closely related FPLV species, based on both the coding and effector sequences of the viral coat protein. We propose to use a simplified arithmetic model to calibrate the molecular clock for CPV and FPLV. We then show the results from the determination of the rate of nucleotide substitution based on the phylogeny obtained for the viral coat gene. The results using our model correlate well with previous estimates for the molecular rate. We are further able to date the fictive early ancestors of CPV and hence, to give a refined estimate of the CPV divergence time from the FPLV variant species. Together with the analysis of synonymous vs. non-synonymous substitutions, we are able to evaluate whether the evolution of CPV and FPLV are under different selective forces. We also evaluate the dynamics of epidemic vs. pandemic phases in the recent evolution of CPV antigenic subtypes. We show a correlation of different values for the molecular rate to phases of host range shift and adaptive evolution. Our results are supported by the results of a recent study from another group. Finally, we show the results from identifying taxon-specific residues in the viral coat protein. Mapping of those residues on the virion surface helps in understanding the specific behaviour of certain species and serves as an indicator for biological implications of replacement of specific residues in virus-cell interaction.

Zusammenfassung. Seit dem ersten Auftreten des Caninen Parvovirus (CPV) versuchen Wissenschaftler zu klären wie, wann und woraus es entstanden ist. Dieses einzelsträngige DNA-Virus ist ein beliebtes Objekt für *in vivo* und *in vitro* Experimente, aber auch besonders geeignet für *in silico* Studien. Wir präsentieren die phylogenetische Analyse von mehr als 100 CPV und nah verwandten FPLV Spezies basierend auf den codierenden und Effektorsequenzen des viralen Hüllproteins. Wir schlagen ein vereinfachtes arithmetisches Modell zur Eichung der molekularen Uhr für CPV und FPLV vor und zeigen die Ergebnisse der Ermittlung der molekularen Rate der Nukleotidsubstitution basierend auf der Phylogenie des viralen Strukturgens. Die Resultate, die wir mit unserem Modell erhalten haben, korrelieren gut mit vorherigen Schätzungen. Wir ausserdem in der Lage, die fiktiven Vorfahren von CPV zu datieren und können so eine verbesserte Schätzung des Zeitpunkts der CPV Abspaltung von den FPLV Spezies geben. Anhand der Analyse von synonymen und nichtsynonymen Substitutionen untersuchen wir, ob die Evolution von CPV und FPLV unter verschiedenem selektiven Druck erfolgt und bewerten die Dynamik von epidemischen und endemischen Phasen in der jüngeren Evolution CPV-antigenischer Subtypen. Wir zeigen eine Korrelation unterschiedlicher molekularer Raten mit der Erweiterung des Wirtsspektrums und evolutionärer Anpassung auf. Unsere hier vorgestellten Ergebnisse werden durch die Ergebnisse einer jüngsten Studie bestätigt. Schließlich zeigen wir die Resultate der Bestimmung von taxon-spezifischen Resten im viralen Hüllprotein. Die Lokalisierung jener Aminosäuren auf Oberfläche des Virions gibt Hinweise auf ein mögliches spezifisches Verhalten einzelner Spezies und dient als Indikator für mögliche biologische Implikationen des Aminosäureaustauchs bei der Virus-Zell-Interaktion.

Contents

List of Tables

List of Figures

Abbreviations

CPV	Canine Parvovirus
LCPV	Leopard Cat Parvovirus
FPLV	Feline Panleukopenia Virus
MEV	Mink Enteritis Virus
RPV	Rat Parvovirus
PPV	Porcine Parvovirus
AAV	Adeno-Associated Virus
ssDNA	single-stranded DNA
VP1 (VP2)	Viral Protein 1 (2)
NS1 (NS2)	Non-Structural Protein 1 (2)
TfR	Transferrin receptor
HRV	host range variant
NJ	Neighbor-Joining
MP	Maximum Parsimony
ML (LH)	Maximum Likelihood
K2P model	Kimura-2-Parameter model
TrN model	Tamura-Nei model
GTR model	General Time Reversible model
HKY model	Hasegawa-Kishino-Yano model
nt(s)	nucleotide(s)
aa	amino acid(s)

Chapter 1

Introduction

"It is better to know some of the questions
than all of the answers"
James Thurber

1.1 Motivation

The emergence of Canine Parvovirus some 40 years ago quickly gained the interest of scientists working in the fields of virology, veterinary medicine, and, recently, molecular evolution and bioinformatics. Although solved incompletely, the evolution of CPV is well documented by means of DNA sequences available from isolates of different outbreaks all over the globe, being sampled from its first appearance to date. It is the goal of this thesis to contribute to solving some of the still unanswered questions of CPV evolution at the level of molecular data. First, we will evaluate different hypotheses for CPV emergence proposed in previous studies and are looking for confirmation of either of these hypotheses. Thus, a major part of this work deals with reconstructing phylogenetic trees based on large data sets and using different evolution models, phylogenetic methods and parameters. We are tracking down the evolution of Canine Parvovirus to its

original ancestor and establish the corresponding "true" phylogenetic tree that represents its development during the past decades. We are also tracking temporally and geographically independent outbreaks by comparing sequence data from isolates of different origin. The analysis of phylogenetic trees might not reveal such geographical or historical clusters solely, but also batches of host range expansion followed by selective adaptive evolution. Ideally, we try to answer the question whether the shift from fast evolution[1] in epidemic phases to slow evolution in endemic phases is corresponding to particular nodes in the phylogenetic tree. Identifying the time—and consequently, the place—in the "true" CPV phylogenetic tree where a shift from faster to slower molecular evolutionary rates or vice versa has happened then allows to set up a species-specific molecular clock. We are then able to distinguish between isolates from endemic and epidemic phases of CPV based on DNA sequence and we might also deduce trends in the CPV phylogeny to learn more about the essential principles underlying viral evolution based on evolutionary rates. Analyses like this have been performed for other viruses like HIV [36] and Influenza [17, 23], adding to a better understanding of the emergence and adaption processes of new viral pathogens. It is a vital part to understand the evolution of viruses in order to fight the pathogenic potential in their hosts. Hence, recent viruses like CPV represent a unique chance to observe evolution at work.

Second, we emphasise the effect of single molecular substitutions on the viral capsid structure, and how these changes specify the interaction with the host. CPV itself has emerged through a host range shift, subsequently favouring the adaptation to a new host. Once we understand the

[1]The terms "fast" and "slow" evolution are used with respect to high and low values for the molecular rate of nucleotide substitution, respectively.

structural mechanisms involved in virus-cell interactions, the possibility of another host range shift will be discussed. Then we can take a closer look at the mechanisms responsible for infection with Canine Parvovirus and its emergence [18, 21, 41]. A summary of the current knowledge and recent findings about general and specific properties of Canine Parvovirus is given in the following part, including details about the taxonomy, genomic organisation, infectious mechanisms and evolution of Canine Parvoviruses.

1.2 Origin, Past & Recent Evolution

The Canine Parvovirus is considered a host range variant of Feline Panleukopenia Virus, sometimes referred to as Feline Distemper Virus. It is transmitted among dogs orally by encounter with contaminated feces. The virus requires host cell S-phase machinery; thus, it preferentially invades proliferating cells of the intestinal tissue, and efficiently replicates in these rapidly dividing cells. From the intestine, it can spread throughout the body. Infection with CPV leads to severe symptoms i.e. septicaemia, myocarditis, and gastroenteritis—or may even have lethal consequences for the host. Young individuals, i.e. puppies, and also hosts with a weak or depressed immune status or suffering from other intestinal infections, are particularly susceptible to CPV. CPV has spread worldwide quickly after its emergence in the late 1970s, and since then has been globally endemic in both wild and domestic dog populations. Vaccines have been developed swiftly to combat the disease, but it is still prevalent as shown e.g. by the number of new CPV infections per year in the US. The major reasons for CPV's persistence are the quick antigenic replacement of struc-

tural epitopes through antigenic drift and increasing infection efficiency and binding specificity through adaptive selective evolution. CPV still represents a serious threat to dog populations, and is also able to bind and infect cat cells *in vitro* and *in vivo* [33] and human cells in culture [45]. Thus, research has addressed the question of CPV emergence and its evolutionary tendencies. Initially, three theories for the emergence of CPV as a host range variant of Feline Panleukopenia Virus (FLPV) have been proposed and investigated. First, CPV may have arisen in cat or dog populations as a natural variant of FPLV. However, CPV is considered very unlikely a direct descendant of FPLV since no intermediate properties i.e. ancestral sequences of canine and feline capsid genes have been found in CPV and FPLV sequences yet. Indeed, phylogenetic analyses using tight bootstrapping parameters have revealed that the split between CPV and FPLV is highly significant. Second, the emergence of CPV from FPLV vaccines has been studied, but so far, no evidence for the emergence of CPV from a FPLV live virus vaccines has been found in neither serological nor *in silico* analyses [73]. The role of vaccines in the emergence of CPV remains unclear however, especially since FPLV vaccines have shown to fall into the CPV clade close to the "old" i.e. initial CPV variants. A third model proposed CPV to have emerged indirectly from an FPLV-CPV intermediate host, suggesting wild carnivores—e.g. foxes and wolves [74]. Steinel, Truyen *et al.* [64,65] have recently addressed this question in their studies, and with a Parvovirus sequence from an European Red Fox isolate have found a true intermediate between CPV and FPLV [64,72,75]. One part of this thesis will emphasise on looking for evidence for this hypothesis.

The original CPV type 2 (CPV-2) has been present worldwide in dog populations within a year after the host range shift to dogs. CPV-2, how-

Table 1.1 Specificity of FPLV and CPV antigenic (sub-)types for binding and infecting feline and canine cells.

	Felines	**Canines**	**Amino Acid Replacements**
FPLV	+++	-	-
CPV 2	-	+	6
CPV 2a	+	+++	3
CPV 2b	+	+++	1
CPV 2c	+	+++	1

ever, initially had lost its ability to bind and infect feline cells (see Table 1.1). Comparisons of the novel CPV with its ancestor FPLV using the VP2 gene sequence have revealed six amino acid differences between the species. These mutations have extended the host range from cat to dog populations. CPV type 2 was then soon replaced by its antigenic variants CPV-2a which acquired mutations affecting three additional residues, and later CPV-2b, differing from CPV-2a in one amino acid position only (see also Table 1.1). Three amino acid changes between CPV-2 and CPV-2a have increased the specificity of CPV for its new host, and also endowed the virus with the ability to infect feline cells. Within a few years, CPV has acquired ten specific amino acid differences to FPLV, which allow the high-specific identification of CPV types and discriminate those against their host range variant (HRV) FPLV [74]. Both CPV-2a and CPV-2b have re-gained the ability to bind and infect feline cells, causing the disease in either cats and dogs, whereas FPLV has remained specific for feline hosts solely [76]. In contrast to the original CPV-2, which is considered extinct to date, its antigenic types 2a and 2b can also co-exist in cat and dog populations without being replaced by each other. Since CPV-2a and 2b differ at a single position solely, we will have a closer look at the differences between

these antigenic types of CPV. Recently, occurrences of CPV antigenic type 2c have been reported in Asia [33]. We will include this novel antigenic type into phylogenetic analysis and examine its properties on nucleotide level. More details about the nucleotide and amino acid changes affecting the host range properties of CPV are given in Section 1.5.

1.3 Taxonomic Lineage

The genus Parvovirus is a subgroup of the *Parvoviridae*, a single-stranded DNA (ssDNA) virus family. Two subfamilies of *Parvoviridae* are known, *Parvoviridae* and *Densovirinae*, genera belonging to the second group will not be subject of this study. Parvovirus species have been identified and named from isolates, corresponding to the infectious potential for and the occurrence in specific animal hosts belonging to the mammalian orders *Carnivora* and *Rodentia*. Hence parvoviruses have been classified according to their host range: e.g. Canine, Porcine or Human Parvovirus. Figure 1.1 gives the partial taxonomic lineage of parvoviruses as given in the ICTVdb [1].

1.4 Genomic Organisation & Gene Products

Parvoviruses are assigned to the class of single-stranded DNA (ssDNA) viruses. Their unsegmented genome is approximately 5kb big and may equally occur as positive- (sense) or negative-sensed (antisense) DNA. The positive form however is prevalent in CPV populations. The viral genome encodes for two major proteins: the non-structural protein (NS), and the viral coat protein (VP) (see also Figure 1.4). Both genes are translated into

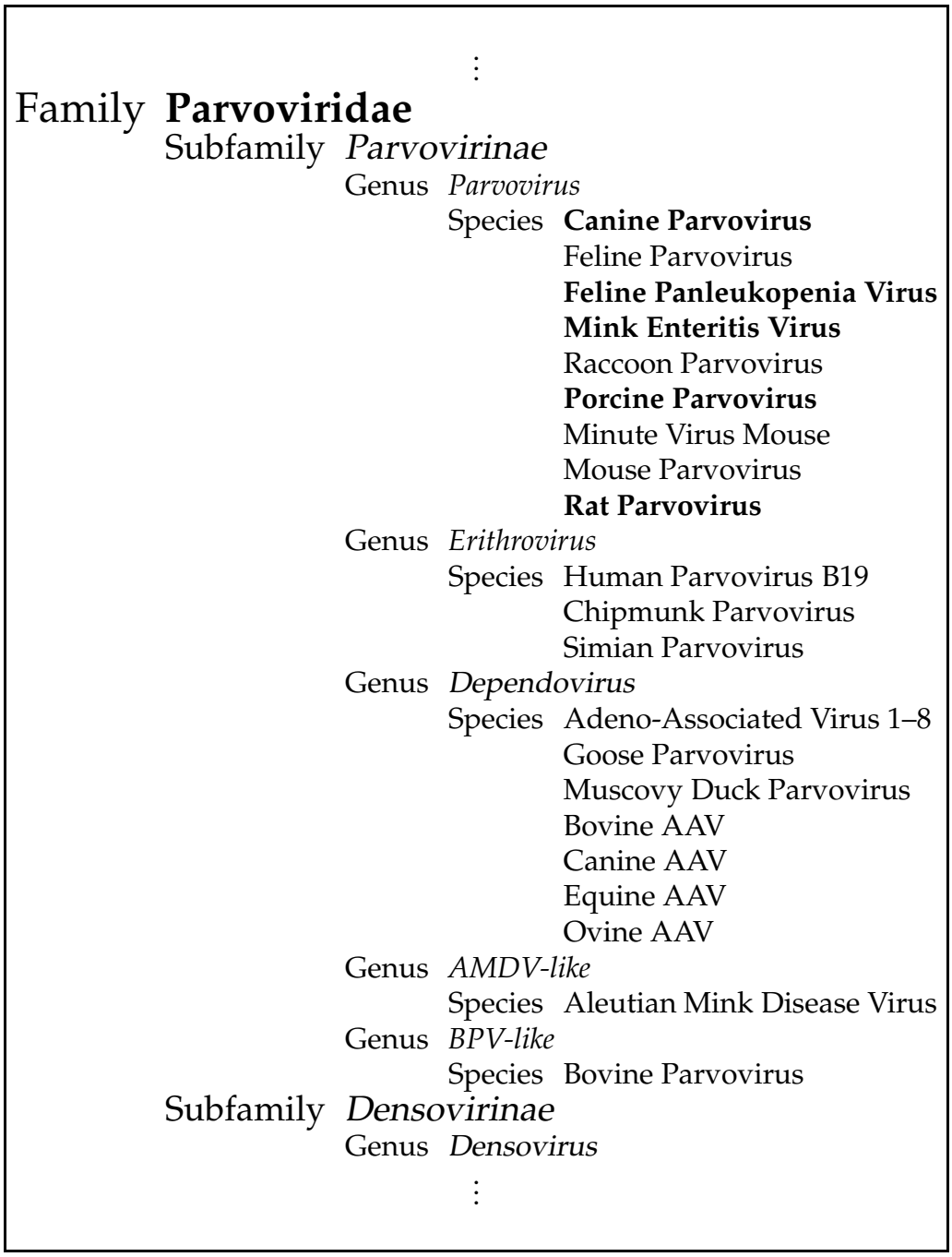
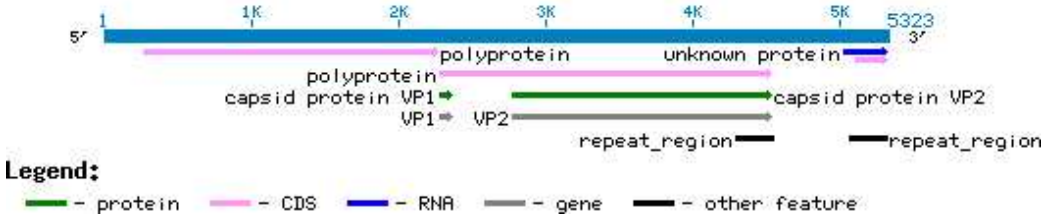


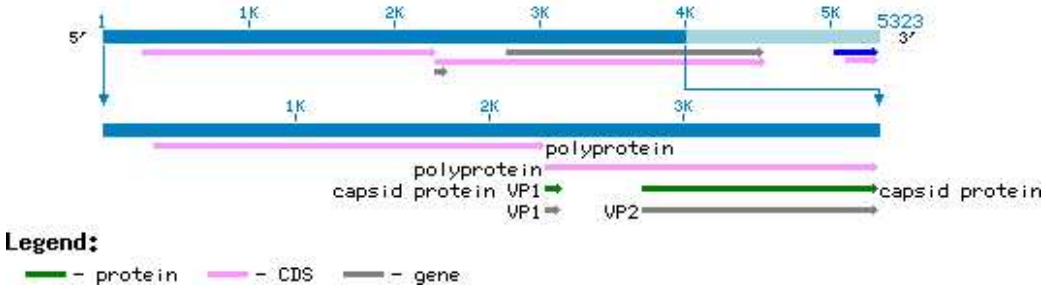
Figure 1.1 Abridged version of the taxonomic lineage of *Parvoviridae* and *Parvoviridae*, showing the positioning of Carnivore Parvovirus members used in this study (shown in bold face) within the taxonomy (taken from ICTVdb).

two proteins, enumerated 1 and 2, respectively. The NS1 protein is the major product of the NS gene, and exercises ATPase and DNA helicase activity *in vitro* and *in vivo*, hence it is involved in viral replication [83]. Similar proteins—with respect to the primary structure of the NS1 protein—have been described and are involved in virus replication in BoPV, SiPV, AMDV, Densoviruses, and other viruses [55]. The significantly shorter NS2 protein is an alternative splicing product of the NS mRNA sharing the N-terminal domain with NS1. Its function has not yet been described in detail, but it is very likely involved in viral replication, too [83].

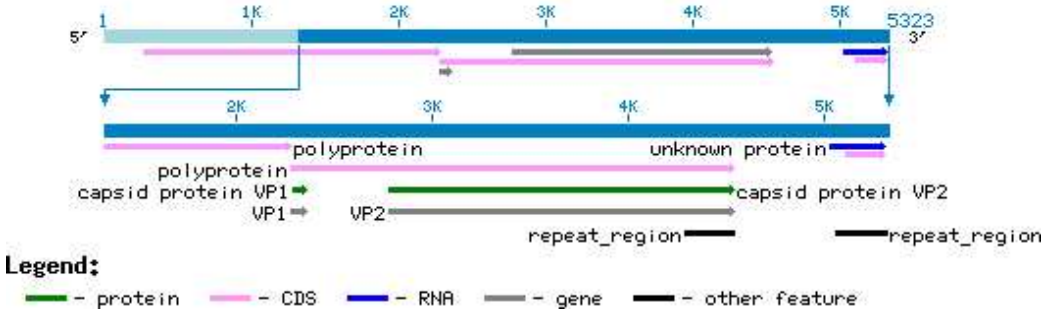
The viral capsid gene (VP) is equally translated into gene products VP1 and VP2 due to alternative splicing of the VP mRNA. The VP3 variant, however, is not a transcriptional or translational product itself but a post-transcriptionally proteolytic product of VP2. In this study, nucleotide and protein sequences of both VP genes are used to reconstruct phylogenetic trees, emphasising on the VP2 gene sequence. The VP2 gene spans a 1755 nucleotide region in the 3′ part of the genome, starting at position 2.783 with respect to the 5.323nt complete genome sequence of CPV "Norden" (PUBMED accession number `M19296`, [55]). The gene product is a 585 amino acid protein of 67kDa size, and together with VP1 builds up the viral capsid. The ratio of VP1 to VP2 monomers in the biological unit is 1:10. A detailed description of the biological properties and functions of VP2 *in vivo* and *in vitro* is given in the paragraph below. Figure 1.4 shows a schematics of the CPV genome (Canine Parvovirus strain "Norden"). It shows the location of the NS gene near the 5′ end, tightly followed by the VP gene near the 3′ end, with respect to the positive strand.



(a) 0–5.323kB



(b) 0-4k



(c) 1.321–5.323kB

Figure 1.2 Organisation of the CPV genome. **Top** Complete Genome **Middle** NS gene and translation products **Bottom** VP gene and translation products (taken from the NCBI [2]).

Table 1.2 Some general properties of Canine Parvovirus viral proteins 1 and 2. The values in parentheses indicate sequence lengths before splicing of the intron.

Property	**Viral Protein**	
	1	2
DNA sequence length [nt]	2184 (2256)	1755
Splicing of primary transcript	+	
Amino acid sequence length [aa]	728 (752)	585
Molecular weight [kDa]	83.3	67.3
Ratio of monomers in virion	1	10
Function: Host range/targeting	+	+
Function: DNA packaging	?	+

1.5 Morphology & Biological Properties

Sixty capsomer units together build the non-enveloped virion's icosahedral nucleocapsid with a diameter of about 255–260Å (Figure B.1). The full capsid is equally of VP1 and VP2 units wit a ratio of 1:10. The majority of VP2 monomers is modified, sometimes designated as VP3 units. VP1 and VP2 are the primary translation products of the VP gene and undergo alternative splicing, resulting in two proteins differing 26 residues in size and their N-terminal composition (Table 1.2). Assemblies of VP1 and VP2 can build the empty, i.e. non-DNA-containing viral coat. In addition, the role of VP1 in nuclear transport and its role in cell infection has been shown [45, 80]. Mature particles, however, consist of three different i.e. modified monomers. In contrast to the major proteins VP1 and VP2, VP3 is not a translational product itself but a post-translational proteolytic product of VP2. All monomers share the same composition of the C-terminal domain and differ in their N-terminal conformation solely.

The major structural difference between completely assembled empty

Table 1.3 Nucleotide and amino acid substitutions between FPLV, CPV-2 and its subtypes controlling the canine host range.

Protein Level		**DNA Level**		**Codon**	**Specificity**
Position	Substitution	Position	Substitution	Position	
80	K → R	239	A → G	2	CPV-2 †
87	M → L	259	A → C/T	1	2/2a
93	K → N	279	C/T → A/G	3	FPLV/CPV
103	V → A	308	T → C	2	CPV-2 †
300	A → V V → G	889	C → T T → G	1	2/2a
305	D → Y	913	G → T	1	2/2a
323	D → N	967	G → A	1	FPLV/CPV
426	N → D	1276	A → G	1	2a/2b
564	N → S	1691	A → G	2	CPV-2 †
568	A → G	1703	C → G	2	CPV-2 †
265	T → P	793	A → C	1	Italy
297	S → A	889	T → G	1	2a/b (EUR)

and full capsid particles has been correlated to VP2 [12]. Packed DNA binds to the inner surface of the capsid in the amino-terminal region of VP2 which is is exposed inwards in the empty capsids. Mature particles, however, expose the N-termini of VP2 monomers on the outer surface due to conformation changes, indicating VP2 plays a major role in DNA packaging [46]. The capsid surface plays a crucial role in host invasion and determines the host range [29, 47, 50]. Recently, the infection of host cells in culture and *in vivo* has been correlated to specific structural properties of the VP2 protein [62]. Indeed, very few residues of VP2 are responsible for targeting and effective interaction with host cells by binding to the transferrin receptor [19, 27, 44].

The ability to specifically bind the canine or feline transferrin receptor has been postulated a major driving force in the evolution of antigenic CPV types [28]. The mutation of very few structural determinant residues

interacting with the host cell transferrin receptor has led to the host range shift of CPV from feline to canine hosts, followed by adaption to its new host i.e. efficient binding of the canine receptor. The change of two amino acid residues only has been responsible to introduce the CPV-specific antigenic epitope [71]. Two epitopes are directly involved in either host range determination by interaction with the transferrin receptor and antigenicity through interaction with specific antibodies. In studies using antigenic variants and structural determinants of CPV VP2 the history of CPV and its subtypes was reconstructed by means of immunological and *in silico* assays. The role of specific residues in the capsid structure is currently studied using mutants [19, 27]. The interaction between the parvovirus VP2 and the transferrin receptor determines the specificity for feline and canine hosts. While Feline Parvoviruses have the ability to specifically bind to the feline transferrin receptor and infect feline cells in culture and *in vivo*, they bind to the canine TfR at low levels solely and cannot infect canine cells. The Canine Parvovirus, however, binds and infects both feline and canine cells due to the ability to specifically bind the canine TfR (gain of function) without loss of function to bind the feline TfR [76]. The canine host range correlates to three crucial amino acid substitutions of a total of 10 mutations resulting in different residues. These mutations affect three TfR-interacting epitopes. Residue 93 of the native VP2 chain has been changed from N to K, located in epitope 1, residue 323 replaces N with K in the nearby epitope 2, and a third epitope is affected by the substitution of N by K at position 300. While the epitopes formed by residues 93 and 323 interact directly with the canine TfR, residue 300 in the third epitope is involved in another, less well-defined yet crucial virus-cell interaction. Other residues may also play a role in virus-cell targeting; their

function is currently being studied using mutants. An overview of amino acid substitutions affecting the host range shift from feline to canine hosts is given in Table 1.3.

Chapter 2

A Classical Approach

"It is best to do all things systematically."
Hesiod

2.1 Phylogenetic Analysis

Sequence Selection and Retrieval. A total of 120 different canine and feline parvovirus VP2 gene sequences has been analysed in this large-scale phylogenetic analysis, considering sequence samples from all continents[1]. A significant proportion of 25 sequences, i.e. 21% of the total number of sequences, has been analysed in previous phylogenetic studies partly solely [26, 48, 74], and has been deposited neither to GENBANK nor EMBL sequence databases. Twenty-five sequences from German and US CPV, FPLV and MEV isolates were retrieved from the original sequencing files. These isolates have been collected between 1993 and 2000. Ninety-five sequences were identified and evaluated in a BLAST search using the complete genomic sequence of CPV strain "Norden" (GENBANK accession number `M19296`) as a reference and subsequently retrieved from GEN-

[1]The detailed listing of sequences used in this study is given in Table C.1 in the appendix section.

BANK [2]. To find support for the hypothesis of CPV emergence from modified live FPLV vaccines, we considered sequences from FPLV and CPV vaccines as well as live host isolates. The oldest CPV isolates date back to 1978, and the most recent ones have been isolated in 2003, i.e. the divergence time of CPV from FPLV is at least 36 years. Yet CPV has presumably emerged significantly earlier than 1978. By that time, it may have already spent a considerable time in a reservoir host animal before its first actually documented occurrence in domestic dogs in the late 1970s. Isolates have been collected in Canada, the USA, Australia, South Africa, Germany, France, Finland, Italy, Poland, the UK, the Soviet Union, China, India, Japan, Taiwan, and Thailand. Previous studies investigating the emergence and evolution of CPV in Brazil, Italy, and South Africa [7, 8, 51, 79] included several virus sequences not deposited to GENBANK or another nucleotide database and hence, could not be included into this study. The authors have shown that—with respect to CPV sequences publicly available—there is, however, no evidence for distinct lineages of CPV in Brazil and Italy, respectively [8, 51, 64, 79]. We have chosen a porcine and a rodent parvovirus sequence as appropriate outgroup representatives for phylogenetic analyses as suggested in previous studies [5, 13, 74, 82] and by BLAST hits.

Sequence retrieval from original files or GENBANK, and subsequent alignment suggested to consider a shortened VP2 gene sequence due to the high number of only partial gene sequences available. Hence, the first 162 nucleotides have been removed from the full-length 1755nt VP2 gene at the 5′ end of the sequence, corresponding to the first 54 amino acids at the N-terminus of the protein (see Figure 2.1). The shortened sequence,

[2] To ensure the reproducibility and for clarity reasons, any sequences deposited to sequence databases after March 2004 have not been considered for this thesis.

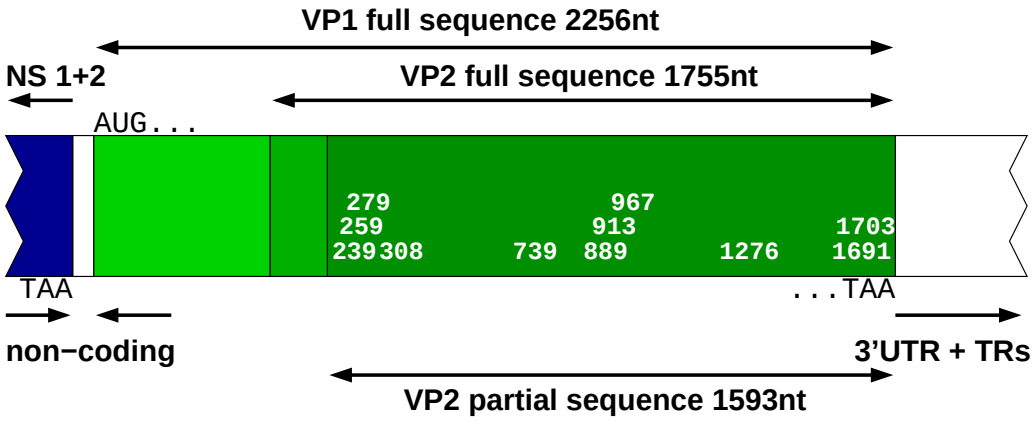


Figure 2.1 Target region of the VP2 gene sequence considered for phylogenetic analyses. For the genomic and translated sequences of the VP gene see appendix D.

consisting of 1593 nucleotides leading to a N-terminally truncated protein of 530aa length[3], includes all antigenically informative positions (Figure 2.1, [26,64]). We thereby reduce the number of phylogenetic uninformative positions and, more importantly, gaps drastically. So we expect to improve the quality of our results using the 5'-truncated sequences compared to those obtained using the full-length VP2 sequence.

Phylogeny Reconstruction. Nucleotide phylogenetic trees[4] were reconstructed using the UNIX-based PAUP* package (Version 4.0 Beta 1, [68]) and MEGA (Version 3.0 Test 7 [37, 38]) for Microsoft Windows. The appropriate nucleotide substitutions models and the corresponding optima parameters for usage with PAUP* were selected using MODELTEST (Version 3.06, [52]). We wanted to test whether the usage of evolution models, the setting of optional parameter values or the method for phylogeny re-

[3]The 531th codon is the non-sense stop-codon.

[4]Throughout this thesis, the term "phylogenetic trees" refers to the phylogeny of CPV reconstructed using the partial sequences of the surface antigen, i.e. the Viral Coat Protein 2 (VP2).

construction affects the results obtained. Hence the phylogeny was reconstructed with and without parameters for nucleotide γ-distribution correction, invariant sites, and unequal base frequencies. Neighbour-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML, or LH) methods were used for deference of phylogenetic trees, and a total of seven trees per method were compared (see Table 2.1). The most suitable evolution model for deference of nucleotide trees using PAUP* has been determined. The HKY+I+G model for nucleotide substitution [22] considers variable base frequencies and two different rates of nucleotide substitutions. It allows to specify parameters for the proportion invariant sites and for γ-correction of nucleotide distribution. We compared the results of HKY model trees to trees obtained using the GTR model considering six different rates [39, 56] for nucleotide substitutions, and "natural" trees obtained with plain distance, parsimony and likelihood methods using no optional parameter settings I+G. The quite sophisticated HKY model, the generic GTR model and the "reduced" model using default settings produced stable tree topologies with respect to NJ, MP, and LH methods. We observed similar topologies among the tripartite NJ and LH sets[5], with minor differences among the three NJ and LH trees, respectively (figures not shown). Since those trees' topologies did not vary significantly among NJ, MP and LH trees, the NJ tree based on the HKY+I+G model has further been considered as the reference tree. This approach allows to increase the significance of the results obtained enormously, analysing a total of 19 phylogenies. It also compensates for random choice of parameters and decreases the probability of possible biases due to intrinsic proper-

[5]With NJ and LH methods, PAUP* allowed to specify the usage of I+G parameters. We produced three "equal" trees using (1) I+G, (2) I solely, and (3) neither I nor G. The MP algorithm does not allow specification of neither I nor G, so we produced a single MP tree solely.

Table 2.1 The $e^{3^{3^3}}$ approach: Summary of different parameter settings for CPV DNA trees in PAUP*.

Model	**Reconstruction Method**		
	NJ	MP	LH
HKY	HKY+I+G HKY+I HKY w/o	MP w/o	HKY+I+G HKY+I HKY w/o
GTR	GTR+I+G GTR+I GTR w/o	MP w/o	GTR+I+G GTR+I GTR w/o
None	+I+G HKY+I w/o	MP w/o	+I+G +I w/o

ties of methods for phylogeny reconstruction. Although neither method produces the "true" phylogenetic tree, the combined analysis of different topologies for the CPV and FPLV phylogeny adds high significance to both the method and the application on the specific data.

MEGA did not require specific input parameters to select an evolutionary model. The Kimura-2-parameter (K2P, [34]) and Tamura-Nei (TrN [69]) models were selected for the deference of NJ MP, and LH nucleotide tree: the models are not identical, yet equivalent to the HKY and GTR models described above, and, in contrast to those used with PAUP*, were implemented in MEGA, and vice versa. The model parameters were estimated by MEGA. The same parameter for γ-correction has been specified in PAUP* and MEGA. In total, we produced five more phylogenetic VP2 DNA trees. A (sub-)tree or branching pattern was considered stable if it occurs at least two, yet preferably in all three types (NJ, MP, and LH) of trees. We found such support for the topology obtained using different tree reconstruction methods. Within all trees, a clade of 13 or 14 taxa, respectively, assigned as CPV-2 branches first from the root, followed by the

2a and 2b subtypes, according to their geographical origin. We also observed a weak temporal clustering of taxa with respect to their isolation date. There were only few taxa where the unique isolation date was inconsistent with the temporal pattern. Such temporal clustering was usually loosely correlated to distinct "outbreaks", i.e. rather "collections" of taxa, which are assigned to different points in time but have been sampled together. The temporal order within outbreak clusters was weak, yet no major discrepancies were detected. Sequencing artefacts cannot be entirely ruled out, but are considered insignificant: neither the geographical nor temporal clustering patterns are stringent enough to support sequencing, editing or alignment artefacts of newly sequenced isolates. The antigenic patterning within nucleotide trees was strong, separating original CPV-2 types from 2a and 2b subtypes. The geographical clustering of taxa, however, was found to be even stronger, separating 35 CPV-2a and 2b species from Japan and Asia from 17 2a and 2b species collected in the USA. Within the Asian cluster, newly designated CPV type LCPV from leopard cats [32] formed a distinct subclade, as well as a three-member subclade shows up in the US cluster, possibly from wild canines. European taxa are equally present in the Asian and US subclades. The detailed figures of NJ, MP, and LH nucleotide trees are given in the appendix (Figures A.1–A.3), and a schematic summary of taxon clustering is given in Figure 2.3b.

Since in phylogenetic and sequence analyses (see also Section 2.3) the sequence diversity of the viral species has been shown to be very low, we created a graphical representation of the "relatedness" to visualise the sequence similarity. The reticular structure of the relationship between Carnivore Parvovirus species is shown in Figure 2.2. We used the SPLITSTREE

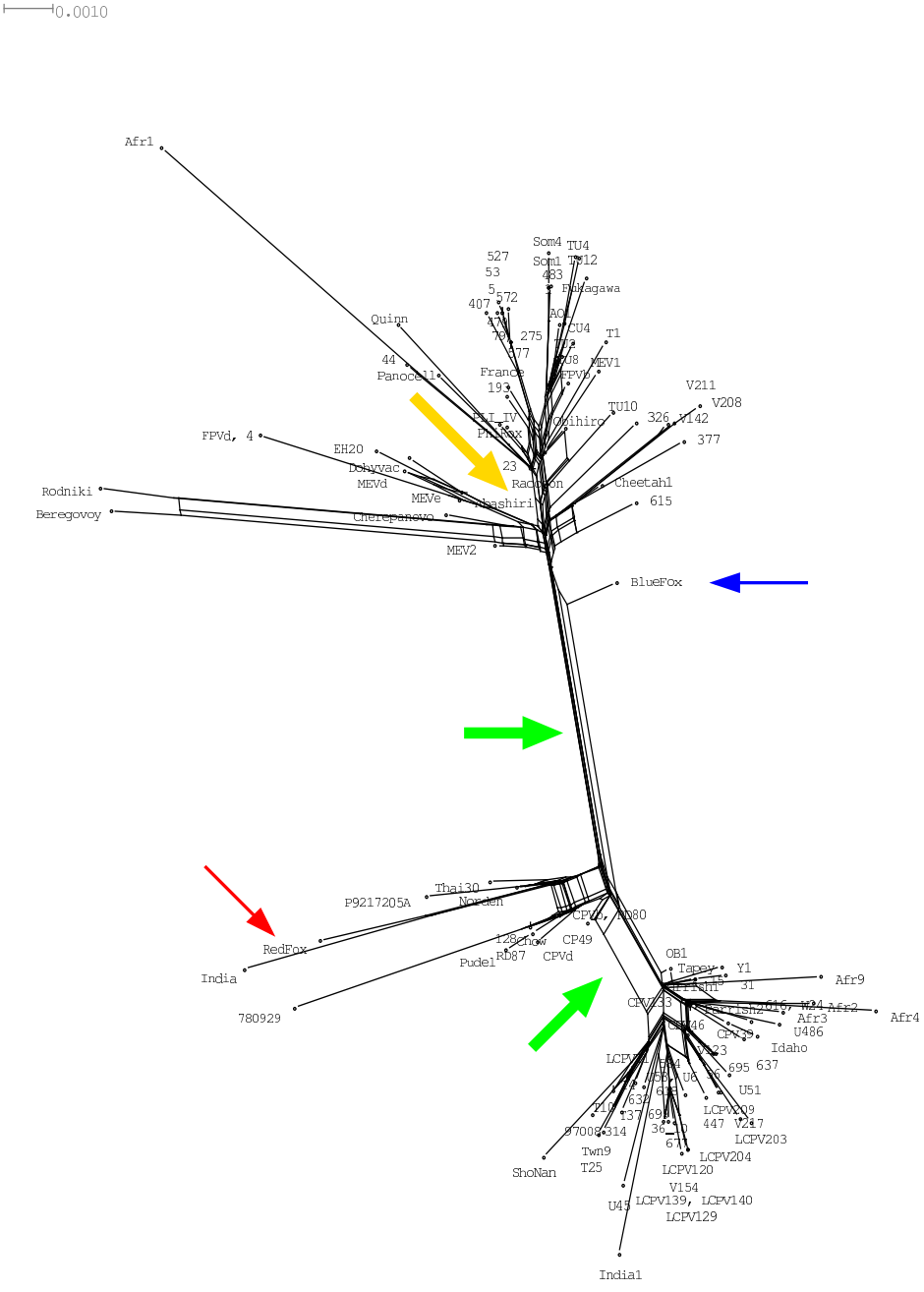


Figure 2.2 Phylogenetic network representation of sequence data as measure for sequence diversity (created with SPLITSTREE using the NEIGHBORNET method).

package (Version 4 Beta 10, [31]) to create a phylogenetic network of the sequence data, using the recently implemented NEIGHBORNET method [10, 15]. As shown in Figure 2.2, the FPLV (upper cluster) and CPV (lower cluster) subclades form distinct sub-graphs, separated by a long internal edge (upper green arrow). This internal edge is well supported by means of a significant split between the subclades. Within the subgraphs, the species show a high degree of interconnectedness, resulting from high sequence relatedness[6]. Close to the upper FPLV subgraph, we found the FPLV strain "BlueFox" (blue arrow), confirming it is an FPLV-CPV intermediate strain. The CPV isolate "RedFox" (red arrow), however, does not appear at such distinguished position along the internal edge representing the FPLV-CPV split. The FPLV subset also did not show a significant split between FPLV and MEV species, which would indicate that FPLV and MEV are distinct antigenic subtypes. However, seven out of eight MEV species were found in a larger cluster of eleven FPLV and MEV species. This is possibly an indication that MEV species might share some properties that have not been detected by the analysis of FPLV-MEV phylogenetic subtrees. Within the CPV subset in the lower part of the figure, the older CPV-2 members form another distinct subgraph. Those 14 CPV-2 species, that have also been observed as the first branching subtree in phylogenetic trees, are separated from the rest of the CPV data set by a significant split (lower green arrow). Some indication for typical long-branch effects[7] was detected in both FPLV and CPV subgraphs, although neither NJ, MP,

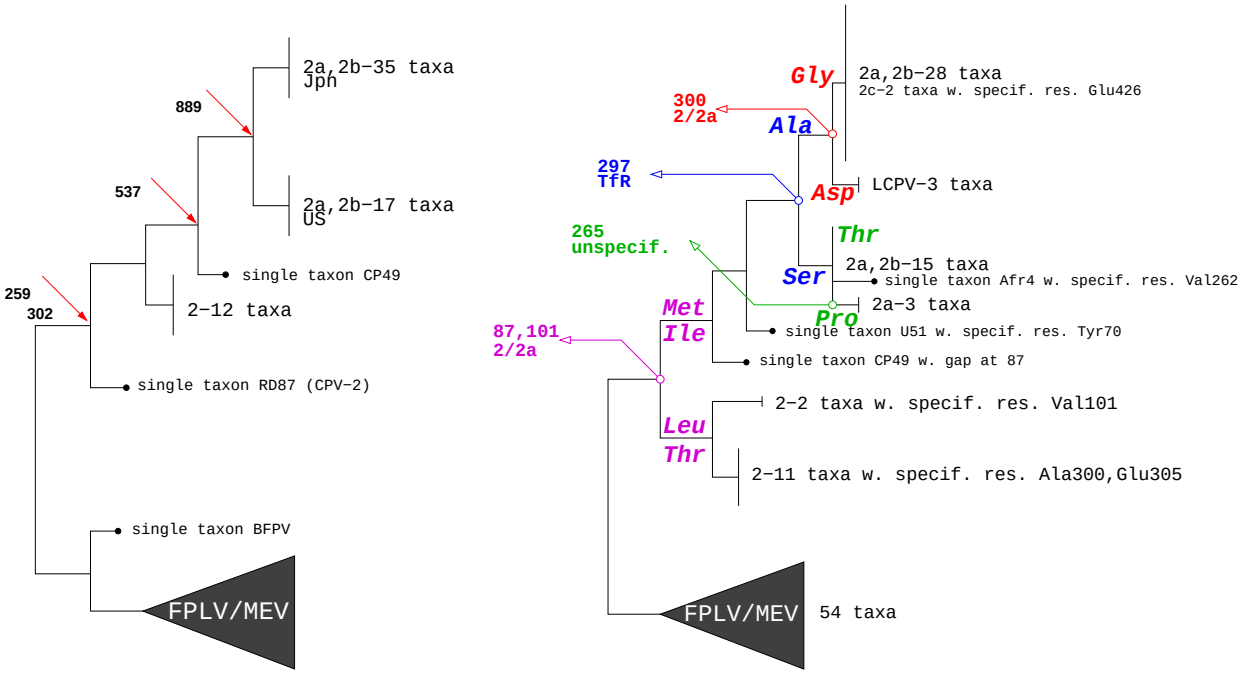
[6]This is—informally, of course—also called an ominous "phylogenetic tuft" [53]

[7]Species that have accumulated a high number of substitutions are consecutively assigned a high value for the respective branch length. Such long branches tend to be clustered by some reconstruction methods except for NJ ("long-branch-effect" or "long-branch-attraction"). They might be misleading, though, and therefore, should be avoided.

nor ML methods produced a notable long-branch-attraction effect in the phylogenetic trees.

Protein Phylogenetic Trees. We evaluated the significance of results obtained for the VP2 coding sequence using the corresponding translated sequences for reconstruction of a phylogenetic tree. The phylogeny was obtained using MEGA and PAUP* considering either the DAYHOFF substitution model and the uncorrected p-distance, i.e. number of amino acid differences, as distance measure. The topology of the protein trees could be projected onto the topology of the corresponding nucleotide trees without causing inconsistencies. However, the strength of the antigenic patterning was observed to be very low as seen by a subtree of 2a, 2b and 2c subtypes not separated from each other by an internal edge. Interestingly though, the higher-ranking geographical patterning could be observed. The temporal patterning of distinct outbreaks was shown about as strong as in the nucleotide trees. The original MP tree is shown in Figure A.4, and a schematics is given in figure 2.3a.

Considered an intermediate species, the FPLV strain "BlueFox" has shown a special behaviour in both nucleotide and protein trees. It has been found to separate CPV from FPLV strains in all trees created. In contrast, the CPV strain "RedFox", which has also been shown to possess both CPV– and FPLV-specific properties, did not appear at such a distinguished position within the phylogenetic trees. This sequence, however, is the shortest considered in our study. The low significance found for a special role of the wild carnivore virus in CPV evolution is possibly related to the incomplete sequence data.



(a) Protein (b) DNA

Figure 2.3 Schematics: Number, type and position of nucleotide and amino acid acid replacements corresponding to single branch points of the CPV VP2 nucleotide and protein trees.

Consistency of Results. Basically, using different methods and different parameters did not produce contradictory results. Differences in tree topologies are usually not well supported as indicated by very low bootstrap values. The majority of low and zero bootstrap values also corresponds to isolates with identical nucleotide sequence as identified in pairwise calculation of the number of transitional and transversional substitutions[8]. We therefore do not have to pay further attention to this kind of topology differences in NJ, MP and LH trees. An overview of taxa with identical sequence of the partial VP2 gene region is given in Table 2.2. Taxa identified as identical strains are listed first, showing both taxon names equally in use, and marked by an equals sign ("="). Taxa showing sequence identity in the VP2 gene region from nucleotide position 163−1755 (with respect to the 1755nt full-length VP2 gene) are marked by an asterisk ("*").

A significant number of internal edges and terminal branches is supported by their occurrence in two or all three types of trees. Since the methods differ considerably, the resulting tree topologies indicate that the data set is plausible and there are only few taxa showing different behaviour in different trees. The core tree topology, however, is stable among all three types of trees. Bootstrap values for subtrees and terminal branches are comparable with a standard deviation of 6.09 and a standard error of 4.524 over all internal branches with at least one corresponding internal branch or furcation pattern in another topology. The general sequence variation is low within the data set, hence bootstrap values have low significance (see also Figure 2.2).

[8]A transitional mutation refers to the replacement of a purine (A or G) with another purine base, or a pyrimidine (C or T) with another pyrimidine base, leading to four possible types of transitions. A transversional mutation refers to the replacement of a purine

Table 2.2 Sequence identity of CPV and FPLV strains in the 162–1755nt region of the VP2 gene. The equals sign "=" indicates complete sequence identity of taxa with more than one taxon name equally in use, whereas the asterisk "*" refers to different taxa with partial sequence identity in the VP2 target region.

FPLV strains		**CPV strains**	
Taxon 1	Taxon 2	Taxon 1	Taxon 2
MEV1 (MEV)	=MEVa	CPVb	CPVd*
MEV2 (MEV)	=MEVb	CPVb/CPVd	RD80*
MEV8 (MEV)	=MEVd	CPV46	CPV133*
MEV1 (MEV)	=MEVe	LCPV120	LCPV129*
EH20 (FPLV vaccine)	=Carlson	LCPV139	LCPV140*
1 (FPLV)	=941	LCPVT1	T4*
1 (FPLV)	Som1*		U53*
4 (FPLV)	FPVd*	U6	584*
Rodniki (MEV)	Beregovoy*		618*
527 (FPLV)	479 *	W24	616*
	53*	695	56*
275 (FPLV)	577 *		
	79*		

VP1 Sequence and Phylogenetic Trees. Retrieval of VP1 nucleotide sequences from GENBANK and subsequent alignment with CLUSTALW revealed a minor—yet interesting—detail. The mature VP1 and VP2 mRNAs of the VP gene are considered alternative splicing variants of the VP primary transcript of 2256nt length [48]. Splicing off consecutive 501 nucleotides at the 5′ end produces the VP2 mRNA, whereas the VP1 mRNA lacks a significantly shorter intron at the 5′ end. Approximately half the VP1 sequences deposited to GENBANK indicate different splice consensus sites for VP1 than annotated in the other moiety. There is a 2214nt and 738 amino acid VP1 variant, with splice consensus sites at nucleotide positions 78 and 121 (`join 1..78,121..2256`). Yet, there

with a pyrimidine base or vice versa, leading to eight possible types of transversions.

is another 2184nt and 728 amino acid variant, with splice sites predicted at positions 31 and 104 (`join 1..31,104..2256`). Both nucleotide sequences can be translated into protein sequences *in silico*, resulting in proteins differing only in ten residues in length and 30 different sequence positions (residues 11–40) in an alignment of both variants. Interestingly, translations for both splicing variants exist for all VP1 sequences deposited to GENBANK, regardless the specific splice consensus site indicated in the respective GENBANK file. It is also possible to translate VP primary transcripts of CPV respective outgroups used (rat and pig) into both the 728aa and 738aa variant. It is unknown whether the annotation of splice consensus sites contains any errors or in fact both VP1 variants exist *in vivo* and *in vitro*. Re-sequencing of the VP1 protein isolated from infected tissue and biochemical and genetic experiments in the lab may shed more light on this problem and identify the VP1 variant occurring *in vivo*. Furthermore, the analysis of splice sites and RNA secondary structures present in the intron region might reveal which mRNA variant is folding into the correct structure for recognition of the splice apparatus. In our study, we considered the shorter 728aa variant of VP1 to ensure correct results from phylogenetic analyses. The use of sequences of equal length and the correctness of resulting trees based on the VP1 sequence in previous studies should also be confirmed. We found that the phylogenetic trees based on the VP1 sequence are consistent with those based on VP2 sequences. The VP1 nucleotide and protein trees were superimposed to the VP2 trees showing no discrepancies in the topology (data not shown).

2.2 A New Approach

A specific feature of Canine Parvovirus evolution is the appearance of distinct lineages in the phylogeny, according to CPV antigenic subtypes. One of our main interests was to resolve the CPV phylogeny to the amino acid and nucleotide level, i.e. to identify cluster-specific substitutions along internal branches. The knowledge about the occurrence of single substitutions at specific branching points in the topology provides valuable insight into the evolutionary process. It enables us to learn more about the mechanism of host range shift at a molecular level. Although the need for such analysis tools is obvious, so far no phylogenetic software package has included this feature in neither phylogenetic nor sequence analysis tools. So we implemented a UNIX command line tool that, among other, is able to perform the following queries:

1. Given a specific subset of taxa, either specified by taxon names or the subset size, which are the nucleotide (or amino acid) positions in which the subset differs from the rest of the data set?

   Which character state replacement has occurred, separating the subset from the rest of the data, i.e. which nucleotide (or amino acid) character is specific for the subset and which character is the corresponding one in the data set?

2. Given a specific position in the sequence, what are the different character states for this position, and what are the split(s), or subsets, created with respect to this position?

3. Given a specific group of clustered taxa, what is the significance for the occurrence of the subset, i.e. are there further positions producing

the same split of the data set?

2.2.1 Implementation

The principal idea was to create a derivative tool of the UNIX `diff` command that compares files and yields the lines in which the files differ. The input file read by `vdiff` is in aligned FASTA format, ensuring equal sequence length required for position-specific queries. The phylogeny resulting from the data is not required as input; the user provides the subset for a specific query manually, instead. Appendix Section F gives a comprehensive summary of the modus operandi and the features of `vdiff`.

`vdiff` is a very simple tool, but it is able to answer important questions quickly. It can be applied to any kind of semantic problem, including molecular sequence analysis. It is especially suited for application on phylogenies with distinct lineages of closely related species, e.g. viral antigenic subtypes. The subsequent application of `vdiff` to the data set, corresponding a walk from the root to the terminal leaves, continuously reduces the subset size. Thereby, the phylogenetic patterning is resolved to single (nucleotide or amino acid) differences.

2.2.2 Application to the Data Set

The phylogenetic analyses have revealed some well-known character differences, but also some unknown positions. Apart from the well-known differences between FPLV and CPV species, we concentrated on the CPV subset. Our goal was to analyse specific edges in the graph, shown in Figure 2.3a. The first edge, representing the terminal branch for the FPLV-CPV intermediate "BlueFox" was assigned a transversional replacement

of A by T. The resulting codon change CCA to CCT both encode for proline. Since the nucleotide replacement is silent, the "BlueFox" edge was not observed in neither the phylogenetic tree (see Figure 2.3b) nor detected with vdiff. We next analysed the corresponding replacement(s) separating FPLV from CPV. Interestingly, we detected two replacements with equal significance corresponding to the first branching pattern in the condensed topology. This branching pattern separates CPV-2 species from the rest of the data set. It is no surprise we detected nucleotide position 259 as a character-state splitting position. CPV-2 members show an A at this position, whereas CPV-2a members show a T. This replacement lead to the non-silent replacement of methionine by leucine at position 87 in the VP2 protein, as it has been sufficiently proven. The "twin" replacement at position 302 shows that CPV-2 have a T, whereas CPV-2a members have a C. The corresponding change in codons from ATT to ACT at codon position two is non-silent, replacing isoleucine 101 by threonine. Both nucleotide replacements are non-synonymous, as the split was also present in the protein tree and identified by vdiff. The distinguished positioning of CPV-2 taxon "RD87" is due to two exclusive nucleotide replacements at third-codon positions 354 (G instead of A) and 1623 (A instead of C). Since a single edge for "RD87" does not occur in the protein tree, and neither a specific character position has been identified by vdiff, the replacements are considered silent. Codons 118, either GGT or GGA, both encode for glycine; codons 541, either GCG or GCA, both encode for alanine. We further detected a character state difference at nucleotide position 537 (indicated by a red arrow in the figure). It further separates the CPV-2a subset from the CPV-2 subset by a silent replacement of T by C. The codon change at position two does not affect the encoding for serine 179. The distinguished

positioning of CPV taxon "CP49" in both the nucleotide and protein tree, as well as the positioning of "U51" (C 208 replacement by T), is described in detail in Chapter 4. Another replacement at nucleotide position 889, geographically separating 2a and 2b taxa, was identified. One subset shows a T, the other one a G. The resulting codons, TCT and GCT, are responsible for a non-silent replacement at amino acid position 297 because of the first codon position is affected. The resulting amino acid replacement of alanine by serine is also observed by the split pattern in the phylogenetic tree and supported by translated sequence analyses with `vdiff`. We also performed a high-resolution character-state analysis for the Asian and US subclades, as described earlier. Three taxa each formed distinct subclades in those geographically separated 2a- and 2b-clusters. Within the Asian cluster, three LCPV members, i.e. CPV species isolated from leopard cat hosts, show an aspartate at CPV 2-/2a-specific position 300, as encoded by an A 899, altering codon position two (GAT). CPV-2 members show an alanine at position 300 encoded by **GC**T, and CPV-2a species a glycine, encoded by **GG**T. In the US cluster, another three CPV species, possibly from wild canids, are separated from another 15 species. The replacement of residue 265 has recently been shown to become prevalent in CPV populations, although its role is currently unknown. The branching pattern was correlated to a difference in amino acid position 265. The small US-European subset is divided into a three-member CPV-2a set showing a proline 265 (**C**CA), and a fifteen-member 2a- and 2b-set showing a threonine 265 (**A**CA). The role of the LCPV cluster and the wild canid cluster is also discussed in Chapter 4. The single taxon positioning "Afr4" was correlated to an exclusive replacement of alanine 262 (G**TC**) by valine (G**CT**).

All non-silent replacements identified in the translated sequence data

set were confirmed in the nucleotide sequence data set, whereas silent nucleotide replacements were not detected by `vdiff` in the protein sequence data set. Application of combined phylogenetic and semantic analyses on the FPLV subset did not reveal significant character replacements for internal branching patterns. This indicates FPLV evolves more randomly than CPV, which shows a highly ordered evolution pattern of consecutive character replacements. Distinguished single taxon positions in the phylogenetic tree were correlated with exclusive character replacements at specific positions. A more detailed discussion of single non-synonymous nucleotide and the resulting amino acid replacements, respectively, is given in Chapter 4.

2.3 Sequence Analysis

Transition-to-Transversion Ratio. We were looking at the pairwise rates of transitions and transversions[9] to investigate general properties of the whole data set. The number of transitions is expected to decrease with increasing genetic distance between two species. With closely related species, however, transitional substitutions occur more frequently than transversional substitutions [81]. The overall rate $R = s/v$ was calculated as 2.5476. The value for R would be exactly 2, assuming completely random substitutions. The plot in Figure 2.4 shows that the data do not behave—as expected—a^x-like, but R is rather constant with p-distances (green curve). In total, more than 5000 pairwise calculations were performed for 120 parvovirus species, however, producing 74 unique data points for the complete data set solely. The data set was then split into

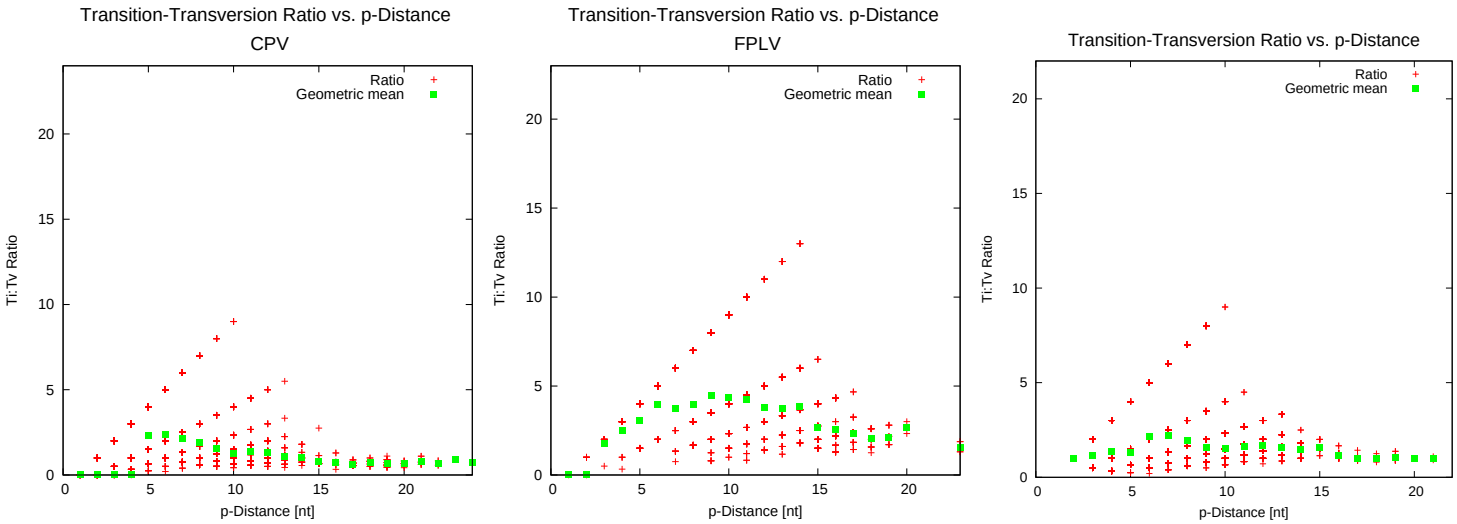
[9] Ti:Tv, or $R = s/v$, with s the number of transitions observed, and v the number of transversions observed.

CPV and FPLV subsets, and the analysis was performed for the smaller subsets separately to investigate whether the data show different properties.

The geometric mean for the Ti:Tv ratio corresponding to a specific p-distance is also shown in the figure (green curve). The overall geometric mean for CPV alone was determined to be 4.24, whereas for FPLV, the value for R is 2.21. With CPV species, the rate R is 1.9 times higher than with FPLV species, and the number of transitions is higher for CPV species. These results are consistent with the fact that Canine Parvoviruses are more closely related to each other than drifting Feline Parvoviruses. Interestingly enough, considering separate data subsets, the geometric mean values for CPV show the expected a^x-like behaviour, whereas for FPLV, the values cannot be approximated with a square, cubic, or another higher power polynomial function.

Analysis of Synonymous vs. Non-Synonymous Substitutions. We performed a dS/dN analysis to test whether CPV and FPLV are under significant selective pressure. Since FPLV has been suggested to evolve with random genetic drift, we tested the null hypothesis [10] $H_0 : dS \neq dN$ and the alternative hypothesis $H_1 : dS = dN$. In contrast, we tested the CPV population for positive selection assuming $dN > dS$ as alternative hypothesis. The results from separate dS/dN analyses are summarised in table 2.3. *dS* and *dN* values indicate the rate of synonymous and non-synonymous substitutions in the VP2 gene, respectively. The corrected values of possible substitution rates per synonymous and non-synonymous sites N

[10] Neutral evolution is considered to show equal average rates of synonymous and non-synonymous changes, whereas non-silent changes dominate in positive selection and silent changes dominate in purifying evolution.



(a) CPV (b) FPLV (c) All

Figure 2.4 Plot of the transition-to-transversion ratio vs. the uncorrected nucleotide distance **Left** CPV only **Middle** FPLV only **Right** Complete data set.

Table 2.3 Results from analysis of synonymous vs. non-synonymous nucleotide substitutions in the VP2 gene of CPV and FPLV species.

	Population			**Factor**
	CPV	FPLV	Total	CPV:FPLV
dS	0.010022	0.016497	0.022131	0.61
Sd/S	0.009931	0.016279	0.021711	
dN	0.003503	0.002974	0.006666	1.18
Nd/N	0.003488	0.002966	0.006625	
dS/dN	**5.83**	**8.48**	**4.99**	**0.69**

($pS = S_d/N$, $pN = N_d/N$) is also given in the table.

We determined the overall ratio[11] of dS/dN for CPV to be 5.83 and for FPLV to be 8.48, i.e. the rate of synonymous vs. non-synonymous substitution is about 1.5 times higher within CPV than within FPLV species. When analysed separately, the rate of synonymous substitutions of CPV is about half the rate of synonymous changes in FPLVs. In contrast, the number of non-synonymous changes is about 1.2 times higher in CPVs than in FPLVs. The calculations of the rates for dS, dN, and the ratio dS/dN were performed for FPLV and CPV species separately, as were the tests hypotheses for neutral and positive evolution. The hypothesis for $dN > dS$ was confirmed with $p = 1.00$, and the hypothesis for $dS < dN$ and $dN = dS$ were rejected with probabilities p of 0.01 and 0.02, respectively, for CPV. The hypotheses $dN > dS$ and $dS < dN$ were rejected for FPLV ($p = 0.02$ and $p = 0.00$), and the hypothesis $dN = dS$ was confirmed with a probability p of 1.00. Figure 2.5 shows a plot of the cumulative number of silent and non-silent changes per site of the VP2 protein. Indels are not shown in the figure, but have been observed to occur at the beginning and at

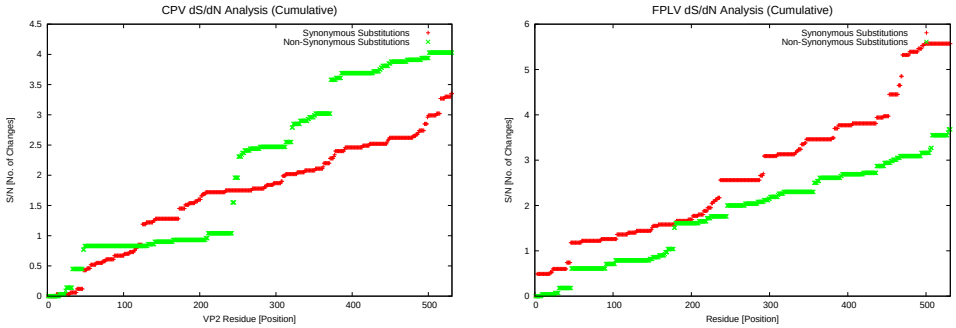
[11] In contrast to the codon-per-codon based method to determine the number of non-synonymous and synonymous changes, the number of S and N per coding site was determined starting from the first to the last position of the partial protein sequence as the average over the data set.

the end of the sequence solely, i.e. in regions of partial sequence availability. With FPLV (shown in figure 2.5b), the number of non-synonymous substitutions is higher that the number of non-synonymous substitutions throughout the protein sequence. There is only a narrow sequence window where both rates are nearly equal. Yet with CPV (shown on the left side in the figure), there are three major, distinct windows of considerably larger size. In the first part, the number of non-silent substitutions is larger (until around residue 90). The number of non-silent changes then remains relatively constant for about 200 residues. Around residue 280, however, there is a sudden increase in the number of amino acid replacements, indicating some kind of "hot spot(s)" for non-silent changes. From this region on, the number of non-silent replacements prevails the number of silent changes throughout the sequence until the end of the chain. Although smaller in size, another "jump" in the number of non-silent changes was observed for CPV around residue 330[12].

2.4 Phylogenetic Considerations

CPV and FPLV Do Not Share a Common Evolution Pattern. We constructed phylogenetic trees for CPV based on the partial sequence of the surface antigen, considering all antigenically informative positions affecting the host range. In total, more that 20 different phylogenetic trees based on the nucleotide sequence have been constructed, using different software packages and methods for phylogenetic deference. We used the Neighbor-Joining [57] method as representative for distance methods, the maximum parsimony method and the maximum likelihood method. The

[12]Those target sites are possibly corresponding to host-range determining residues 93, 297 and 323, but have not yet been confirmed.



(a) CPV (b) FPLV

Figure 2.5 Diagram showing the cumulative number of synonymous (red) and non-synonymous (green) amino acid changes in the 531aa region of the VP2 protein. Indels are not shown in the figure.

idea of using all methods available for phylogenetic reconstruction was to detect stable branching patterns in different topologies. The calculations for phylogenetic reconstruction were repeated five times with no major discrepancies in the resulting tree topologies with respect to the previous round. The trees have proven stable with minor subtree differences. The NJ tree was selected as representative optimal tree obtained. The following discussion refers to the NJ tree in particular (see Figure A.1 in the appendix), but for the most part may also be applied to MP and LH trees (Figures A.2 and A.3). The figures of the NJ, MP and LH trees are given in Appendix A.

FPLV together with MEV species, and CPV species form distinct clades in the phylogenetic tree. The split of feline and canine sequences is highly significant. The sequence of a CPV isolate from a Blue Fox (BFPV) has been found to show intermediate properties of FPLV and CPV [74] in nucleotide, but not in protein trees. In the study by Truyen *et al.*, BFPV has been found in between FPLV and CPV in a phylogenetic tree. Our re-

sults confirm those obtained by Truyen *et al*. BFPV has been found to be stably positioned in between FPLV and CPV with bootstrapping values > 90. It is stably found closer to FPLV than to CPV species with bootstrap values of ~ 50. This finding further supports the hypothesis that CPV has arisen from an intermediate virus adapted to a wild carnivore host, which is closely related to Feline Parvovirus species (see also Figure 2.3). At present, there are several theories for the transmission of the virus from a feline via a wild carnivore to a canine host. One hypothesis considers North European animal husbandries where wild and domesticated felids and canids are in close contact [U Truyen, personal communication, 2004]. The fact that the BFPV isolate showing FPLV and CPV intermediate properties origins from Finland adds further weight to this theory.

A Tripartite Clustering Hierarchy for CPV. We classified viral species into three groups according to their geographical origin: sequences from Europe and Asia, Sequences from Japan and Thailand, and Sequences from North America and Australia. In Figures A.1–A.3, the terminal branches and subtrees are coloured according to these geographical groups. We observed that the majority of sequences formed geographical clusters rather than clusters with similar isolation dates. This is, again, a consequence of difficulties with exact dating of the isolates. Geographical clusters are not strictly limited, indicating that the topology is not corrupted by sequencing artefacts. Sequences from Japan[13] have shown to be less susceptible to clustering with European and insusceptible to clus-

[13]For convenience, the term "Europe" is also used including both the European and Central Asian species, "Japan" refers to the group of Far East viruses, and the terms "US" and "North America" are equally used for sequences with origin from the USA and Australia.

tering with North American sequences, though[14]. Another obvious clustering pattern for CPV was according to subtypes. A simplified view of the phylogeny obtained is given in figure 2.3b. It is interesting that CPV species from Asia and species from Europe and North America cluster together rather than CPV subtypes 2a and 2b. CPV-2b subtypes are assumed to have emerged from CPV-2a, which has replaced the original CPV type 2. This pattern can be found in all trees for European and North American CPV species, but the restricted pattern of "Eastern" vs. "Western" sequences was not observed for FPLV. It is possible that sequencing artefacts have an effect on the tree topology, but they are very unlikely to be the reason for the drastic isolation of Far East sequences. There is also no internal edge in the topology of the FPLV clade separating Japanese from other sequences, although FPLV species are roughly clustered in geographical subtrees. This is another indication that geographical clustering is not a consequence of local isolation and sequencing. Consequently we do not have to consider sequencing and alignment artefacts to be the cause for the geographical separation of CPV subtypes. We also resolved this clustering pattern to amino acid and nucleotide level. The exact function of the residues responsible for the split has to be identified in classical genetic studies.

A Novel Antigenic Subtype. CPV-2c which has been reported a new antigenic subtype in Japan [33] could not be confirmed in our study. LCPV140, designated a CPV 2c/2a type was found with other LCPV species annotated as type 2a and LCPV203, designated a 2c/2b virus

[14]Hypothetically, the geographical split might correspond to the host phylogeny, indicating different immunological properties of the host. However, no such co-evolution analyses of CPV and its host has been performed with respect to the geographical origin.

grouped with other 2b LCPV species rather than together forming an 2c cluster. Yet after a re-assignment of nucleotide substitutions specific for CPV-2c, we found evidence for the occurrence of a true CPV02c type in Europe. Three CPV species, previously assigned 2c, have been shown to possess CPV-2c specific properties, yet neither of earlier designated types 2c/a and 2c/b show this property.

Is There a True MEV? In contrast to CPV, we observed no strict clustering pattern for FPLV and MEV species. First, there are no subtypes defined for FPLV, and second, the evolution of FPLV is not considered directed, resulting in an—at first sight—unsystematic subtree. The MEV species are divided into two low significant clusters in the subtree, one containing the Central Asian MEV species, and another with MEV clustered together with two vaccines, "Dohyvac" and "Panocell", and an early FPLV isolate, FPLV-d. The MEV species, showing FPLV properties, cluster by emergence date and geographical origin. The Central Asian species "Rodniki", "Beregovoy" and "Cherepanovo" are new isolates from the late 1990s, whereas the North American MEV species are from the mid 1950s to late 1970s, when the first isolates have been sequenced. In general, FPLV sequences have shown to cluster geographically and only very roughly with respect to their isolation dates. We also spotted subtopologies for old and recent FPLV species, divided into European–North American and Asian subtrees, yet the significance as indicated by bootstrap values and sequence analysis is low.

The Role of Vaccines. The role of vaccines in the emergence of CPV is unclear. The hypothesis that CPV has emerged from live FPLV vaccines is hardly supported, but neither could have been completely abandoned yet.

We found no evidence that either of five FPLV vaccines is a putative ancestor of CPV in the CPV subtree. Neither have CPV vaccines (three strains), which have been added to the data set later, been shown to be close to the root, i.e. the FPLV-CPV split. Within the FPLV clade, FPLV vaccines are widely distributed in the topology. Interestingly though, a European CPV isolate designated "Quinn" is stably located within the FPLV clade. Together with the FPLV-derived vaccine strain "Panocell", it forms a stable pair of terminal taxa with a bootstrap value of ~ 60. Since in previous studies [73], CPV "Quinn" has been shown to fall into the CPV clade and represents one of the old type 2 isolates dated 1980, the results presented might provide partial support for the theory of CPV emergence from FPLV vaccines. This mystery has been solved, however, revealing that CPV "Quinn" is actually a FPLV species, and that the sequence retrieved from the database is incorrect. A detailed discussion of the "Quinn" CPV vs. FPLV contradiction is given in Chapter 4, including structural considerations. The true sequence of CPV "Quinn" was, however, not included into this study due to the late discovery of the mix-up of sequence data.

Distinct Evolution of the Surface and Internal Antigens. We were looking for confirmation of the results obtained for nucleotide sequence trees combining the NS and VP gene. Thereby, we were able to increase the number of phylogenetic informative sites using a longer combined sequences of 3762nt length. The trees obtained with different methods have proven to be stable, i.e. we observed the same or a similar branching pattern with few differences solely. We also got support for internal and terminal branches by similar bootstrap values. Due to the low number of CPV NS1 sequences (five taxa) available, we were not able to get further

support for the phylogeny of CPV. We could not confirm the topology for CPV exactly, but we found support for the topology obtained for FPLV (17 taxa available). Since CPV undergoes directed evolution which is immediately associated with evolution of the surface antigen, combining external and internal antigen sequences naturally alters the phylogeny. With FPLV, whose surface antigen is not under such selective pressure, there are obviously less differences in the topology. The phylogeny obtained by the partial VP2 gene, however, is close to the phylogeny based on the complete gene sequence. Similar to VP2 partial gene trees, CPV and FPLV form stable clades with bootstrap value of 99 and 100. With a bootstrap value of $/sim70$, taxa with partial identity of the VP2 gene, i.e. "Som1" and "94-1", were better resolved than in the VP2 partial gene tree due to the additional sequence information. The results obtained from combination of NS1 and VP2 are consistent with those obtained by Horiuchi *et al.* [26] using branch-and-bound maximum parsimony analysis. Two thirds of the data origin from Japanese isolates, but we could not find direct evidence for local sequence artefacts, which, however, must not be ruled out completely. We also did not detect another specific clustering pattern for FPLV as with partial VP2 gene trees.

The Phylogenetic Role of Silent and Non-Silent Replacements. We also analysed phylogenetic trees based on the amino acid sequences of the VP2 gene. The majority of nucleotide substitutions is synonymous, thus we were looking for differences in the phylogeny based on the effector sequences. Basically, the phylogenetic reconstruction based on the translated partial VP2 sequence did not conflict with that based on nucleotide sequences. However, with a bootstrap value of ~ 70, the sep-

aration of FPLV and CPV is less well supported than in nucleotide trees. The separation is nevertheless considered significant enough. We detected a nearly identical "old" subtree of CPV type 2-like sequences in the protein tree. The German isolate from a Red Fox is found in this clade. The virus is supposed to have emerged early, presumably even before the first CPV-2 isolates from domesticated canids. The Red Fox isolate is not dated but it is known that it was collected in the 1990s. The antigenic and biological properties in a wild carnivore host are poorly understood, and the virus appears to evolve slower in the wild carnivore host. It appears that a CPV type 2-like viral species is able to spend a significant period of time in a wild carnivore. We suggest to confirm the wild carnivore hypothesis collecting more tissue samples from infected wild canids, and sequencing of the isolates containing viral DNA. The local separation of 2a and 2b CPV subtypes—i.e. especially the fact that no North American and Japanese sequences were found together—is not apparent in the protein tree. The canid taxa, however, group roughly according to 2a and 2b subtypes with few exceptions. The exceptional sequences show 2a- and 2b-specific features—i.e. residues—although they fall into the opposite sub-clade, respectively. After analysis with our sequence analysis tool, we actually detected some GENBANK mis-assignment of taxa.

Within the FPLV clade, we found support that FPLV evolves by random genetic drift. Sequences from different geographical origin, and likewise, with different isolation date are clustered together. Vaccines or MEV species were not identified as distinct subtrees. However, there are subtrees showing an accumulation of different old FPLV isolates, and a high proportion of geographically related species, respectively. This supports the the idea that FPLV evolves rather slowly with low molecular rate, and

that antigenic properties are less susceptible to amino acid variation. Interestingly though, the sequence from a Finnish Blue Fox is found deeper in the protein tree than in the nucleotide tree. This behaviour has been found to be due to a single exclusive nucleotide substitution, which is silent.

Our final conclusions from classical phylogenetic and sequence analysis, combined with statistical analysis of character-state specific subsets are:

1. CPV lineages 2, 2a, 2b and 2c differ from each other in more nucleotide and corresponding amino acid positions than hitherto known.

   The majority of newly identified nucleotide replacements results in non-silent amino acid replacements, and hence, might play a significant role in host cell targeting.

2. The re-evaluation of CPV-2c specific properties revealed true 2c types in our data set, whereas neither of earlier designated types 2c/a and a 2c/b were shown to possess 2c-specific residues

3. Significant branching patterns separating antigenic lineages, or isolating single taxa, can easily be correlated to single character replacements. In the latter case of single taxon branches, those replacements are exclusive to the taxon rather than shared with another taxon or taxa.

4. In contrast to FPLV, CPV evolution is associated with a strictly consecutive order of nucleotide and amino acid substitutions, resulting in distinct lineages. Such orderly replacement does not occur in FPLV.

5. The arbitrary choice of parameters for phylogenetic reconstruction considered optimal for the target data is compensated in combined use with the statistical character-state approach. The character-state approach does not take into account weighting parameters considered in character substitution models and therefore, cannot be corrupted by model artefacts. The significance of splits observed is very high.

2.5 Tools

Origin of Sequences. We used two different sources to obtain sequences from Parvovirus samples. A total of 95 feline and canine parvovirus species sequences were retrieved from original GENBANK files. Those CPV and FPLV taxa have been selected using a nucleotide BLAST [4] search based on the complete VP2 coding sequence of CPV strain "Norden". Sequences deposited to GENBANK after March 2004 have not been included to this study for convenience. Another 25 unpublished feline and canine parvovirus nucleotide sequences were retrieved from original sequence and corresponding alignment files. The corresponding protein sequences were translated from the coding sequences available and cross-checked for errors by comparing them to amino acid sequences deposited to GENBANK whenever possible. In total, 120 different species have been considered in this study (see also Table C.1 in the appendix). Collected samples cover a period of 50 years, and were isolated on all continents.

Sequence Analysis. The multiple alignment of nucleotide and amino acid sequences was performed using CLUSTALW Version 1.81 [70] and optimised using the `code2aln` package Version 1.1 [66]. The transition-to-transversion ratio of nucleotide substitution was determined using MEGA

2.1 and MEGA 3 (Test7) [37, 38], respectively[15]. The analysis of synonymous vs. non-synonymous substitutions was performed using the SNAP program package [35,36,42]. The Z-test for selection based on a large sample was performed using the Nei-Gojobori method [36,43] using pairwise p-distances, which is implemented in MEGA.

Phylogenetic Analysis. We used two programs for inference of phylogenetic trees, the UNIX based PAUP* (Version 4.0 Beta 1) package [68] and the Windows application MEGA (Versions 2.1 and 3 Test 7) [38]. To increase the support for a stable topology, we compared results obtained by NJ, ML, and MP methods. With nucleotide sequences, we considered both the optimal HKY+I+G evolution model [22] suggested by MODELTEST [52] as well as the more general GTR model [39, 56] for nucleotide substitution rates. Reconstruction of the phylogeny based on the VP2 translated sequence was performed using the DAYHOFF substitution model [14].

Identification of Taxon-Specific Substitutions. It was our goal to keep track of the specific nucleotide and amino acid substitutions in the tree and to assign specific replacements to particular branching points in the phylogenetic tree. Analysis of parsimony informative and singleton sites was hence performed using `vdiff` [25]. `vdiff` is implemented in Python. It performs site-specific and taxon-specific queries in a set of sequences of equal length, i.e. aligned nucleotide or amino acid sequences. For more details, see the manpage in Appendix F or Section 2.2.

[15]Results from the unreleased test version have been confirmed using the earlier release version and did not differ from each other.

Chapter 3

Evolutionary Rates

"And now for something completely different"
Monty Pythons

3.1 A Novel Method

Common molecular clock models may not be applied to CPV populations for two reasons: First, the exact time of CPV emergence is unknown. The oldest isolates date back to the late 1970s, however, the virus is presumably older than isolation dates indicate and may have spent a significant time in an intermediate reservoir host. Second, molecular clock models require molecular rates to be relatively constant over t, which is clearly not the case with CPV viruses: Phases of adaptive evolution to the host and of host range shift then must be associated with different molecular rates. Low molecular rates indicate adaptive evolution, whereas antigenic or host range shift requires raised substitution rates. We subsequently established a straight-forward arithmetic model to estimate CPV and FPLV molecular rates separately.

The basis for the formula derived is shown in Figure 3.1. All pairs of

taxa with (1) $\Delta t > 0$ (2a) $b_1 \neq b_2$ and (2b) $b_1, b_2 \neq 0$ have been considered to calibrate the molecular clock for the species. After calculation of τ, i.e. the estimated time of the fictive common ancestor of taxa 1 and 2, pairs leading to unreasonable values for τ have been removed from the calculation scheme. Such pairs with nearly equal branch lengths b_1 and b_2 have not been considered since values obtained for τ are highly misleading (one hundred years and more) and values for λ are far too low (close to zero at decimal places). The arithmetic model does not account for such pairs and hence, calibration of the molecular clock is not possible with equal and nearly equal branch lengths. The model is based on the simple correlation $t = \frac{d}{2\lambda}$, which can also be written as $d = \lambda \cdot t$ with a specific value for λ for a pair of terminal taxa 1 and 2 rather than for two separate branch lengths b_1 and b_2 specified by the same pair of taxa. We actually assume λ to be constant along the sum of branch lengths b_1 and b_2 instead of branches, and consider all pairs of terminal taxa in the tree with $\Delta T \neq 0$ and $b_1, b_2 \neq 0$. We denote

$$\begin{aligned} \tau &= \Delta T + 2\delta \\ &= \Delta T \cdot \frac{b_1 + b_2}{|b_2 - b_1|} \end{aligned} \tag{3.1}$$

as the total time passed along branches b_1 and b_2 and calculate

$$\lambda = \frac{d}{\tau} \tag{3.2}$$

with d to be the Hamming-distance of terminal taxa 1 and 2, indicating the actual number of nucleotide differences. Considering all pairs i of terminal

taxa fulfilling the condition $\Delta T_i \neq 0 \wedge b_{1i}, b_{2i} \neq 0$, we calculate

$$\begin{aligned} \lambda^* &= \sqrt[n]{\lambda_1 \cdot \lambda_2 \cdot ... \cdot \lambda_i \cdot ... \cdot \lambda_n} \\ \hat{\lambda} &= \sum_{i=1}^{n} \lambda_i \cdot \frac{1}{n} \end{aligned} \tag{3.3}$$

as the arithmetic and geometric mean over all values for λ_i in CPV and FPLV populations separately.

3.1.1 Implementation

We implemented a method to automatically determine the respective branch lengths between a pair of taxa based on a standard NEWICK tree. The principal idea is to represent a leaf by means of the path spanned between the root and the leaf. It determines the length of the path from a terminal leaf to the root by reverse walking along the path and keeps track of the distances between each child and its parent node. The total branch length is calculated cumulatively during the reverse walking process. The distance between a pair of taxa is calculated as the sum of the total branch lengths of taxon A and taxon B, minus the shared part of the paths. We denote

$$dist(taxa) = path(A) + path(B) - 2 \cdot path(parent_{AB}) \tag{3.4}$$

which equals the numerator $b_1 + b_2$ in the fraction in Equation 3.1. We further calculate

$$dist(path) = |path(A) - path(B)| \tag{3.5}$$

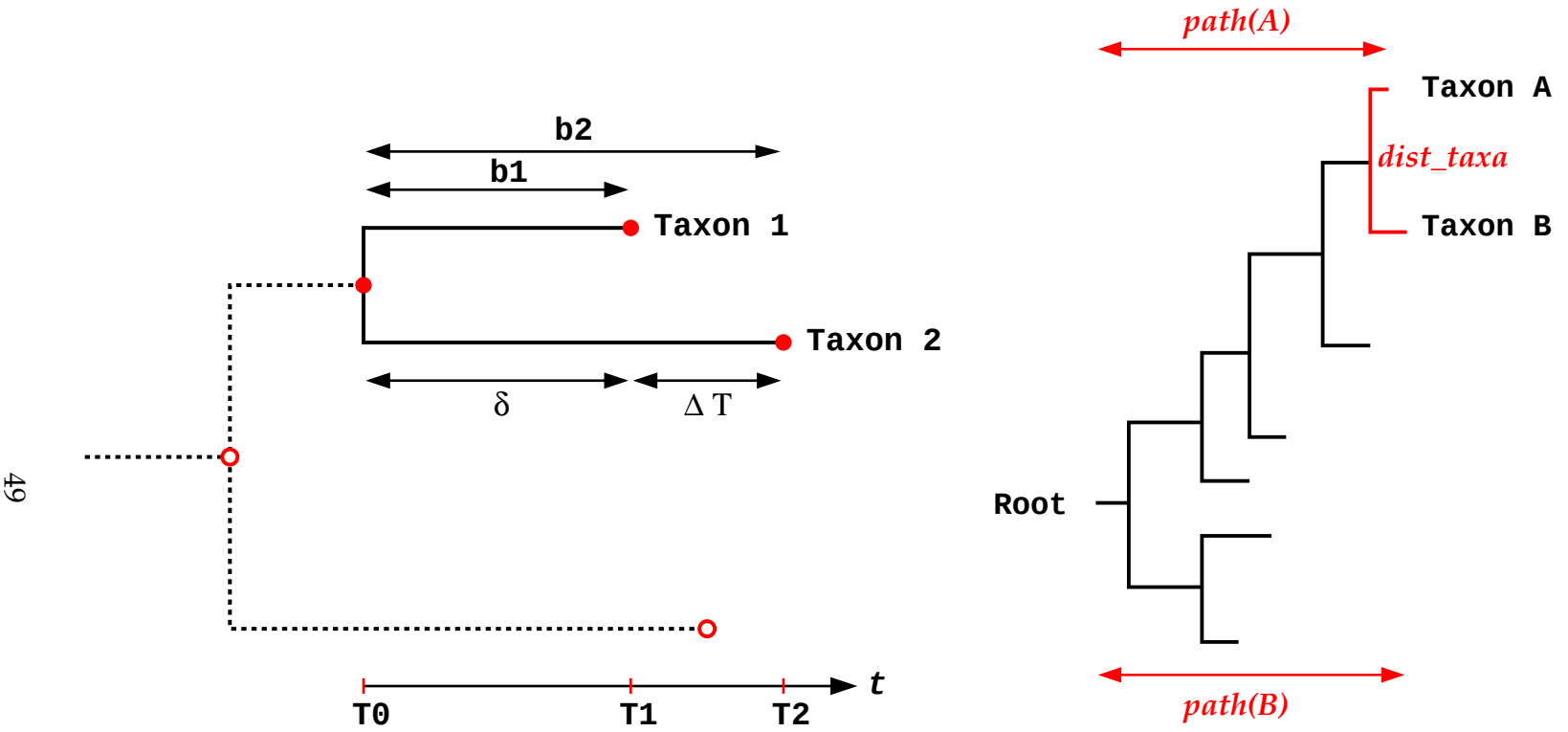


Figure 3.1 Left Schematics: Estimating the molecular rate manually from sequence information. $\mathtt{b}_1$ and $\mathtt{b}_2$ indicate the respective branch lengths of terminal taxa 1 and 2 by means of the number of nucleotide substitutions along the branch, ΔT refers to the difference of isolation dates $\mathtt{T}_2$ and $\mathtt{T}_1$ in years, and δ is an auxiliary variable to compute $\mathtt{T}_0$, i.e. the estimated date of a fictive common ancestral sequence of taxa 1 and 2. **Right** Schematics: Determining the branch lengths between two terminal taxa in a NEWICK standard tree.

providing the denominator $b_2 - b_1$ required by the expression, which equals the absolute distance of path lengths A and B. The values for T_1 and T_2 read from the input file, and all combinations of terminal taxa are considered. We finally obtain a list of $n \cdot (n-1)$ elements, providing $dist(taxa), dist(path), \Delta T$ and $p - dist$ required for the calculation of the auxiliary time correction variable τ, T_0, and the molecular rate λ in table format. The program `treepather` [24] is implemented in `C` and `C++`. It reads the aligned AFA file (see section 2.5) to calculate pairwise distances, and the phylogenetic tree in NEWICK standard format (see Appendix E) as input.

3.1.2 Generating Artificial Sequence Data

The correctness of the assumptions made in the model presented above was verified using a set of simulated sequences. Sequence data using a specified topology of the phylogenetic tree was generated using the ROSE [67], SEQGEN [54] and TREEVOLVE [20] packages assuming different evolution models. Our model for λ calculation was then applied to the artificial sequence data set and we determined the molecular rate λ to rule out artefacts in determining the branch lengths b_1 and b_2 and caused by the higher number of recent isolates.

3.2 Application to the Data Set

The molecular rate for CPV considering 60 different viral species has been determined. As with the phylogeny, the molecular rate calculation is based on the surface antigen and considers all types of nucleotide sub-

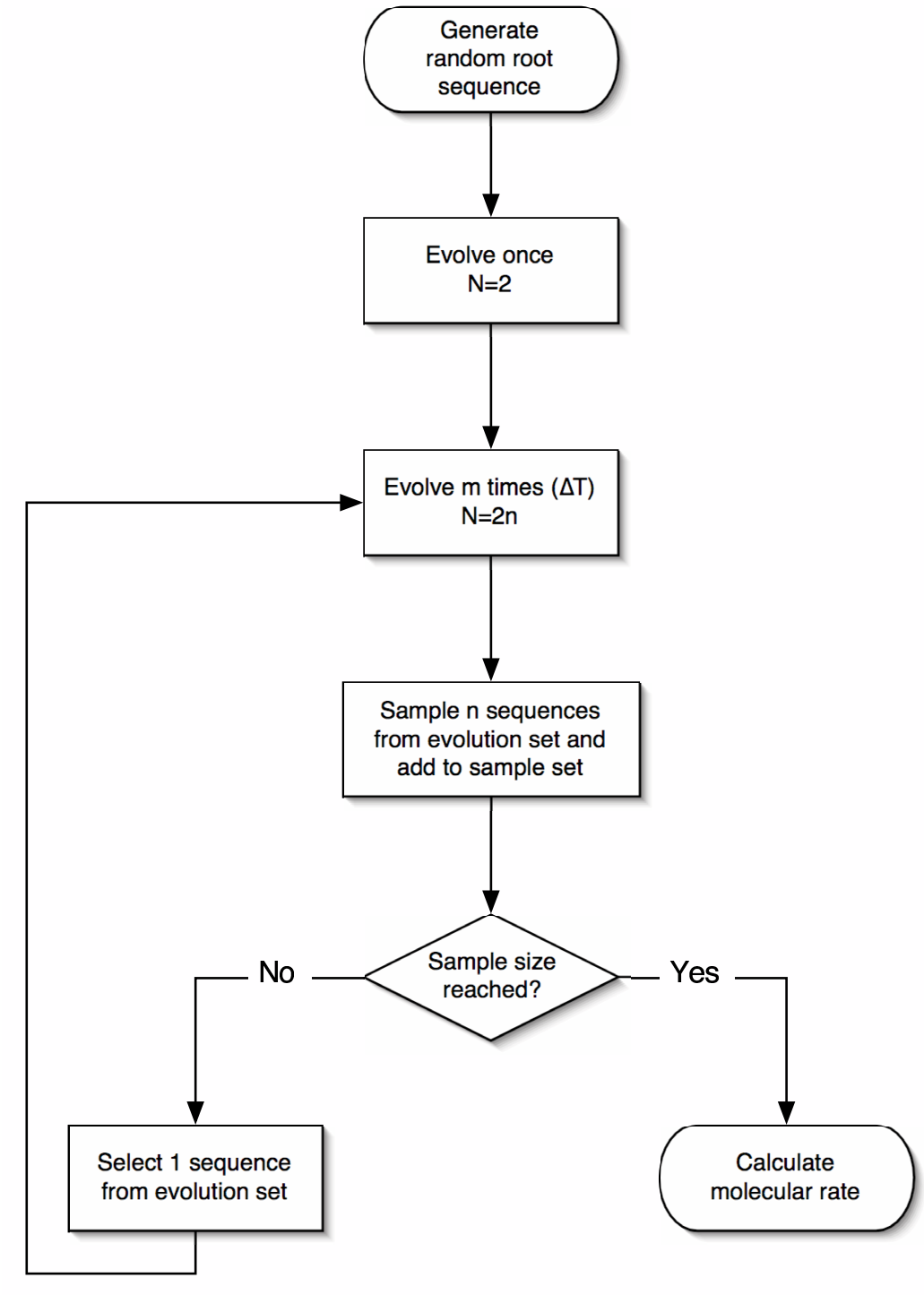


Figure 3.2 Schematics: Mimicry of sequence evolving, sampling, and sequencing using a random sequence generator and sequence evolution tools.

stitutions, silent and non-silent changes[1]. All pairs of taxa with $\Delta T > 0$ and $b_1 \neq b_2$ have been considered to calibrate the molecular clock for the species. After calculation of τ, i.e. a time correction factor for the time passed along the pair of branches, pairs leading to unreasonable values for τ have been removed from the calculation scheme. Such pairs with nearly equal branch lengths b_1 and b_2 have not been considered since values obtained for τ are misleading (higher than 100 years and more) and values for λ are far too low (close to zero at decimal places). The arithmetic model does not account for such pairs and hence, calibration of the molecular clock is not possible with equal and nearly equal branch lengths.

The mean value for the molecular rate $\lambda_{\ CPV}$ has subsequently been determined to be 0.72 ± 0.019 nucleotide changes per year (arithmetic mean $\pm s.e.$) with a standard deviation of 0.459. For FPLV, the values are 0.56 ± 0.018 nucleotide changes per year (arithmetic mean $\pm s.e.$) and a standard deviation of 0.572. A graphical summary of the λ calculation results is given in Figure 3.3, which is described in detail below, and also summarised in Table 3.1.

3.2.1 Statistical Analysis of λ Distribution.

Assuming that molecular rate values for CPV are generally higher that for FPLV ($H_0 : \lambda_{CPV} > \lambda_{FPLV}$), we performed a metric t-test for two independent samples with unknown variances to test whether the distribution of λ values for the CPV and FPLV populations differs significantly from each other. We also performed a non-parametrical Wilcoxon test assuming a

[1]Assuming a positive selection force, the molecular rate is calculated based on non-synonymous substitutions, whereas under purifying selection, synonymous changes are considered. With neutral evolution, both types of substitutions may be considered, also when the selective force is unknown, as it is with CPV and FPLV.

Table 3.1 Results of the computation of CPV and FPLV molecular rates and summary of the statistical analysis of CPV and FPLV molecular rates

Parameter	**Population**		**Unit**	**Factor**
	FPLV	CPV		
Arithmetic mean	0.5579	0.7172	nt/year	1.3
	3.5×10^{-4}	4.5×10^{-4}	nt/year/site	
Standard error	0.0178	0.0189		
Standard deviation	0.5716	0.4585	nt/year	$\overline{0.8}$
	3.6×10^{-4}	2.8×10^{-4}	nt/year/site	
Geometric mean	0.3686	0.5912	nt/year	1.6
	2.3×10^{-4}	3.7×10^{-4}	nt/year/site	
Median	0.3645	0.6402	nt/year	1.3
	$\mathbf{2.3 \times 10^{-4}}$	$\mathbf{4.0 \times 10^{-4}}$	nt/year/site	
Two-sided t-test	$t = 5.716$, $df = 1613$,$p = 1.155 \times 10^{-9} < 0.05$			
Wilcoxon test	$W = 400982$			

different null hypothesis. Both results from the parametrical ($t_{1613} = 5.716$, $p = 0.05$) and non-parametrical test ($W = 400982$, $p = 0.05$) indicate with high significance that the distribution of λ is indeed different in CPV and FPLV populations. The results from the statistical analysis of the distribution of the molecular rate of nucleotide substitution in CPV and FPLV populations are summarised in Table 3.1. We observe that first, median and mean values are lower for FPLV, whereas second, the standard deviation is higher for FPLV. The standard error is nearly equal in both populations.

λ Distribution. To investigate the significance of results obtained from λ calculation, we first analysed the distribution of λ frequencies. The graphs of λ frequency distributions for CPV and FPLV is shown in the upper left panel of figure 3.3. The distribution of λ values for CPV is shown in red, and the highest frequency is observed for the median. We assumed a stochastic distribution for the molecular rate and observed that low λ values occur more frequently in the FPLV population, as also indicated by

a median value about half as for the CPV population. Two statistical tests were performed to obtain significant support for the difference of λ values for the two species (see above).

λ over t. Based on the model presented earlier (see Formulae 3.1 and 3.2), $T_0 = T_2 - (\Delta T + \delta) = T_1 - \delta$ is considered the date when a fictive common ancestor of taxa 1 and 2 has emerged. We determined T_0 as T_{0i} for all pairs i in the tree which fulfilled the same conditions as for the calculation of λ. Each value for λ_i has been plotted against the corresponding value for T_{0i} to investigate whether λ is constant over t or has been subject to major variation. Panel b the upper right of figure 3.3 shows that for CPV, there is a small peak in molecular rate values around 1980, a second one between 1985 and 1990, and since then, a continuous increase can be observed. An increase of the variance of λ values can also be observed, correlating to the sample size with one exception around 1977. We analysed the distribution of FPLV values for the molecular rate. In contrast to CPV, the geometric mean values appear relatively constant over time, with the highest values among late isolates. However, no trend was observed for FPLV as it was seen for CPV. Although late ancestors show higher λ values than early ancestors do, they alternate with significantly lower values. The significance of these results was tested using a simulated data set of artificially generated sequences (see Section 3.2.2).

λ vs. ΔT. Each value for λ has been calculated pairwise, and is corresponding to a value for ΔT of each pair i.e. the difference of isolation dates of taxa in years. With increasing values of ΔT the number of nucleotide substitutions is expected to increase equally, resulting in values for λ to be nearly constant. As can be seen in panel C of figure 3.3, λ is

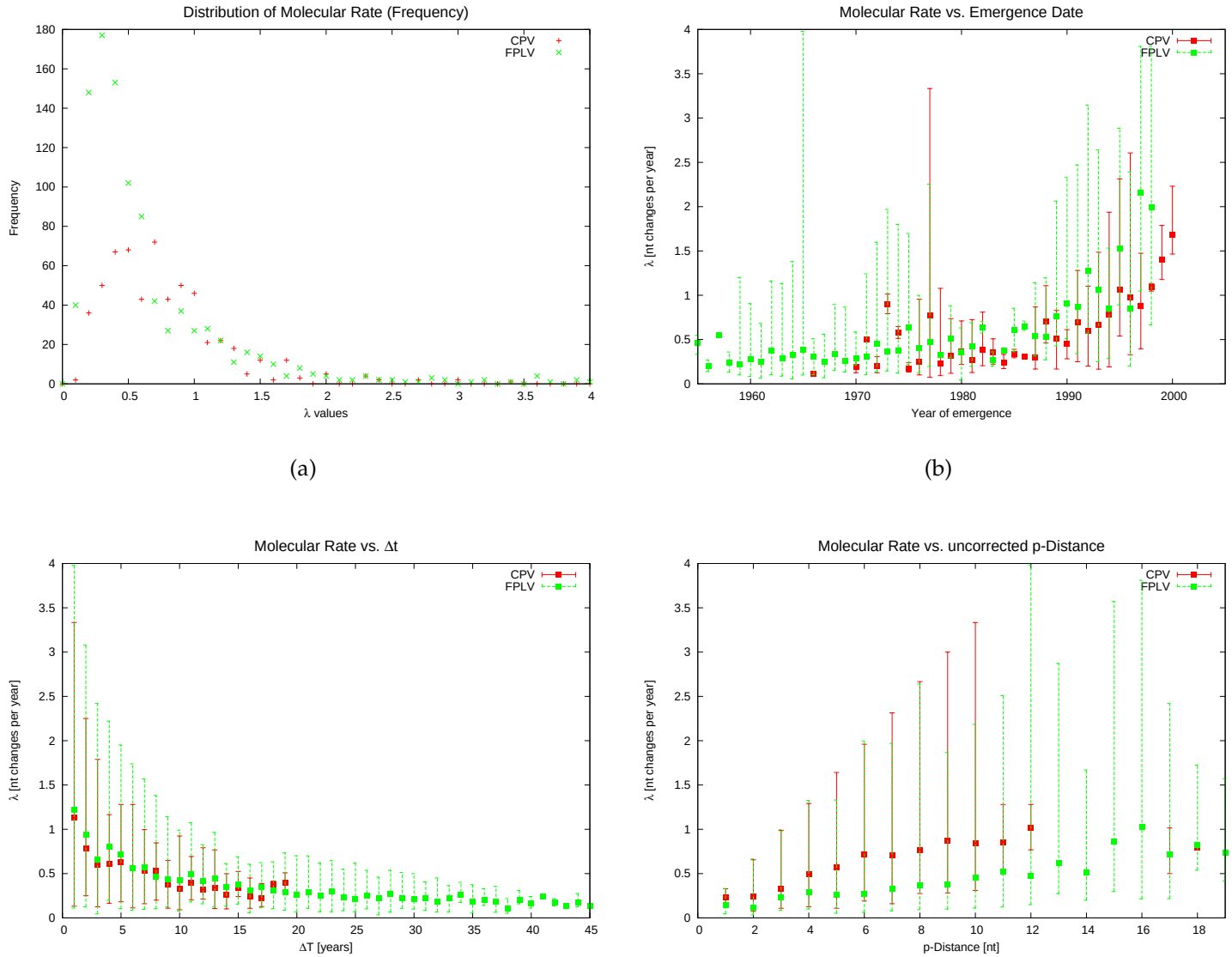


Figure 3.3 Molecular rate for CPV (red) and FPLV (green) species. λ values are plotted as single data points and geometric mean value including error bars indicating the variance. Groups have variable sample size.

relatively constant when extreme values, i.e. very low and high ΔT, are not considered for fitting a linear regression curve to the data (not shown in the figure). λ values have also been fitted using an exponential regression function $y = a^x + d$ with λ to become asymptotic to a value of 0.25 nucleotide changes per year. With FPLV, this correlation becomes even more obvious. Since isolation dates of taxa as indicated in GENBANK files tend to be fuzzy though, naturally the most variation of λ ids found with small ΔT values. As indicated by the error bars in the figure, the variation decreases when values for ΔT become bigger. The decrease of variance has also been found to correlate to a decrease of the number of λ values per ΔT, i.e. the sample size, for a specific ΔT. Values for λ corresponding to ΔT values bigger than 20 are not shown in the figure due to the small sample size of data points (less than five λ_i per ΔT).

λ vs. p-Distance. Finally, panel d in the lower right of Figure 3.3 shows the plot of λ against the uncorrected pairwise distance (p-distance[2]) of 588 out of 657 pairs of taxa considered for λ calculation. Values for λ are expected to be linear with p-distance, assumed that ΔT is constant. This is clearly not the case with our data set considering all possible pairs of taxa with variable ΔT, since a high number of nucleotide substitutions does necessarily correlate to large ΔT values. However, we observe a slight increase of λ with the p-distance for CPV. The slope of the regression curve is lower for FPLV, however. With the variance of λ vs. the p-distance, no correlation to the sample size of the group has been observed. Values for λ corresponding to p-distance values bigger than 12 are not shown in the

[2]In this case, the uncorrected p-distance equals the Hamming-distance, indicating the number of different character states, yielding the number nucleotide differences between two sequences.

figure due to the small sample size of data points (less than three λ_i).

$\lambda_{\mathbf{CPV}}$ vs. $\lambda_{\mathbf{FPLV}}$. The molecular rate for FPLV species has been calculated. On average, it has been determined to be 0.6 times lower than for CPV (0.3645–0.5579). The standard deviation, however, has been shown to be 1.3 times higher than for CPV, indicating a higher variance of λ values. The standard error determined has been shown to be nearly equal in both subsets of data. Characteristic features of the distribution of λ_{FPLV} in contrast to the distribution of λ_{CPV} have been given in given above.

Divergence Time $\mathbf{T_0}$ for CPV. Based on λ and τ/T_0 calculations, the divergence time of CPV has been estimated. The model suggests that the ancestor of CPV has emerged not later than in 1966, i.e. the emergence of CPV is set back more than ten years in the actual timetable. First CPV isolates date back to 1978 and 1979, and the oldest FPLV isolates are from 1960. The host range shift of FPLV to CPV or an intermediate host appears to have happened earlier than assumed until now.

3.2.2 Simulated Data Set

To verify that our model, as described earlier, is applicable to the calculation of the molecular rate and that it is free of conceptual and technical artefacts, we tested it on a simulated data set of artificial nucleotide sequences. The sequences were generated considering CPV specific parameters, e.g. the nucleotide substitution model and substitution rate. Given the disjoint topologies of the CPV and FPLV subtrees[3], we created 60 and

[3]To calculate the rate of nucleotide substitution, we established a NJ tree using the uncorrected p-distance as distance parameter with 1000 bootstrap replicates. The figure of the tree is not shown.

57 sequences, respectively, with SEQGEN and ROSE. The corresponding sequence alignment was then used to create a plain p-distance NJ tree. The molecular rate was re-calculated for CPV and FPLV separately based on the NJ tree.

The ROSE package allows to specify parameters for sequence evolution, including the substitution model, the rate R of transitions to transversions, the average substitution rate etc. A tree topology to generate a complete data set can also be specified. We used ROSE to evolve a randomly generated DNA sequence of comparable sequence length for a given number of rounds, specifying the number of sequences per time interval required[4] to mimic our original data set. From the sampled sequences per round, we chose one sequence to evolve, and the appropriate number of samples required ("sequencing"). We also kept track of the true "history", since sample sequencing did not necessarily include the evolving sequences. The specimens for evolving and sequencing were randomly selected (see Figure 3.2). In test case I, we mimicked our data set with respect to isolation dates, producing identical ΔT values with respect to the original specimens. However, we were assuming a constant molecular rate along the tree. The complete simulated data set consists of 60 taxa. The molecular rate and the corresponding emergence date for each pair of terminal taxa was determined. The statistical analysis of the distribution of λ_i values and λ_i geometric mean values per T_{0i} revealed no correlation between the molecular rate and the emergence date determined by any pair of taxa. Although the rate was specified as a constant value in each round of selection, evolution, sampling and sequencing, the graph-

[4]According to SHANNON's information theorem, we produced twice as many sequences as required for sample sequencing. Half the number of specimens produced were sampled for the phylogenetic reconstruction and molecular rate calculation.

ical analysis shows some minor deviation from a constant function over time. However, a raise in the number of isolates per site does not correlate with large sample sizes. (see also correlation coefficient analysis below). Apart from these observations, the distribution of the molecular rate λ and of λ vs. time[5] shows similar properties to the distribution obtained for the original data set.

In test case II, we used less stringent parameter settings. We used SEQGEN to evolve a root sequence along a given tree topology. In our case this was the topology of the NJ tree used for the initial calculation of CPV molecular rates. We did not mimic our data set with respect to the distribution of the number of isolates per year, but used a constant number of specimens per sampling date instead. No correlation between the molecular rate and the number of equally distributed ΔT and T_{0i} values, respectively, has been detected. The distribution of the molecular rate was found to decrease with fictive emergence dates T_{0i}, whereas the number of specimens per T_{0i} produced a scatter-plot of uniformly distributed data points.

The final test case III considered no specific evolution model at all. The data set was produced assuming a truly random evolution process, using random isolation dates leading to equally distributed values for ΔT. The molecular rate distribution over time could not be fitted with a linear, square, or higher power function, and the distribution of the number of specimens sampled per year was completely random. No significant correlation has been found for T_{0i} λ_i values per specific T_{0i}. All test cases—each using fictive absolute isolation dates—were also evaluated using the

[5]Throughout this section, the term "distribution of molecular rates vs. isolation dates" refers to the distribution of geometric values for λ_i per specific T_{0i} rather than the distribution of all λ_i per specific T_{0i}.

Table 3.2 PEARSON correlation coefficients for molecular rates and emergence date of fictive common ancestors for pairs of taxa

	Original Data		**Simulated Data**		
	FPLV	CPV	CPV I	CPV II	CPV III
Correlation coefficient	0.40	0.60	0.33	-0.01	0.27

linear regression coefficient R^2 for the scatter-plot with respect to the identity function $y = x$, which generally was very low.

We compared the significance of results obtained using a statistical correlation test for λ_i and T_{0i} values. The PEARSON correlation coefficient was determined for the original data subsets. The correlation coefficient for CPV between fictive dates T_{0i} and λ_i for all pairs i of taxa is 0.73 (0.60 for the condensed data set of geometric mean values per specific T_0). For FPLV and MEV, the correlation coefficient is 0.71 (0.40 for the condensed data set). No bias was observed for all artificial data sets. Neither the data set with biased isolation dates and a constant molecular rate, the data set with a constant number of specimens per sampling date along a given topology, nor a perfectly random data set along a given topology revealed a temporally dependent bias for the molecular rate. The use of different evolutionary constraints proves that an increase of λ over time cannot be reliably reproduced by pure chance.

3.3 Molecular Rate Considerations

CPV Evolution Associated with Surprisingly High Rates. This is the first time that an estimate for the CPV molecular rate, based on the VP2 surface antigen sequences of a large set of CPV species, is given. Horiuchi *et al.* [26] have given an approximation for the temporal rate of FPLV

nucleotide variation as $1.1nt/year$, using a geographically restricted and significantly smaller set of 17 FPLV species. We have calculated the molecular rate λ for 60 CPV species as 0.72 ± 0.019 nucleotide changes per year (mean $\pm s.e.$) with a standard deviation of 0.459, which with respect to the sequence length considered, equals a λ of $4.5 \times 10^{-4} nt/year/site$. For FPLV, the molecular rate is $3.5 \times 10^{-4} nt/year/site$. The difference in the FPLV λ values is most likely due to the different sample size and the rough estimation method used by Horiuchi *et al*. Our estimate is higher than a previous estimate for λ based on a smaller sample size of species, given as $1 - 2nt/year$ [74] for CPV only and $4nt/year$ for CPV and FPLV together. It is also higher than the estimate given by Parrish *et al.* [49] as $1.69 \times 10^{-4}/nt/year$. Interestingly, in the study by Truyen *et al.*, the molecular rate is approximately the same for CPV and FPLV, but is about double when CPV and FPLV are considered together for estimating λ. However, no major difference in molecular rates was observed considering either all nucleotide changes or substitutions at phylogenetic informative positions solely. In this work, the estimate for the molecular rate λ is about twice as high than in previous studies. It is possible that this difference is due to the different arithmetic models used. We have set up a simple arithmetic model representing the basic principles of the evolutionary process. It does not take into account multiple substitutions and assumes the molecular rate to be constant for a pair of terminal taxa[6]. It is, nevertheless, more precise than previous models, in which for each taxon, nucleotide differences from the root are simply plotted against the year of isolation. In these cases, the molecular rate was then indicated as the slope of the

[6]The statistical approach of the model compensates for the effect of this assumption. In common molecular clock models, the molecular rate is considered to be constant along an internal branch only.

linear regression curve. This model has several disadvantages, e.g. fitting of a linear regression curve to a scatter-plot or a sigmoid distribution is not possible, and the also, dating of isolates is somewhat diffuse. Our model has been shown to be free of major artefacts, and the significance of results obtains is discussed in the paragraph below.

It must be remembered that surface antigens are assumed to show higher molecular substitution rates than internal antigens e.g. the CPV non-structural protein 1, since they are under significantly higher positive selective pressure. We can, however, compare the CPV VP2 molecular rate to those obtained for other surface antigens. For an RNA virus, e.g. Influenza A, the molecular rate is $6.7 \times 10^{-4} nt/year/site$ [17], whereas for another DNA virus, e.g. the HSV type 1, λ was determined to be $3.5 \times 10^{-8} nt/year/site$ [58]. Interestingly enough, the molecular substitution rate for the DNA virus CPV is close to that of RNA viruses, rather that DNA viruses. It is currently unknown by which mechanism CPV achieves to maintain such high rates[7]. Considering the high molecular rate together with the low sequence diversity for CPV together and evidence for a positive selection force, the occurrence of multiple substitutions that might explain a high number of transitional substitutions is very unlikely. The VP2 gene is presumably rather under another selective pressure, e.g. maintenance of RNA secondary structures, base composition, or another yet unknown property.

Using the Molecular Rate for Back-Dating the Divergence Time. Our model is based on a phylogenetic tree which has been selected as

[7] A major reason for low substitution rates in DNA viruses is the DNA repair system provided by the host cell. Such low error-prone proofreading mechanisms are unknown for RNA targets.

the most stable, i.e. it represents the correct or most likely phylogeny of CPV. Based on all pairwise combinations of taxa, we are able to give an estimated date for every internal node, which is usually supported by more than one dating. Since we are able to consecutively back-date internal branch points, we can also back trace CPV to its most distant ancestors. We have determined that CPV has emerged before 1966, i.e. more than ten years earlier than its actual appearance in canine hosts as supported by first pathogenic isolates from 1978. The dating of CPV emergence is equally important for solving the question what source it has emerged from. Our phylogenetic and molecular analyses include CPV species which infect various other carnivores, e.g. racoons, lynx, foxes, and wolves. The difference between CPV emergence and its appearance in canine hosts may support the hypothesis of a putative carnivore intermediate host although no sequences and tissue isolates are yet available from such carrier.

The Simple Model Reveals CPV Evolutionary Dynamics. We have shown that CPV (1) evolves with a generally higher evolutionary rate than FPLV (2) shows an obvious pattern of epidemic and endemic phases in dog populations (3) molecular rates show a gradual increase over the last few years. We observed a high rate of molecular substitutions associated with both the emergence and recent appearances of CPV. Together with the analysis of transitional vs. transversional and synonymous vs. non-synonymous substitutions, we found support that CPV is under weak, yet obvious positive selection as it has been shown for Influenza A [11, 16]. In contrast, the results for FPLV support the feline HRV is subject to neutral evolution, evolving by random genetic drift. The conclusions about

evolutionary dynamics of CPV in contrast to FPLV are well supported by molecular rates, and the rate of synonymous and non-synonymous replacements, as well as by a recent study using a far more sophisticated coalescent evolution model [59].

The dating of isolates represents a minor problem though, since our model requires to calculate ΔT for a given pair of taxa. However, a ΔT value of one rarely indicates that both isolates were collected twelve months one after another. In the worst case were collected $1 - 23$ months after another, i.e. ΔT may take values from $0.08 - 1.92$. Furthermore, the dating of an isolate does not take into account whether the isolate was taken from an acute infection, so the virus has possibly already been passaged in the host for several times by the time it was isolated. It may likewise not be the first actual occurrence of a viral species but the first observed occurrence instead, leading to some mis-annotation of isolation dates. It must therefore be remembered that the isolation date does not necessarily correspond exactly to the emergence date of a viral species. Also, the analysis of the molecular rate distribution over time considers the emergence dates T_0 of the common ancestor of a given pair of taxa. The emergence date, as well as the common ancestor it corresponds to, are fictive. Temporarily dependent artefacts in determination of λ do, however, not affect the quality of our results, as shown by the use of artificial sequences.

3.4 Tools

Calculation the Molecular Rate. The UNIX command line tool providing raw data for the calculation of the molecular rates was implemented

in C/C^{++} (`treePather`, see Appendix E).

Statistical Analysis. Statistical analyses, correlation analyses, and paired sample tests were performed using the SPSS statistics package [63].

Chapter 4

Structural Implications

"Form ever follows function."
Louis Henri Sullivan

4.1 Structural Mapping

The major questions arising from the phylogenetic analysis of the CPV VP2 sequences was whether it is possible to determine the temporal succession of single amino acid replacements in the evolution of the VP2 protein and whether a correlation between amino acids substitutions and the antigenic and geographical branching pattern described above can be deduced from the phylogeny. Furthermore, we were interested whether the specific positioning of certain taxa within the protein tree can be explained by taxon-specific replacement of residues involved in determining the host range. Therefore, we investigated the parsimony informative sites and singleton sites[1] in the protein alignment and established a correlation to the protein tree obtained. The complete multiple alignment of 120 amino

[1] A site is parsimony-informative if it contains at least two types of nucleotides (or amino acids), with at least two of them occurring with a minimum frequency of 2. A site is called a singleton site if it contains at least two types of nucleotides (or amino acids) with at most one of them occurring multiple times.

Table 4.1 Specific nucleotide and amino acid replacements in the VP2 gene and protein sequences of selected taxa.

	CPV			**FPLV**	
	RedFox	2c U51	Quinn	BlueFox	Quinn*
Nucleotide	557 928 958	208 1278	176 219 898 1562	864	658 1039 1535 1529 1680
Amino Acid	N186 R310 R320	Y70 E426	G59 I73 S300 N521	none	S220 T347 R512 D517 K560

acid sequences spans 531 of the total 585 positions of the VP2 protein. A total of 82 sites is variable, 36 sites of which have been determined parsimony informative and 46 sites are singleton sites. Basically, the diversity of the CPV-FPLV sequence data set has been determined to be quite low (6.8%) with an approximately twice as high proportion of total variable sites (15.5%). We split the data set into two sets according to their host range specificity (CPV and FPLV, respectively), and analysed the clustering of taxa within those distinct subsets. We also counted the number of variable sites in the FPLV and CPV sets separately: the total number of variable sites in the CPV clade is 40, whereas the total number of variable positions in the FPLV set has been determined to be 47, i.e. a 4% higher number of sequence differences.

At the terminal branches within the CPV set, we found that three species considered as LCPV ("LCPV203", "LCPV139", "LCPV140") show a specific residue at position 300. They are differing from another 30 CPV-2a, 2b, and 2c species in this residue solely, replacing the 2a-specific glycine

with an aspartate. The glycine at position 300 has previously been shown to occur in 2a subtypes and its antigenic variants only but not in the original CPV-2, which has an alanine at position 300 in the VP2 protein. Together, these 33 taxa represent a subclade of many Asian sequences (75%) and a few European sequences, but no sequences from neither Africa nor America. The rest of CPV-2a and CPV-2b species shows the specific glycine at position 300 that distinguished between CPV-2 and its antigenic variants. We found another three terminal CPV species with a specific residue, namely a proline at position 265. This residue has not been characterised before with respect to its potential antigenic properties. However, CPV species showing this replacement have been identified to be prevalent in Italian dog populations [7, 8]. No other taxa show this residue, but the canonical threonine instead. CPV species "W24", "U486" and "616" differ in no other position from the rest of the data set. We observed that one species ("W24") has been isolated from a wild carnivore (*C. lupus*), but we could not yet verify whether the other two species were also isolated from wild carnivores. Moving one level of branching points closer to the root, we observed another amino acid replacement corresponding to a stable branching pattern. This branching pattern was significant to separate Asian sequences from American sequences and African, with sequences from Central Europe distributed over the two branches. We determined that all Japanese taxa[2] show an alanine at position 297, whereas US sequences[3] show a serine residue. It has been show in earlier studies that the replacement S297A has become more frequent in recent CPV isolates[4] [7, 8]. The European sequences show either of these residues. The

[2]This is a generalisation of sequence origin solely, roughly summarising taxa from a large geographical area with respect to the most frequent country.

[3]Same as above.

[4]A hypothesis for the role of residue 297 is that it slowly, but gradually becomes re-

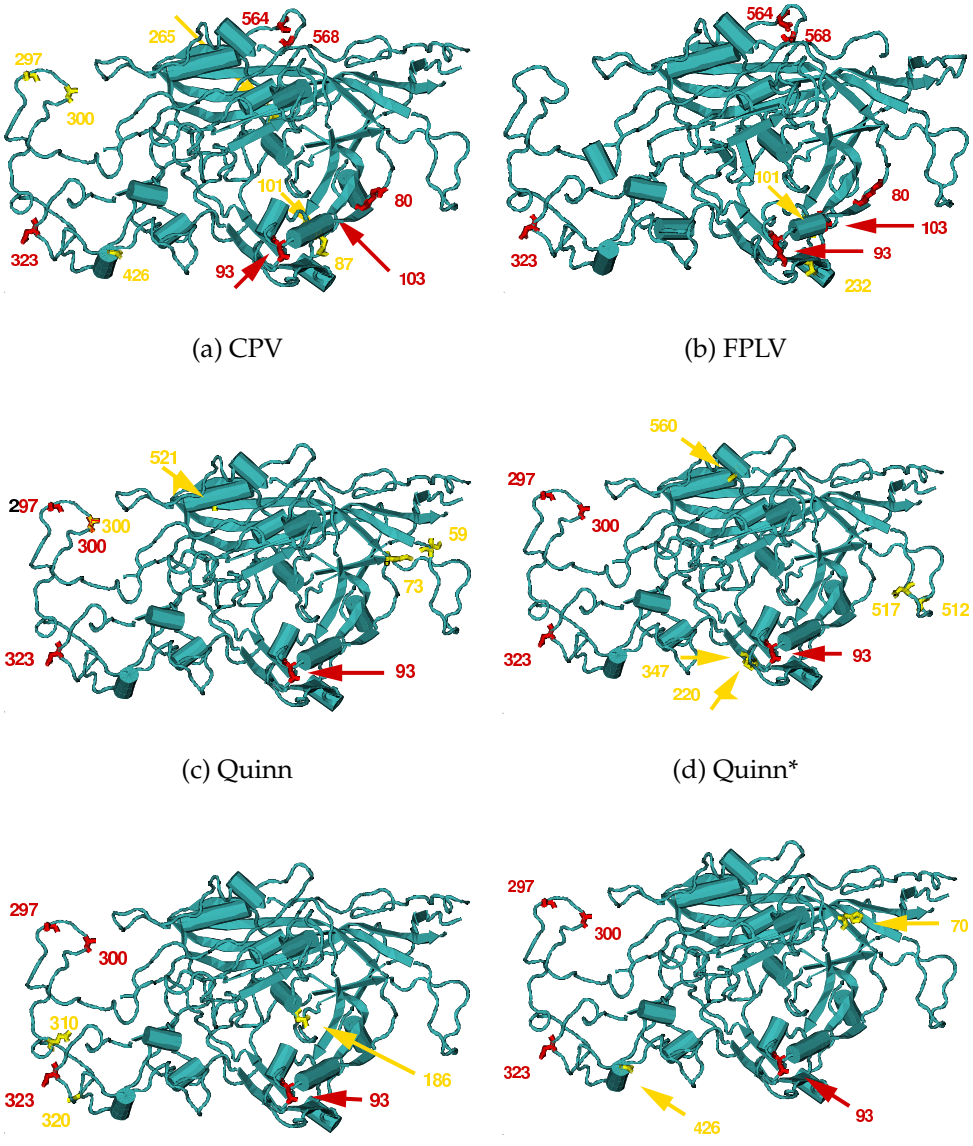
isolated single taxon "U51" also shows an alanine, as well as "Thai30", a CPV-2 species. Finally, we determined the amino acid substitutions causing the branching of 13 CPV-2 type species. They are differing in two residues from the rest of the CPV set. First, they show a methionine at position 87, which was replaced by a leucine that has previously been shown to be specific for the initial shift from CPV type 2 to CPV type 2a. Two CPV species ("CP49", "OB1") have gaps at this position due to partial sequence availability solely and cannot be classified exactly. Second, the same 13 taxa have a specific isoleucine at position 101 which has been replaced with a threonine in the later CPV-2a type. Residue 87 has been previously shown to be responsible for the antigenic shift, whereas residue 101 has not been antigenically characterised yet. However, we used CSU [61] to determine interaction residues and we found that residues 87 and 101 interact directly within the asymmetric unit. The interaction explains why both residues are changed and together are a pair of phylogenetic important residues. Isolated taxa "U51" an "CP49" in the protein tree (see Figure A.4) have been analysed for showing specific residues, but only CPV-2c "U51" has a specific tyrosine at position 70 that distinguishes it from the rest of the CPV species showing a histidine at position 70. In addition, we determined that U51 belongs to novel antigenic subtype 2c, which has only been isolated and described in Vietnam [40] so far. We also found another two taxa ("56", "695") with a 2c-specific glutamate at position 426, which has previously been shown to distinguish between antigenic subtypes 2a (asparagine) and 2b (aspartate). CPV species "CP49" has no specific residues but shows gaps at positions identified to correlate

placed in dog populations worldwide due to an unknown selection mechanism; another hypothesis considers geographical differences in dog host populations that might explain the split at position 297 among CPV subtypes 2a and 2b according to their area of isolation.

with the branching pattern. The specific amino acid replacements of selected taxa is shown in Table 4.1. The amino acid replacements were also detected as non-synonymous nucleotide changes when the same analysis was performed for the set of 66 CPV coding sequences (data not shown). Within the nucleotide tree, the branching pattern was correlated to non-silent replacements solely. Only FPLV species "BlueFox" showed a significant silent nucleotide replacement at position 864 in the VP2 gene.

The same analysis was performed for the FPLV clade. However, only one significant amino acid replacement could be assigned to the topology. The first branching point from the root separates 19 FPLV taxa with an isoleucine at position 232 from another 35 taxa showing a valine at the same position. Although the branching pattern observed within the FPLV clade looks reasonable, the subsets separated by the character state at position 232 do not have any other significant commonness. The smaller subset, however, shows a larger proportion of older FPLV species and taxa with exotic features like vaccine strains. Residue 232 has not been assigned an antigenic property so far. There is also no substitution that distinguishes FPLV strains from MEV strains. No other specific replacement shared by subclades were determined in our analysis based on the VP2 effector sequence tree. Indeed, only one terminal subclade comprising 12 taxa showed a common specific residue, namely isoleucine at position 101, which is also shared by four other taxa located in different subclades.

We subsequently mapped those residues identifies as phylogenetic relevant onto the structure of the asymmetric unit of the canine and feline biological molecule, respectively. The mature virion consists of 60 VP monomers, together forming the icosahedral capsid. For convenience, single amino acid substitutions were mapped onto the monomer structure.



(a) CPV

(b) FPLV

(c) Quinn

(d) Quinn*

(e) RedFox

(f) 2c U51

Figure 4.1 Structural mapping of residues on the asymmetric unit. **Top** Mapping of residues correlating to phylogenetic branching patterns. **Middle** and **Bottom** Mapping of taxon-specific substitutions of residues on to the VP2 structure. FPLV/CPV host range determining residues are shown in red, and taxon-specific residues are shown in yellow.

Their relative location to each other in the mature particle is described below. Figure 4.1 shows the positioning of the relevant single amino acid substitutions in the VP2 protein of CPV and FPLV, respectively. Residues previously identified as responsible for the host range shift are shown in red, whereas residues identified to cause a particular branching pattern are shown in yellow. Residue 101 showing a CPV specific substitution is found right next to the 2/2a determining position 87 and close to the FPLV/CPV determining position 103. Positions 80 and 93 which also discriminates between FPLV and CPV are less close to CPV-specific positions 87 and 101. We also identified residue 265 as phylogenetic relevant for CPV. This residue is located on the "back side" of the molecule[5] right at the start of a large β strand region. Due to the symmetrical assembly of many molecules, residue 265 is also exposed to the viral surface. Positions 297 and 300 which have been shown to feature CPV specific substitutions are located in the same loop on the opposite side of the molecule. Residue 300 has been shown to discriminate between CPV-2 and the subsequent 2a subtypes, and residue 297 identified in our phylogenetic analysis is located right next to residue 300. The figure also shows position 426 which discriminates between three CPV antigenic subtypes, 2a, 2b, and 2c, at the tip of a surface exposed α helix. There are less residues that have been identified for FPLV, and their location within the VP2 monomer is shown in the right panel of Figure 4.1. One identical position to CPV, residue 101, has been identified. The second position, residue 232, is FPLV specific. It is not immediately next but very close to other FPLV/CPV determining positions 93 and 103.

[5]The term is used with respect to the two-dimensional represnatiation of the three-dimensional molecule in Figure 4.1.

CPV vs. FPLV "Quinn". The sequence of parvovirus species Quinn assigned as CPV was obtained from GENBANK (accession number `AJ002929`). However, the taxon was found to fall into the FPLV clade in the phylogenetic trees several times. We subsequently analysed CPV Quinn for FPLV and CPV specific changes in the VP2 sequence using `vdiff`. The results proved that CPV Quinn indeed shows five FPLV-specific residues but no CPV-specific amino acids. We conclude that in fact, CPV Quinn is a FPLV species and has mistakingly been assigned to CPV. The original CPV Quinn sequence was then extracted from original sequence and alignment files and analysed for CPV specific residues. Both sequences were compared and shown to be non-identical. In contrast to the sequence deposited to GENBANK, the sequence from original files showed CPV-specific residues. We did not perform further phylogenetic analyses including the original CPV type "Quinn" for convenience. The sequences have obviously been interchanged, explaining the appearance of a CPV species within the FPLV clade. We were able to map the specific residues of selected taxa onto the VP2 structure as shown in figure 4.1 and evaluate their relative positions to antigenic relevant epitopes. The original CPV "Quinn" has been assigned four specific changes, one of which at position 300. CPV "Quinn" has a specific serine at the position that determines the canine vs. feline host range. This residue is located near the canyon. Two other residues (G59 and I73) are located in a loop close to the three-fold axis, whereas residue N521 is located on a beta sheet close to the FPLV/CPV host range determining region (residues 564 and 568 in alpha helix) near the five-fold axis. The FPLV species "Quinn" obtained from GENBANK however, has five specific amino acid replacements. One of them, K560, is found in the same region as the CPV "Quinn"

specific residue N521. The serine at position 220 and the threonine at position 347 on the other hand are found closely to each other in a beta sheet near the FPLV/CPV determining region of residue 93. Two other replacements, R512 and D517, are located on the very tip of the same loop as CPV "Quinn" specific residues G59 and I73 described above. The specific amino acid replacements in the VP2 protein of the CPV isolate from a "RedFox" are shown in the lower left of the figure. We only found three specific residues for the isolate from a wild carnivore, an asparagine at position 186 which is located at the tip of a small loop, and two residues, K310 and R320, close to the FPLV/CPV determining epitope around residue 323. Since the sequence from the red fox isolate is only about 50% of the total sequence analysed, we were not able to determine other residues that might be specific for a wild carnivore intermediate species. Finally, figure 4.1d shows the specific amino acid replacements of the CPV subtype 2c ("U51"), the 2c-specified glutamate at position 426 and U51-specific tyrosine 70.

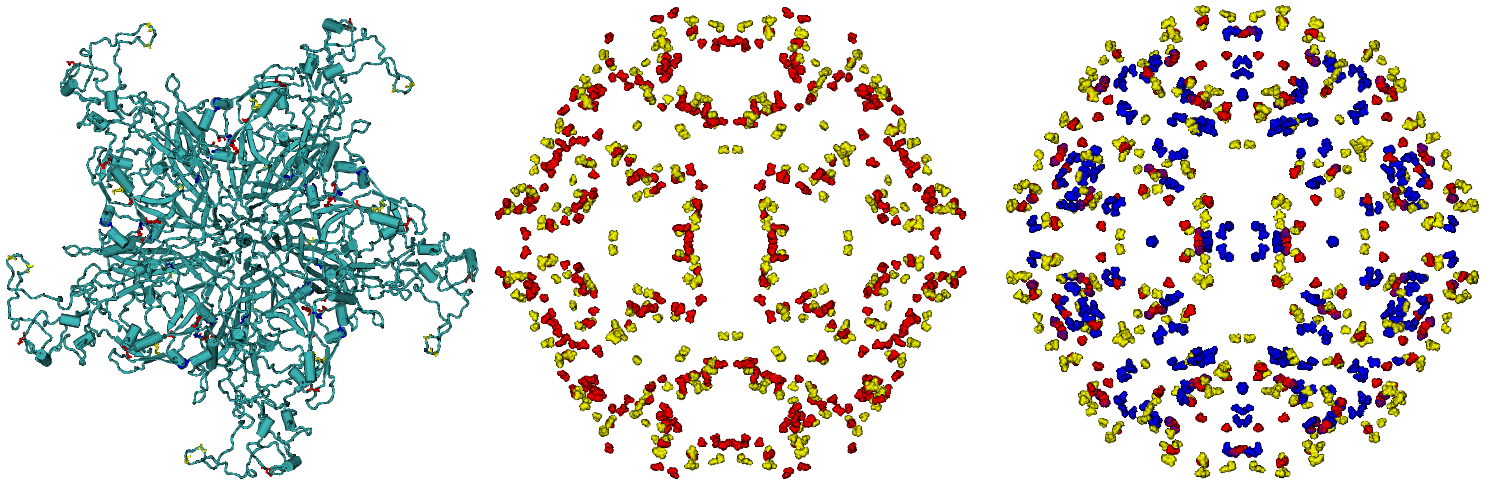
Interacting Residues. We also analysed the position of CPV-specific and taxon-specific residues with respect to the interaction of monomers. Figure 4.2b shows the arrangement of residues identified in our replacement-based phylogenetic analysis (in yellow) relative to the host-range determining residues (in red). The three-dimensional pattern in the distribution of host-range determining residues on the viral surface is obviously corresponding to the geometrical organisation of the coat protein. The residues identified in our approach are following a non-identical, but partially similar distribution pattern on the surface. Here we show that those five residues are in fact in close contact with the six host-range deter-

mining residues, together forming crucial epitopes for virus-cell interaction (see the close combinations of yellow and red residues in the figure). Some directly interacting residues have also been determined using VMD; residues 48, 80, 238, 248, 558, and 565 are shown in blue in Figure 4.2.

FPLV vs. MEV. The classes MEV and FPLV are assigned to viral specimens according to their occurrence in either mink or feline hosts [6]. They are considered different subtypes of Feline Parvoviruses. However, we found no evidence for distinct lineages of MEV and FPLV in neither phylogenetic trees, phylogenetic networks, nor in character substitution analysis using `vdiff`. Hence, FPLV and MEV species may be considered equal viral types, in contrast to CPV subtype 2 and its descendant lineages. The indication that MEV species may share a common feature that has not been revealed by phylogenetic and sequence analyses, as mentioned earlier, was not confirmed. However, a group of seven out of a total of eight MEV species share a common substitution at nucleotide position 1410 with another eleven FPLV species, as revealed by `vdiff`. The split is indicated by an arrow (dark yellow) in Figure 2.2. Those 19 species show a common silent replacement of a thymidine with a cytosine. Both resulting codons `TAC` and `TAT` encode for the tyrosine at position 470 in the VP2 protein.

4.2 Structural Considerations

Altered Host Range Properties Through Single Point Mutations. It is obvious that neither phylogenetic nor sequence analysis alone leads to a comprehensive knowledge of the evolutionary process. In the case of Canine and Feline Parvovirus evolution, phylogenetic trees and character



(a) (b) (c)

Figure 4.2 Structural Mapping of residues on the biological unit. **Left** Combined view of three CPV monomers indicating the interacting residues. (Colour code same as in figure 4.1, residues interaction with other monomers are shown in blue). **Middle** and **Right** Projection of the distribution pattern of phylogenetic and antigenic relevant residues only onto the viral surface. The rest of the biological unit is not shown.

substitutions have been intensively studied. The use of tools for analysis of molecular sequence data provided valuable insight into the history of CPV. However, antigenic replacement by antigenic shift and antigenic drift, as it readily occurs in CPV populations, plays an important role. In contrast, random genetic drift is considered the major driving force in FPLV populations. Unfortunately, we were lacking an available and easy-to-use "power-tool" to answer the question which specific replacement has occurred along an internal edge or at an internal branching point in the phylogenetic tree. So we implemented a straightforward tool to assign sequence properties to specific properties of the corresponding topology. This tool was not only helpful to achieve a more detailed resolution of CPV phylogeny as described in a earlier, but also provided data for mapping of amino acid replacements onto the structure of the surface antigen. Those residues identified to have occurred consecutively during the evolution of CPV are not only the same residues which have been identified to be involved in antigenic subtype-specific host cell targeting. In contrast, we detected three more residues with phylogenetic, antigenic, and semantic significance. The function of those residues has not been determined exactly yet, but is currently being studied.

Mapping of all residues corresponding to distinguished elements in the phylogenetic tree topology showed that (1) the distribution of antigenic relevant residues on the biological unit follows a strict geometric pattern similar to the highly geometric configuration of monomer units (2) the distribution of phylogenetic relevant residues on the biological unit follows a non-identical, yet similar geometric pattern with a comparably high degree of organisation (3) evidence for direct interaction of antigenic, phylogenetic and character-state relevant residues has been found. We con-

clude that residues that have not been identified previously as decisive for the specific behaviour of antigenic subtypes are equally, yet possibly indirectly, involved in virus-cell-interactions. The high-order geometric distribution of specific replacements on the spheric surface does not only correspond to the highly structured composition of the virion from 60 subunits. It also emphasises the importance of single non-silent point mutations involved in cell targeting, since the host targeting process occurs in a highly concerted manner.

An Intra-Molecular Compensatory Mutation? The identification of a kind of compensatory mutation[6] of residues 87 and 101 was confirmed by analysing intra-monomeric interactions. Since residues 87 and 101 are in immediate contact, interacting via a covalent bond, we can explain why (1) in antigenic replacement experiments, only residue 87 has been identified and studied because it is surface-exposed, yet residue 101 is not, and (2) in our analysis using a global character-state based approach, both residues 87 and 101 were identified at an early branching point with equal significance.

It is, actually, unknown which mutation might have occurred previous to the other. However, here we hypothesise two scenarios: (I) If residue 87 altering the host range has occurred first, the change in the structural conformation of the capsid allowing to target different host cells possibly could only be maintained if also residue 101 has changed. The replacement at position 101 then can be considered a truly compensatory mutation to maintain the epitope conformation. (II) If residue 101 mutated by random replacement, the compensatory replacement at position 87, or a

[6] A compensatory mutation is conserving a functional structure of the molecule in either RNA or proteins by replacement of the interacting nucleotide or amino acid.

hypothetical parallel replacement of residues 87 and 101, might have introduced a structural feature to the capsid altering its host range properties. The structure then would act as a selective advantage, accelerating the host range shift from CPV-2 to CPV-2a. The host range shift then would be considered an indirect consequence of the replacement at position 101. However, considering that (1) residue 87 has plays a more important role as surface-exposed residue, introducing novel antigenic subtype 2a, and (2) the antigenic shift from CPV-2 to CPV-2a has occurred very quickly in less than a year, we favour scenario I. If in a more detailed analysis, evidence for a purifying selection force on residue 101 can be found by means of synonymous vs. non-synonymous replacements, we will be able to confirm that the replacement of residue 101 is actually due to a compensatory mutation.

4.3 Tools

Identification of Taxon-Specific Substitutions. Specific nucleotide and amino acid replacements corresponding to specific data (sub-)sets were identified using `vdiff`. Taxa or groups of taxa of special interest, i.e. with respect to antigenic or phylogenetic clustering, were selected after analysis of phylogenetic trees and subsequently analysed with `vdiff`.

Structural Mapping. Phylogenetically relevant positions identified by `vdiff` were mapped on the viral coat surface structure using the VMD 3D molecular modelling tool [30] Version 1.8.2. The structure of the FPLV and CPV molecules was retrieved from the Protein Database (PDB) [3, 9] (accession numbers `1C8E` [60] and `4DPV` [84]). Files included coordinates for both the asymmetric unit, i.e. VP2 monomers, and the biological unit, i.e. the fully assembled viral capsid consisting of 60 asymmetric units.

Interaction of Residues in the Biological Unit. Inter-molecular interacting residues, i.e. residues interacting with residues in another VP2 unit, were identified based on the X-Ray crystallography coordinates of the biological unit [78, 84]. Analysis of selected monomer frames was performed using VMD.

Interaction of Residues in the Asymmetric Unit. Intra-molecular interactions between residues of the same monomer were identified using CSU [61].

Chapter 5

Summary

"What if this weren't a hypothetical question?"
Unknown

5.1 Concluding Remarks

We shed some light—although, still, incompletely—on some peculiarities of CPV evolution. At a molecular level, we analysed the evolution of Canine Parvovirus (CPV) and its closely related variant, Feline Panleukopenia virus (FPLV). Our results show that the consecutive non-silent replacement of a few nucleotides can be tracked in the phylogenetic tree we established based on the sequence of the surface antigen. In addition to a handful of residues that have been described previously to be involved in the original host range and subsequent antigenic shifts, we presented evidence that a few more residues are likely to play a crucial role in CPV evolution. The biological function of those amino acid replacements at specific positions is unclear, but e.g. residue 297 is currently being studied in great detail to explain its role in host cell targeting by interaction with the host cell receptor. The results of our analyses show that residues

of phylogenetic and antigenic relevance can easily be identified, thereby providing novel starting points for genetic and biochemical analyses of mutants. This approach combining sequence and phylogentic analyses is neither limited to CPV and nor to virus populations, but instead, can be applied to any problem using molecular sequence data. The phylogeny established for the surface antigen served as the basis to determine the rate of nucleotide substitution. We observed that the molecular rate differs in CPV and FPLV populations, and that different selective forces are acting on the host range variants as indicated by the rate of silent and non-silent changes of the viral coat gene. Our results add further support to the hypothesis that FPLV undergoes neutral evolution, whereas CPV is under positive selection and its epidemic pattern includes both phases of fast and slow mutation. With our simple model for calculating the molecular rate, we observed that CPV shows different molecular rates in phases of host range or antigenic shift and adaption to the host. Similar results were obtained by other groups using more sophisticated models of nucleotide sequence evolution. The recent increase in the rate of nucleotide substitution is possibly correlating to the appearance of novel antigenic subtype CPV-2c in Vietnam and also, in Europe. Another hypothesis considers the currently ongoing amino acid replacement at position 297 in the protein chain, which to date has been identified to be prevalent in dog populations worldwide. However, the exact dynamics of how and when CPV is changing its molecular rate remains unclear. Artefacts in our model were ruled out confirming our results using a set of artificial nucleotide sequences. With our approach to identify arbitrary subtree-specific nucleotide or amino acid replacements, we were able to map taxon-specific residues of the surface antigen on the three-dimensional structure of both

the asymmetric monomer and the icosahedral biological unit. We showed that taxon-specific replacements are located close to residues of surface exposed structures involved in receptor targeting and residues in sites interacting with other copies of the molecule.

5.2 Outlook

We will look for confirmation of the hypothesis that the current changes in antigenicity at single sites are the reason for the rise in molecular rates. In this context, a major topic is the refinement of the model for the calculation of molecular rates. It will be an interesting task to be investigate the dynamic evolution pattern of CPV, allowing to consider different molecular rates at different times. Since CPV evolves with a rate similar ro RNA viruses rather than other DNA viruses, it will be important to learn more about the mechanism maintaining a high nucleotide substitution rate. Ideally, we should be able to describe the evolutionary dynamics of CPV more precisely. With newly sequenced samples available, we will be able to analyse an even larger data set, emphasising on the appearance of novel CPV subtype 2c. It is currently unknown which mechanism was responsible for the antigenic shift of CPV subtype 2a to subtype 2b, as it is unknown for novel subtype 2c. The pattern of independent emergence of identical subtypes is, however, more likely due to selective forces than to a transmission mechanism between different hosts. It will also be interesting to have a closer look at distinct outbreaks and to establish a guideline for the assignment of DNA sequence samples to isolates from such "epidemic bursts". Finally, re-sequencing of CPV "Quinn" samples will resolve the problem of contradicting CPV vs. FPLV sequence assignment this strain

showed in our phylogenetic analyses, and the corresponding GENBANK file will be corrected. In addition, re-sequencing of the large viral protein sequence or laboratory experiments are supposed to resolve the question of the two different splicing variants.

Acknowledgements. I want to thank all people who were, explicitly and generally, contributing to the successful completion of my diploma thesis:

- PF Stadler for excellent academic supervision,
- U Truyen for providing unpublished sequence data and valuable background information about Canine Parvovirus evolution,
- G Fritzsch for 24h phylogenetic emergency stand-by and superb mobile help desk performance,
- I Hofacker for help with structural modelling tools,
- my husband for critical remarks from the computer scientist's department as well as distraction—whenever either of those was necessary,
- my sister for providing expertise in statistical analysis and a role model for persistence,
- everyone at IFI and IZBI Leipzig for social and culinary support, and
- my whole family for their confidence.

Vienna, December 2004

Appendix A

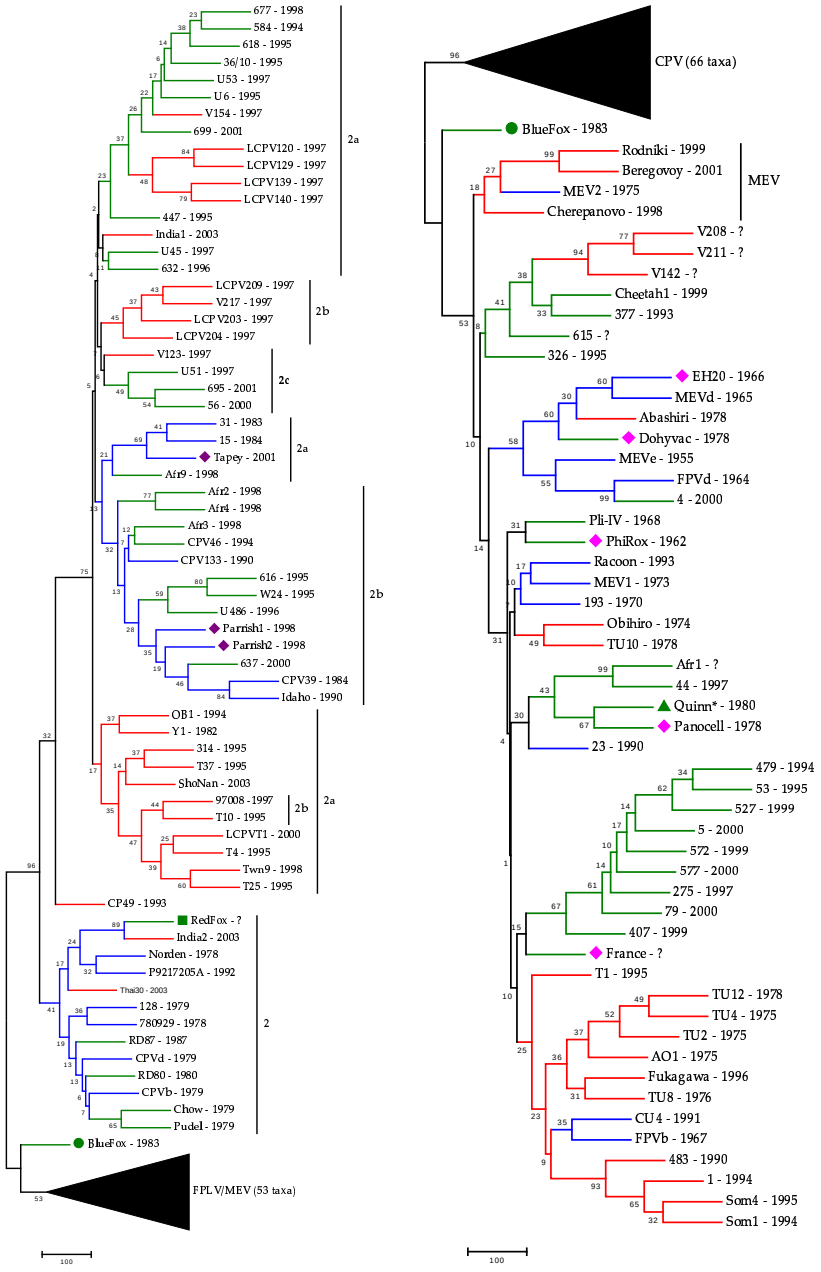
Phylogenetic Trees

A.1 Legend to the Figures

On the following pages, the original Neighbor-Joining, Maximum Parsimony and Maximum Likelihood trees based on the coding (figures A.1–A.3) and effector sequence (figure A.4) of the CPV and FPLV partial VP2 gene are shown. The branches in the tree are coloured indicating their geographical origin.

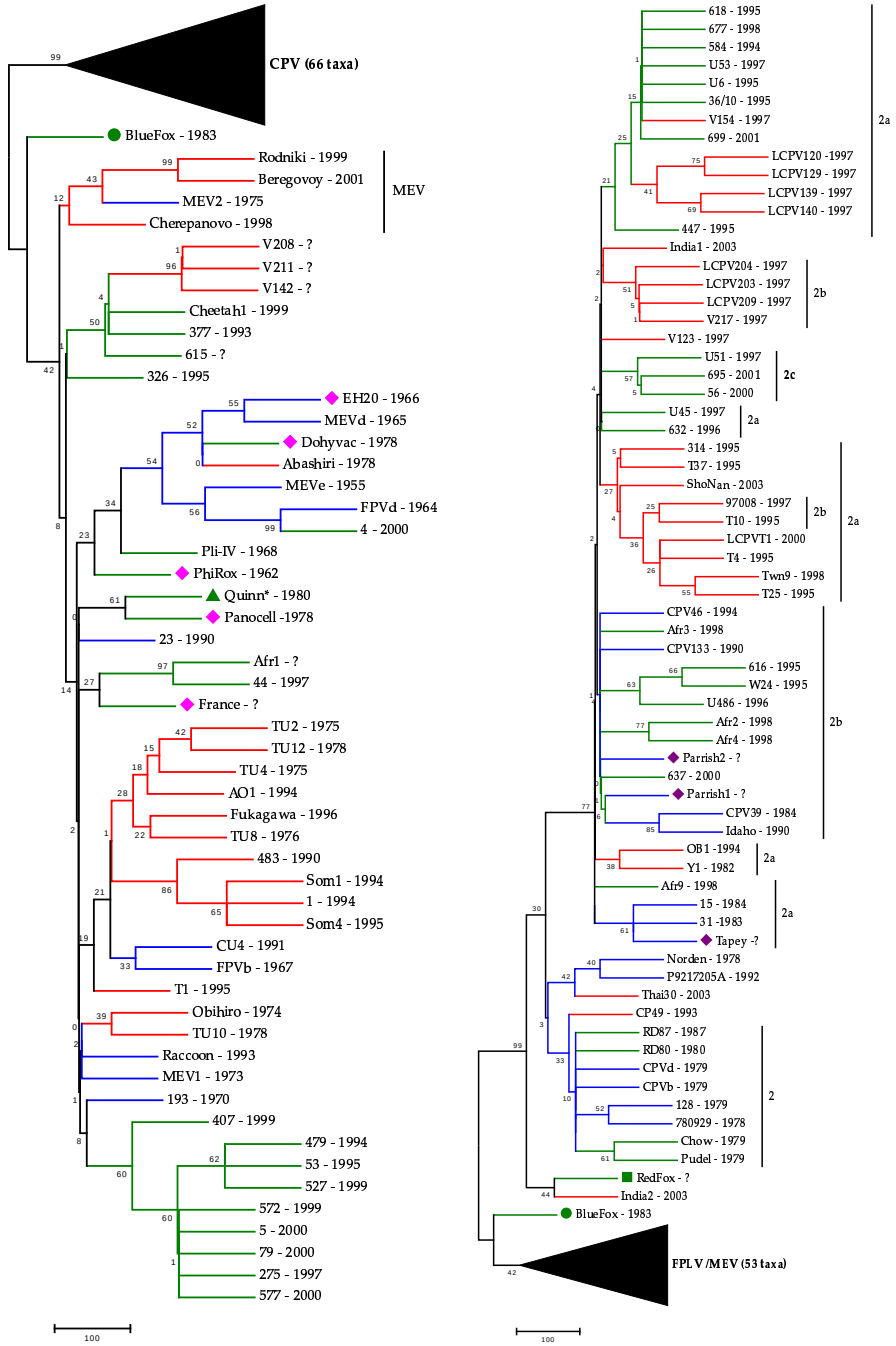
Red branches correspond to sequence origin from Asia (Japan, China, Vietnam, Taiwan, and Thailand). Blue branches indicate sequences have been isolated in the United States (and one, in Australia, respectively). Green branches are corresponding to sequences from European (Germany, France, Finland, UK, Italy, Poland, and Soviet Union) and South African species. FPLV and CPV vaccines are coloured according to their geographical origin and indicated by an additional pink or violet taxon marker (filled diamond $\diamond$). Black lines indicate sequences with unknown origin. The filled triangle ($\triangle$), the filled circle ($\circ$) and the filled square ($\square$) mark taxa of special interest: the CPV-misassigned FPLV "Quinn", and the FPLV-CPV intermediates "BlueFox" and "RedFox".

A.2 Coding Sequence Trees



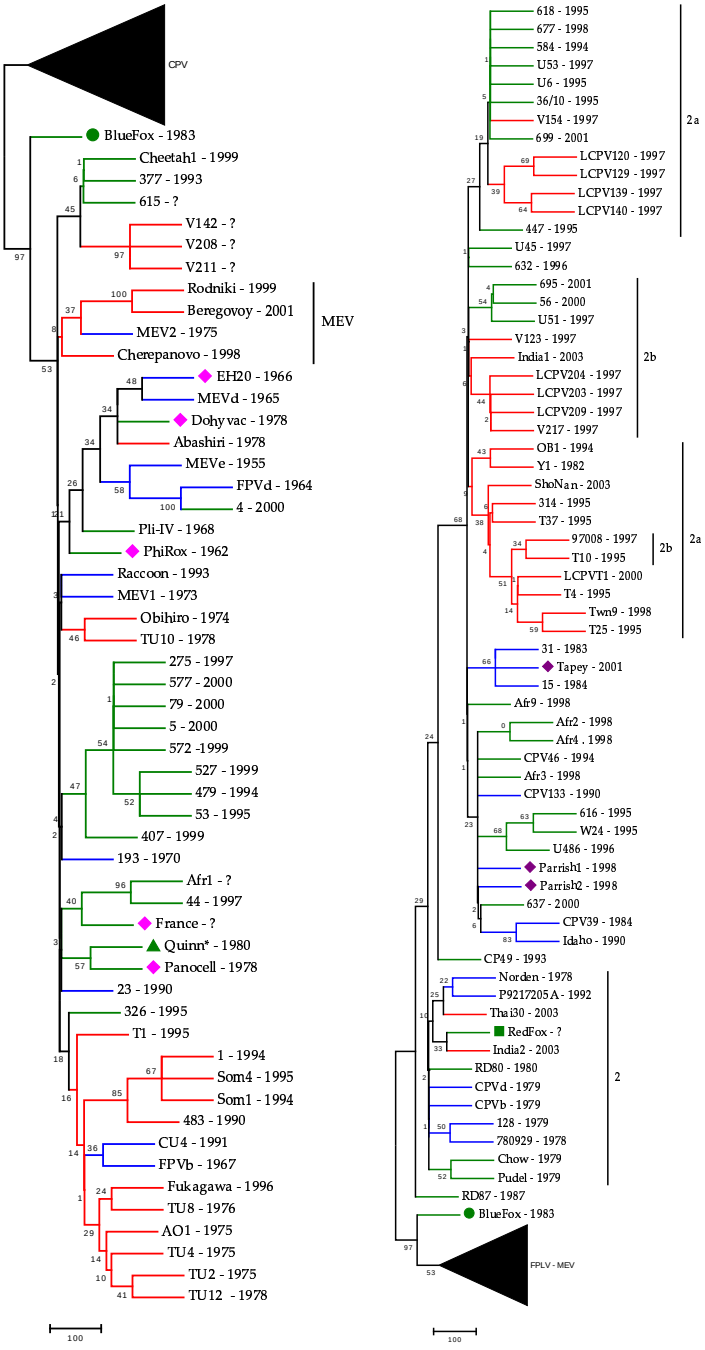
(a) CPV (b) FPLV

Figure A.1 Neighbor-Joining tree for the partial VP2 gene sequence. **Left** CPV **Right** FPLV



(a) CPV (b) FPLV

Figure A.2 Maximum Parsimony tree for the partial VP2 gene sequence. **Left** CPV **Right** FPLV



(a) CPV (b) FPLV

Figure A.3 Maximum likelihood tree for the partial VP2 gene sequence. **Left** CPV **Right** FPLV

A.3 Effector Sequence Trees

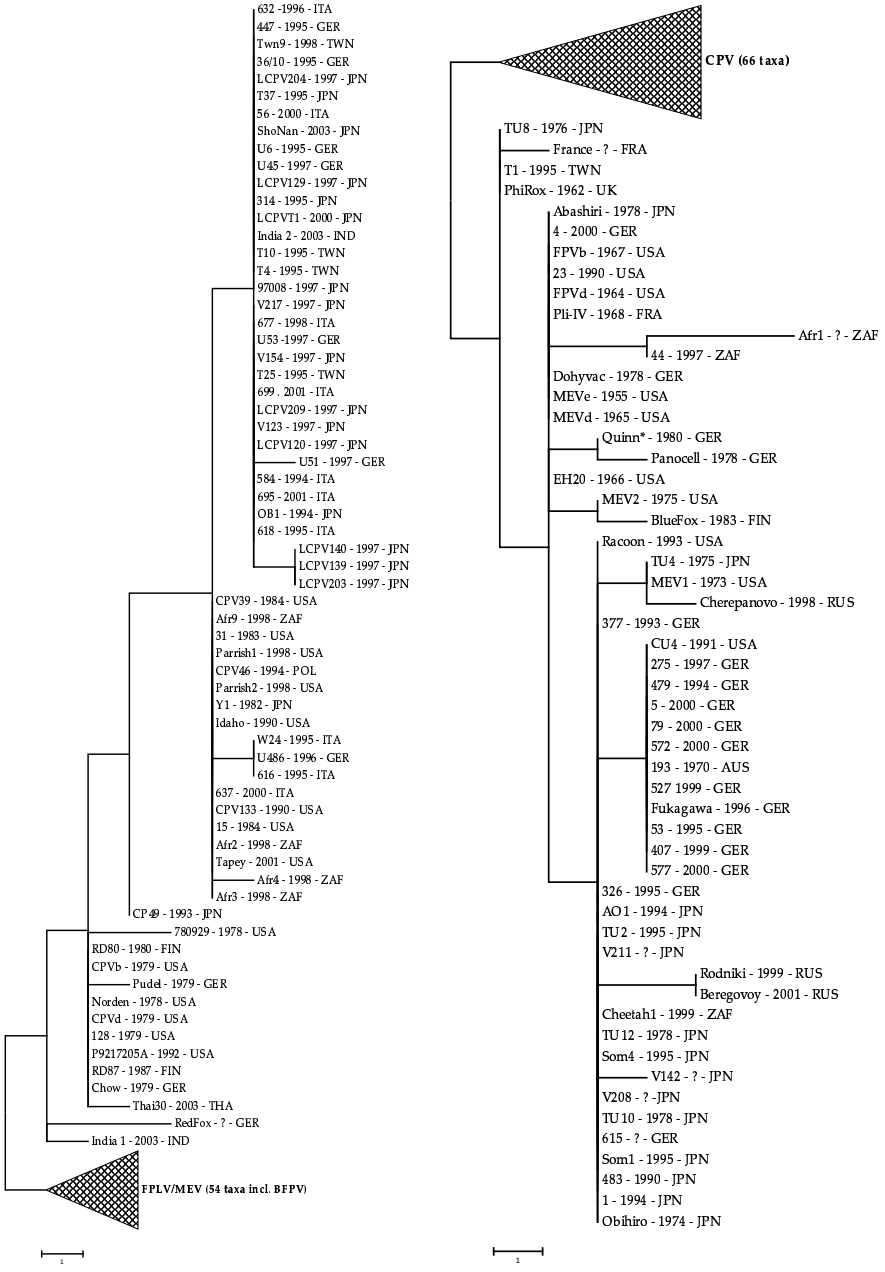
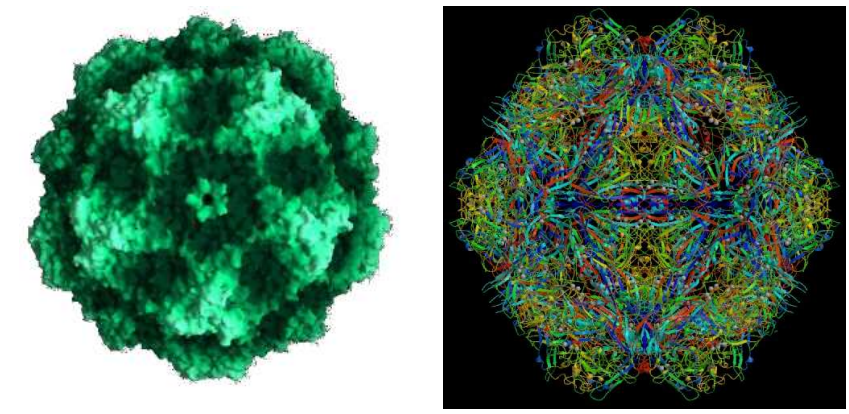


Figure A.4 Maximum Parsimony tree for the partial VP2 protein sequence. **Left** CPV **Right** FPLV

Appendix B

Structure of the Virion



(a) Surface Rendering

(b) Secondary and Tertiary Structure

Figure B.1 Surface structure of the viral coat protein (biological molecule). **Left** The isometric surface structure of the viral capsid (Based on X-Ray data by Tsao *et al.* [77]. Reconstruction and copyright by Jean-Yves Sgro) **Right** The viral capsid, i.e. the biological unit, consists of 60 monomers, i.e. asymmmetric units.

Appendix C

Parvovirus Species List

C.1 Legend

The question mark ("**?**") indicates missing or unknown data. The asterisk ("*") marks sequences excluded due to insufficient sequence length (less than 60% of the 1755nt full-length VP2 gene). The dash ("**–**") stands for "Not applicable". *Slanted labels* are corresponding to vaccine strains.

Sequences from **Africa** (7 used)

Name	**Type**	**Subtype**	**Origin**			**GenBank**
			Year	Country	Host	
Afr2	CPV	2b	1998	ZAF	dog	AJ007497
Afr3	CPV	2b	1998	ZAF	dog	AJ007498
Afr4	CPV	2b	1998	ZAF	dog	AJ007499
Afr9	CPV	2a	1998	ZAF	dog	AJ007500
44	FLPV	–	1997	ZAF	cat	AJ249556
Afr1	FPLV	–	?	ZAF	dog	–
Cheetah	FPLV	–	1999	ZAF	cheetah	AJ249557

Sequences from **America** (24 used)

Name	**Type**	**Subtype**	**Origin**			**GenBank**
			Year	Country	Host	
15	CPV	2a	1984	USA	dog	M24003
102/10*	CPV	?	1986	USA	dog	M11871
128	CPV	2a	1979	USA	dog	U22186
780929	CPV	2	1978	USA	dog	M10989

31	CPV	2a	1983	USA	dog	M24000
P9217205A	CPV	2a	1992	USA	?	A26575
CPV39	CPV	2b	1984	USA	dog	M74849
CPV133	CPV	2b	1990	USA	dog	M74852
CPV-b	CPV	2	1979	USA	dog	M38245
CPV-d	CPV	2	1979	USA	dog	M23255
CPV-d*	CPV	?	1991	USA	dog	M75727
CPV-d*	CPV	?	1986	USA	dog	M12998
Idaho	CPV	2b	1990	USA	lynx	U22896
Norden	CPV	2	1978	USA	dog	M19296
Parrish1	CPV	2b	1998	USA	dog	AR043630
Parrish2	CPV	2b	1998	USA	dog	AR043629
Tapey	CPV	2a	2001	USA	dog	AR130261
23	FPLV	–	1990	USA	wildcat	U22187
CU-4	FPLV	–	1991	USA	cat	M38246
EH20	FPLV	–	1966	USA	cat	M10824
FPV-b	FPLV	–	1967	USA	cat	M24004
FPV-d	FPLV	–	1964	USA	wildcat	U22189
FPLV*	FPLV	–	1991	USA	cat	M75728
MEV-1	MEV	–	1973	USA	mink	M23999
MEV-2	MEV	–	1975	USA	mink	M24001
MEV-d	MEV	–	1965	USA	mink	U22190
MEV-e	MEV	–	1955	USA	mink	U22191
Raccoon	RaPV	–	1993	USA	racoon	M24005

Sequences from **Asia** (45 used)

Name	**Type**	**Subtype**	**Origin**			**GenBank**
			Year	Country	Host	
Y34*	CPV	?	2000	CHI	dog	AF237781
China*	CPV	?	2000	CHI	dog	AF321277
India1	CPV	2a	2003	IND	dog	AJ564427
India2	CPV	2a	2004	IND	dog	AJ698134
314	CPV	2a	1995	JPN	cat	D78585
97008	CPV	2b	1997	JPN	dog	AB115504
CP49	CPV	2a	1993	JPN	?	D26081
LCPV120	CPV	2a	1997	JPN	cat	AB054215
LCPV129	CPV	2a	1997	JPN	cat	AB054216
LCPV139	CPV	2a	1997	JPN	leopard cat	AB054222
LCPV140	CPV	2ca	1997	JPN	leopard cat	AB054223
LCPV203	CPV	2cb	1997	JPN	cat	AB054224
LCPV204	CPV	2b	1997	JPN	leopard cat	AB054221

LCPV209	CPV	2b	1997	JPN	leopard cat	AB054219
LCPVT1	CPV	2a	2000	JPN	?	AB054214
OB1	CPV	2a	1994	JPN	?	D26080
ShoNan	CPV	2b	2003	JPN	?	AB128923
T37	CPV	2a	1995	JPN	dog	U72698
Twn9	CPV	2a	1998	JPN	cat	AB054213
V123	CPV	2b	1997	JPN	cat	AB054218
V154	CPV	2a	1997	JPN	cat	AB054217
V217	CPV	2b	1997	JPN	cat	AB054220
Y1	CPV	2a	1982	JPN	dog	D26079
Thai30	CPV	2a	2003	THA	?	AY262281
T4	CPV	2a	1995	TWN	dog	U72695
T25	CPV	2a	1995	TWN	?	U72697
T10	CPV	2b	1995	TWN	?	U72696
1	FPLV	-	1994	JPN	cat	AB000050
483	FPLV	-	1990	JPN	cat	D88286
AO1	FPLV	-	1994	JPN	cat	AB000052
Fukagawa	FPLV	-	1996	JPN	cat	AB000054
Obihiro	FPLV	-	1974	JPN	cat	AB000056
Som1	FPLV	-	1994	JPN	cat	AB000059
Som4	FPLV	-	1995	JPN	cat	AB000061
TU2	FPLV	-	1975	JPN	cat	AB000066
TU4	FPLV	-	1975	JPN	cat	AB000068
TU8	FPLV	-	1976	JPN	cat	AB000070
TU10	FPLV	-	1978	JPN	cat	D78584
TU12	FPLV	-	1978	JPN	cat	AB000064
V142	FPLV	-	?	JPN	cat	AB054225
V208	FPLV	-	?	JPN	cat	AB054226
V211	FPLV	-	?	JPN	cat	AB054227
Abashiri	MEV	-	1978	JPN	?	D00765
Beregovoy	MEV	-	2001	SU	mink	AF469009
Cherepanovo	MEV	-	1998	SU	mink	AF201477
Rodniki	MEV	-	1999	SU	mink	AF201478
T1	FPLV	-	1995	TWN	leopard cat	AF015223

Sequences from **Australia** (1 used)

Name	**Type**	**Subtype**	**Origin**			**GenBank**
			Year	Country	Host	
193	FPLV	–	1970	AUS	cat	X55115

Sequences from **Europe** (43 used)

Name	**Type**	**Subtype**	**Origin**			**GenBank**
			Year	Country	Host	
RD87	CPV	2a	1987	FIN	raccoon dog	U22193
RD80	CPV	2a	1980	FIN	raccoon dog	U22192
36/10	CPV	2a	1995	GER	cat	–
447	CPV	2b	1995	GER	cat	–
Chow	CPV	2	1979	GER	dog	AJ002927
Pudel	CPV	2	1979	GER	dog	AJ002928
RedFox	CPV	?	?	GER	red fox	–
U45	CPV	2a	1997	GER	dog	–
U486	CPV	2b	1996	GER	dog	–
U51	CPV	?	1997	GER	dog	–
U53	CPV	2a	1997	GER	dog	–
U6	CPV	2a	1995	GER	dog	–
56	CPV	?	2000	ITA	dog	AY380577
699	CPV	2a	2001	ITA	?	AF393506
584	CPV	2a	1994	ITA	dog	AF306446
616	CPV	2b	1995	ITA	dog	AF306449
618	CPV	2a	1995	ITA	dog	AF306447
632	CPV	2a	1996	ITA	dog	AF306445
637	CPV	2b	2000	ITA	dog	AF306450
677	CPV	2a	1998	ITA	dog	AF306448
695	CPV	?	2001	ITA	cat	AF401519
W24	CPV	2b	1995	ITA	wolf	AF306444
CPV46	CPV	2b	1994	POL	dog	Z46651
BlueFox	FPLV	–	1983	FIN	blue fox	U22185
PLI-IV	FPLV	–	1968	FRA	cat	D88287
France	FPLV	–	?	FRA	cat	–
275	FPLV	–	1997	GER	cat	–
326	FPLV	–	1995	GER	cat	–
377	FPLV	–	1993	GER	wildcat	U22188
4	FPLV	–	2000	GER	cat	–
407	FPLV	–	1999	GER	cat	–
479	FPLV	–	1994	GER	cat	–
5	FPLV	–	2000	GER	cat	–
527	FPLV	–	1999	GER	cat	–
53	FPLV	–	1995	GER	cat	–
572	FPLV	–	2000	GER	cat	–
577	FPLV	–	2000	GER	cat	–
615	FPLV	–	?	GER	cat	AJ002930
79	FPLV	–	2000	GER	cat	–

Dohyvac	FPLV	–	1978	GER	cat	AJ002931
Panocell	FPLV	–	1978	GER	cat	AJ002932
Quinn	FPLV	–	1980	GER	dog	AJ002929
PhiRox	FPLV	–	1962	UK	cat	M24002

Outgroup Sequences (2 used)

Name	**Type**	**Subtype**	**Origin**			**GenBank**
			Year	Country	Host	
Rat1a	RPV	–	1998	USA	rat	AF036710
NADL2	PPV	–	1993	CAN	pig	L23427

Appendix D

Canine Parvorvirus Genome Sequence

The complete genome sequence of CPV strain "Norden" is given in the following part. Non-coding or unassigned regions are coloured in black, the specific sequence of the non-structural gene (NS) in blue, and the viral capsid gene (VP) in green. Yellow parts indicate splicing variant A of VP1, and red parts splicing variant B of VP1, with orange parts of overlapping regions of spliced introns. The viral capsid gene VP2 is shown in dark green, and terminal repeats of the VP2 gene are coloured in cyan.

```
>gi|333438|gb|M19296.1|PVCCPN Canine parvovirus strain CPV-N,
complete cds
ATTCTTTAGAACCAACTGACCAAGTTCACGTACGTATGACGTGATGACCCGCTGCGCGCG   60
CTGCCTACGGCAGTCACACGTCATACGTACGCTCCTTGGTCAGTTGGTTCTAAAGAATGA  120
TAGGCGGTTTGTGTGTTTAAACTTGGGCGGGAAAAGGTGGCGGGCTAATTGTGGGCGTGG  180
TTAAAGGTATAAAAGACAAACCATAGACCGTTACTGACATTCGCTTCTTGTCTTTGACAG  240
AGTGAACCTCTCTTACTCTGACTAACCAACCATGTCTGGCAACCAGTATACTGAGGAAGT  300
TATGGAGGGAGTAAATTGGTTAAAGAAACATGCAGAAAATGAAGCATTTTCGTTTGTTTT  360
TAAATGTGACAACGTCCAACTAAATGGAAAGGATGTTCGCTGGAACAACTATACCAAACC  420
AATTCAAAATGAAGAACTAACATCTTTAATTAGAGGAGCACAAACAGCAATGGATCAAAC  480
CGAAGAAGAAGAAATGGACTGGGAATCGGAAGTTGATAGTCTCGCCAAAAAGCAAGTACA  540
AACTTTTGATGCATTAATTAAAAAATGTCTTTTTGAAGTCTTTGTTTCTAAAAATATAGA  600
ACCAAATGAATGTGTTTGGTTTATTCAACATGAATGGGGAAAAGATCAAGGCTGGCATTG  660
TCATGTTTTACTTCATAGTAAGAACTTACAACAAGCAACTGGTAAATGGCTACGCAGACA  720
AATGAATATGTATTGGAGTAGATGGTTGGTGACTCTTTGTTCGGTAAACTTAACACCAAC  780
TGAAAAGATTAAGCTCAGAGAAATTGCAGAAGATAGTGAATGGGTGACTATATTAACATA  840
CAGACATAAGCAAACAAAAAAAGACTATGTTAAAATGGTTCATTTTGGAAATATGATAGC  900
ATATTACTTTTTAACAAAGAAAAAAATTGTCCACATGACAAAAGAAAGTGGCTATTTTTT  960
AAGTACTGATTCTGGTTGGAAATTTAACTTTATGAAGTATCAAGACAGACAAATTGTCAG 1020
```

```
CACACTTTACACTGAACAAATGAAACCAGAAACCGTTGAAACCACAGTGACGACAGCACA 1080
GGAAACAAAGCGCGGGAGAATTCAAACTAAAAAGGAAGTGTCAATCAAATGTACTTTGCG 1140
GGACTTGGTTAGTAAAAGAGTAACATCACCTGAAGACTGGATGATGTTACAACCAGATAG 1200
TTATATTGAAATGATGGCACAACCAGGAGGTGAAAATCTTTTAAAAAATACACTTGAAAT 1260
TTGTACTTTGACTTTAGCAAGAACAAAAACAGCATTTGAATTAATACTTGAAAAAGCAGA 1320
TAATACTAAACTAACTAACTTTGATCTTGCAAATTCTAGAACATGTCAAATTTTTAGAAT 1380
GCACGGATGGAATTGGATTAAAGTTTGTCACGCTATAGCATGTGTTTTAAATAGACAAGG 1440
TGGTAAAAGAAATACAGTTCTTTTTCATGGACCAGCAAGTACAGGAAAATCTATCATTGC 1500
TCAAGCCATAGCACAAGCTGTGGGTAATGTTGGTTGTTATAATGCAGCAAATGTAAATTT 1560
TCCATTTAATGACTGTACCAATAAAAATTTAATTTGGATTGAAGAAGCTGGTAACTTTGG 1620
TCAACAAGTTAATCAATTTAAAGCAATTTGTTCTGGACAAACAATTAGAATTGATCAAAA 1680
AGGTAAAGGAAGTAAGCAAATTGAACCAACTCCAGTAATTATGACAACTAATGAAAATAT 1740
AACAATTGTGAGAATTGGATGTGAAGAAAGACCTGAACATACACAACCAATAAGAGACAG 1800
AATGTTGAACATTAAGTTAGTATGTAAGCTTCCAGGAGACTTTGGTTTGGTTGATAAAGA 1860
AGAATGGCCTTTAATATGTGCATGGTTAGTTAAACATGGTTATGAATCAACCATGGCTAA 1920
CTATACACATCATTGGGGAAAAGTACCAGAATGGGATGAAAACTGGGCGGAGCCTAAAAT 1980
ACAAGAAGGTATAAATTCACCAGGTTGCAAAGACTTAGAGACACAAGCGGCAAGCAATCC 2040
TCAGAGTCAAGACCAAGTTCTAACTCCTCTGACTCCGGACGTAGTGGACCTTGCACTGGA 2100
ACCGTGGAGTACTCCAGATACGCCTATTGCAGAAACTGCAAATCAACAATCAAACCAACT 2160
TGGCGTTACTCACAAAGACGTGCAAGCGAGTCCGACGTGGTCCGAAATAGAGGCAGACCT 2220
GAGAGCCATCTTTACTTCTGAACAATTGGAAGAAGATTTTCGAGACGACTTGGATTAAGG 2280
TACGATGGCACCTCCGGCAAAGAGAGCCAGGAGAGGTAAGGGTGTGTTAGTAAAGTGGGG 2340
GGAGGGGAAAGATTTAATAACTTAACTAAGTATGTGTTTTTTTATAGGACTTGTGCCTCC 2400
AGGTTATAAATATCTTGGGCCTGGGAACAGTCTTGACCAAGGAGAACCAACTAACCCTTC 2460
TGACGCCGCTGCAAAAGAACACGACGAAGCTTACGCTGCTTATCTTCGCTCTGGTAAAAA 2520
CCCATACTTATATTTCTCGCCAGCAGATCAACGCTTTATAGATCAAACTAAGGACGCTAA 2580
AGATTGGGGGGGGAAAATAGGACATTATTTTTTTAGAGCTAAAAAGGCAATTGCTCCAGT 2640
ATTAACTGATACACCAGATCATCCATCAACATCAAGACCAACAAAACCAACTAAAAGAAG 2700
TAAACCACCACCTCATATTTTCATCAATCTTGCAAAAAAAAAAAAAAGCCGGTGCAGGACA 2760
AGTAAAAAGAGACAATCTTGCACCAATGAGTGATGGAGCAGTTCAACCAGACGGTGGTCA 2820
ACCTGCTGTCAGAAATGAAAGAGCTACAGGATCTGGGAACGGGTCTGGAGCGGGGGTGGT 2880
GGTGGTTCTGGGGGTGTGGGGATTTCTACGGGTACTTTCAATAATCAGACGGAATTTAAA 2940
TTTTTGGAAAACGGATGGGTGGAAATCACAGCAAACTCAAGCAGACTTGTACATTTAAAT 3000
ATGCCAGAAAGTGAAAATTATAGAAGAGTGGTTGTAAATAATATGGATAAAACTGCAGTT 3060
AACGGAAACATGGCTTTAGATGATATTCATGCACAAATTGTAACACCTTGGTCATTGGTT 3120
GATGCAAATGCTTGGGGAGTTTGGTTTAATCCAGGAGATTGGCAACTAATTGTTAATACT 3180
ATGAGTGAGTTGCATTTAGTTAGTTTTGAACAAGAAATTTTTAATGTTGTTTTAAAGACT 3240
GTTTCAGAATCTGCTACTCAGCCACCAACTAAAGTTTATAATAATGATTTAACTGCATCA 3300
TTGATGGTTGCATTAGATAGTAATAATACTATGCCATTTACTCCAGCAGCTATGAGATCT 3360
GAGACATTGGGTTTTTATCCATGGAAACCAACCATACCAACTCCATGGAGATATTATTTT 3420
CAATGGGATAGAACATTAATACCATCTCATACTGGAACTAGTGGCACACCAACAAATATA 3480
TACCATGGTACAGATCCAGATGATGTTCAATTTTATACTATTGAAAATTCTGTGCCAGTA 3540
CACTTACTAAGAACAGGTGATGAATTTGCTACAGGAACATTTTTTTTTGATTGTAAACCA 3600
```

```
TGTAGACTAACACATACATGGCAAACAAATAGAGCATTGGGCTTACCACCATTTCTAAAT 3660
TCTTTGCCTCAATCTGAAGGAGCTACTAACTTTGGTGATATAGGAGTTCAACAAGATAAA 3720
AGACGTGGTGTAACTCAAATGGGAAATACAAACTATATTACTGAAGCTACTATTATGAGA 3780
CCAGCTGAGGTTGGTTATAGTGCACCATATTATTCTTTTGAGGCGTCTACACAAGGGCCA 3840
TTTAAAACACCTATTGCAGCAGGACGGGGGGGAGCGCAAACATATGAAAATCAAGCAGCA 3900
GATGGTGATCCAAGATATGCATTTGGTAGACAACATGGTCAAAAAACTACCACAACAGGA 3960
GAAACACCTGAGAGATTTACATATATAGCACATCAAGATACAGGAAGATATCCAGAAGGA 4020
GATTGGATTCAAAATATTAACTTTAACCTTCCTGTAACGAATGATAATGTATTGCTACCA 4080
ACAGATCCAATTGGAGGTAAAACAGGAATTAACTATACTAATATATTTAATACTTATGGT 4140
CCTTTAACTGCATTAAATAATGTACCACCAGTTTATCCAAATGGTCAAATTTGGGATAAA 4200
GAATTTGATACTGACTTAAAACCAAGACTTCATGTAAATGCACCATTTGTTTGTCAAAAT 4260
AATTGTCCTGGTCAATTATTTGTAAAAGTTGCGCCTAATTTAACAAATGAATATGATCCT 4320
GATGCATCTGCTAATATGTCAAGAATTGTAACTTACTCAGATTTTTGGTGGAAAGGTAAA 4380
TTAGTATTTAAAGCTAAACTAAGAGCCTCTCATACTTGGAATCCAATTCAACAAATGAGT 4440
ATTAATGTAGATAACCAATTTAACTATGTACCAAGTAATATTGGAGGTATGAAAATTGTA 4500
TATGAAAAATCTCAACTAGCACCTAGAAAATTATATTAACATACTTACTATGGTTTTTAT 4560
GTTTATTACATATCAACTAGCACCTAGAAAAATTATATTAATATACTTACTATGGTTTTT 4620
ATGTTTATTACATATTATTTTAAGATTAATTAAATACAGCATAGAAATATTGTACTTGTA 4680
TTTGATATAGGATTTAGAAGTTTGTTAGATGGTATACAATAACTGTAAGAAATAGAAGAA 4740
CATTTAGATCATAGTTAGTAGGTTTGTTAGATGGTATACAATAACTGTAAGAAATAGAAG 4800
AACATTTAGATCATAGTTAGTAGGTTTGTTATATGGTATACAATAACTGTAAGAAATAGA 4860
AGAACATTTAGATCATAGTTAGTAGTTTGTTTTATAAAATGTATTGTAAACCATTAATGT 4920
ATGTTGTTATGGTGTGGGTGGTTGGTTGGTTTGCCCTTAGAATATGTTAAGGACCAAAAA 4980
AATCAATAAAAGACATTTAAAACTAAATGGCCTCGTATACTGTCTATAAGGTGAACTAAC 5040
CTTACCATAAGTATCAATCGTTGCGCCCTAATTTAACAAATGAATATGATCCTGATGCAT 5100
CTGCTAATATGTCAAGAATTGTAACTTACTCAGATTTTTGGTGGAAAGGTAAATTAGTAT 5160
TTAAAGCTAAACTAAGAGCCTCTCATACTTGGAATCCAATTCAACAAATGAGTATTAATG 5220
TAGATAACCAATTTAACTATGTACCAAGTAATATTGGAGGTATGAAAATTGTATATGAAA 5280
AATCTCAACTAGCACCTAGAAAATTATATTAACATCTCTAGA                   5323
```

D.1 NS Translated Sequence

```
MSGNQYTEEVMEGVNWLKKHAENEAFSFVFKCDNVQLNGKDVRWNNYTKPIQNEELTSLI   60
RGAQTAMDQTEEEEMDWESEVDSLAKKQVQTFDALIKKCLFEVFVSKNIEPNECVWFIQH  120
EWGKDQGWHCHVLLHSKNLQQATGKWLRRQMNMYWSRWLVTLCSVNLTPTEKIKLREIAE  180
DSEWVTILTYRHKQTKKDYVKMVHFGNMIAYYFLTKKKIVHMTKESGYFLSTDSGWKFNF  240
MKYQDRQIVSTLYTEQMKPETVETTVTTAQETKRGRIQTKKEVSIKCTLRDLVSKRVTSP  300
EDWMMLQPDSYIEMMAQPGGENLLKNTLEICTLTLARTKTAFELILEKADNTKLTNFDLA  360
NSRTCQIFRMHGWNWIKVCHAIACVLNRQGGKRNTVLFHGPASTGKSIIAQAIAQAVGNV  420
GCYNAANVNFPFNDCTNKNLIWIEEAGNFGQQVNQFKAICSGQTIRIDQKGKGSKQIEPT  480
PVIMTTNENITIVRIGCEERPEHTQPIRDRMLNIKLVCKLPGDFGLVDKEEWPLICAWLV  540
KHGYESTMANYTHHWGKVPEWDENWAEPKIQEGINSPGCKDLETQAASNPQSQDQVLTPL  600
```

```
TPDVVDLALEPWSTPDTPIAETANQQSNQLGVTHKDVQASPTWSEIEADLRAIFTSEQLE  660
EDFRDDLD                                                       668
```

D.2 VP Translated Sequence

```
MAPPAKRARRGKGVLVKWGEGKDLITXLSMCFFIGLVPPGYKYLGPGNSLDQGEPTNPSD   60
AAAKEHDEAYAAYLRSGKNPYLYFSPADQRFIDQTKDAKDWGGKIGHYFFRAKKAIAPVL  120
TDTPDHPSTSRPTKPTKRSKPPPHIFINLAKKKKAGAGQVKRDNLAPMSDGAVQPDGGQP  180
AVRNERATGSGNGSGGGGGGGSGGVGISTGTFNNQTEFKFLENGWVEITANSSRLVHLNM  240
PESENYRRVVVNNMDKTAVNGNMALDDIHAQIVTPWSLVDANAWGVWFNPGDWQLIVNTM  300
SELHLVSFEQEIFNVVLKTVSESATQPPTKVYNNDLTASLMVALDSNNTMPFTPAAMRSE  360
TLGFYPWKPTIPTPWRYYFQWDRTLIPSHTGTSGTPTNIYHGTDPDDVQFYTIENSVPVH  420
LLRTGDEFATGTFFFDCKPCRLTHTWQTNRALGLPPFLNSLPQSEGATNFGDIGVQQDKR  480
RGVTQMGNTNYITEATIMRPAEVGYSAPYYSFEASTQGPFKTPIAAGRGGAQTYENQAAD  540
GDPRYAFGRQHGQKTTTTGETPERFTYIAHQDTGRYPEGDWIQNINFNLPVTNDNVLLPT  600
DPIGGKTGINYTNIFNTYGPLTALNNVPPVYPNGQIWDKEFDTDLKPRLHVNAPFVCQNN  660
CPGQLFVKVAPNLTNEYDPDASANMSRIVTYSDFWWKGKLVFKAKLRASHTWNPIQQMSI  720
NVDNQFNYVPSNIGGMKIVYEKSQLAPRKLY                               751
```

D.3 Genetic Code Table

Amino Acid			**Amino Acid**		
1-letter	3-letter	Full Name	1-letter	3-letter	Full Name
A	Ala	Alanine	M	Met	Methionine
C	Cys	Cysteine	N	Asn	Asparagine
D	Asp	Aspartate	P	Pro	Proline
E	Glu	Glutamate	Q	Gln	Glutamine
F	Phe	Phenylalanine	R	Arg	Arginine
G	Gly	Glycine	S	Ser	Serine
H	His	Histidine	T	Thr	Threonine
I	Ile	Isoleucine	V	Val	Valine
K	Lys	Lysine	W	Trp	Tryptophane
L	Leu	Leucine	Y	Tyr	Tyrosine

Appendix E

`treepather` manpage

```
treepather(1)    Customized Bioinf Commands    treepather(1)
```

NAME

TreePath-er - A NEWICK tree **pa**rs**er**
able to determine pa**th** lengths

SYNTAX

```
treepather newick-file afa-file
```

DESCRIPTION

TreePather is a demonstration program for the NEWICK-tree parser and access library. It reads a NEWICK tree, computes its leaves and then calculates various distance measures between leaves. (To compute the Hamming-distance between leaves, a AFA-file is also read).

THEORY OF OPERATION

TreePather is intended as a demonstration on the more general NEWICK tree library, which provides language bindings for C++ and Python. The library provides functions to read a NEWICK tree file into memory and unmarshall it into an object hierarchy, perform path operations on it (including the calculation of the longest common subpath) and can output a tree in NEWICK tree & VCG (Visualization of Compiler Graphs) formats.

The NEWICK tree library is based on a Bison grammar

implementing Gary Olsen's interpretation of the NEWICK tree format. This formal specification is available from

http://evolution.genetics.washington.edu/phylip/newick_doc.html

SEE ALSO

xvcg(1), fa2afa

Appendix F

`vdiff` manpage

```
vdiff(1)          Customized Bioinf Commands          vdiff(1)
```

NAME

vdiff - A Vertical diff-Derivative for Sequence Alignments

SYNTAX

```
vdiff.py [-g|-G] [-o] filename
vdiff.py [-o] [-p pos] filename
vdiff.py [-q name[:threshold]] [-Q] filename
```

DESCRIPTION

vdiff reads aligned sequence input (DNA or protein) and returns variable positions. Specifying command line options allows advanced comparative sequence analysis. Default output is printed to stdout and includes

(1) Total number & raw indices of variable positions
(2) Total number & distribution of different characters (i.e. list of taxa sharing identical positions) for all variable positions*
(3) Summary of specific positions for all taxa*

* using raw indices

OPTIONS

-g value
Additionally prints groups (sorted according

to frequency of occurrence)

-G
Alters dafault output: prints groups only (exclusive use)

-o
Allow positional offset (exclusive use) to convert raw indices into true indices

-p position
Prints results for specified position(s) only

-q name[:value]
Prints results for specified taxon only (using a threshold for the number of taxa sharing a specific position). Default threshold is 1 if no value is specified by the user.

-Q
Uses positions from -q as input for a -p positional query

INPUT FILE FORMAT

Input files must be standard FASTA (.fa) derivatives (almost fasta, AFA).

The .afa file format requires ASCII text in a two column table with columns separated by a space or tabulator and rows separated by a newline character. Cell content must not contain spaces or tabulators.

The first column holds the sequence label names and the second column holds the aligned sequences. Sequences may contain gaps represented as dashes ("-") or dots (".). Sequences must be of equal length.

FASTA and NEXUS (.nex) files may be converted to AFA format using fa2afa and nex2afa, respectively.

SEE ALSO

nex2afa, fa2afa

Bibliography

[1] www.ncbi.nlm.nih.gov/ICTVdb/.

[2] www.ncbi.nlm.nih.gov/Genbank/.

[3] http://www.rcsb.org/pdb/.

[4] SP Altschul, TL Madsen, AA Schäffer, J Zhang, Z Zhang, W Miller, and DF Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[5] LJ Ball-Goodrich, SE Leland, EA Johnson, FX Patyrzo, and RO Jacoby. Rat Parvovirus type 1: The prototype for a new Rodent Parvovirus subgroup. *Journal of Virology*, 74(4):3289–3299, Apr 1998.

[6] IK Barker, RC Povey, and DR Voigt. Response of Mink, Skunk, Red Fox and Racoon to inoculation with Mink Enteritis Virus, Feline Panleukopenia and Canine Parvovirus and prevalence of antibody to Parvovirus in wild Carnivores in Ontario. *Canadian Journal of Comparative Medicine and Veterinary Science*, 47(2):188–197, Apr 1983.

[7] M Battilani, S Ciulli, E Tisato, and S Prosperi. Genetic analysis of Canine Parvovirus isolates (CPV-2) from dogs in Italy. *Virus Research*, 83:149–157, 2002.

[8] M Battilani, A Scagliarini, E Tisato, C Turilli, I Jacobini, R Casadio, and S Prosperi. Analysis of Canine Parvovirus sequences from

wolves and dogs isolated in Italy. *Journal of General Virology*, 82:1555–1560, 2001.

[9] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Schindyalov, and PE Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[10] D Bryant and V Moulton. Neighbor-Net: An agglomerate method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2):255–265, 2004.

[11] RM Bush, WM Fitch, CA Bender, and NJ Cox. Positive selection on the H3 Hemaglutinin gene of Human Influenza virus A. *Molecular Biology and Evolution*, 16(11):1457–1465, 1999.

[12] MS Chapman and MG Rossman. Structural refinement of the DNA-containing capsid of Canine Parvovirus using *rsref*, a resolution-dependent sterochemically restrained real-space refinement method. *Acta Crystallographica Biologica*, D52(1):129–142, Jan 1996.

[13] MS Chapman and MG Rossmann. Sequence and function correlations among parvoviruses. *Virology*, 194(2):491–508, Jun 1993.

[14] MO Dayhoff, RM Schwartz, and BC Orcutt. *Atlas of Protein Sequence and Structure*, volume 2 of *5*, chapter A new model of evolutionary change in proteins, pages 345–352. National Biomedical Research Foundation, Washington, DC, 1978.

[15] A Dress, D Huson, and V Moulton. Analyzing and visualizing sequence and distance data using SplitsTree. *Discrete Applied Mathematics*, 71:95–109, 2004.

[16] WM Fitch, RM Bush, CA Bender, K Subbarao, and NJ Cox. Predicting the evolution of Human Influenza A. *Science*, 286(5446):1921–1925, Dec 1999.

[17] WM Fitch, JME Leiter, X Li, and P Palese. Positive darwinian evolution in Human Influenza-A viruses. *Proceedings of the National Academy of Sciences*, 88(10):4270–4274, May 1991.

[18] M Georgieva. Emergence of "New" viral zoonoses: Filoviral Hemorrhagic Fever. *Experimental Pathology and Parasitology*, 4, 2000.

[19] L Govindasamy, K Hueffer, CR Parrish, and M Agbandje-McKenna. Structures of host range-controlling regions of the capsids of canine and feline parvoviruses and mutants. *Journal of Virology*, 77(22):12211–12221, Nov 2003.

[20] NC Grassly, PH Harvey, and EC Holmes. Population dynamics of HIV-1 inferred from gene sequences. *Genetics*, 151:427–438, 1999.

[21] I Greiser-Wilke and L Haas. Emergence of "New" viral zoonoses. *Deutsche Tierärztliche Wochenschrift*, 106:332–338, 1999.

[22] M Hasegawa, H Kishino, and T Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[23] AJ Hay, V Gregory, AR Douglas, and YP Lin. The evolution of Human Influenza viruses. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 356(1416):1861–1870, Dec 2001.

[24] J Hetzl and PR Tomsich. `treePath-er`: A NEWICK standard tree parser able to determine path lengths, 2004.

[25] J Hetzl and PR Tomsich. `vdiff`: A vertical `diff` derivative for sequence comparison, 2004.

[26] M Horiuchi, Y Yamaguchi, T Gojobori, M Mochizuki, H Nagasawa, Y Toyoda, N Ishiguro, and M Shinagawa. Differences in the evolutionary pattern of Feline Panleukopenia Virus and Canine Parvovirus. *Virology*, 249(2):440–452, Sep 1998.

[27] K Hueffer, L Govidasamy, M Agbandje-McKenna, and CR Parrish. Combinations of two capsid regions controlling canine host range determine canine transferrin receptor binding by canine and feline parvovirus. *Journal of Virology*, 77(18):10099–10105, Sep 2003.

[28] K Hueffer, JSL Parker, WS Weichert, RE Geisel, JY Sgro, and CR Parrish. The natural host range shift and subsequent evolution of Canine Parvovirus resulted from virus-specific binding to the canine transferrin receptor. *Journal of Virology*, 77(16):8915–8923, Aug 2003.

[29] K Hueffer and CR Parrish. Parvovirus host range: Cell tropism and evolution. *Current Opinions in Microbiology*, 6(4):392–398, Aug 2003.

[30] W Humphrey, A Dalke, and K Schulten. VMD—Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[31] DH Huson. SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

[32] Y Ikeda, M Mochizuki, R Naito, K Nakamura, T Miyazawa, T Mikami, and E Takahashi. Predominance of Canine Parvovirus (CPV) in unvaccinated cat populations and emergence of new antigenic types of CPVs in cats. *Virology*, 278(1):13–19, Dec 2000.

[33] Y Ikeda, K Nakamura, T Miyazawa, Y Toyha, E Takahashi, and M Mochizuki. Feline host range of Canine Parvovirus: Recent emergence of new antigenic types in cats. *Emerging Infectious Diseases*, 8(4):341–316, Apr 2002.

[34] M Kimura. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[35] B Korber. SNAP: Synonymous Non-synonymous Analysis Program. `http://www.hiv.lanl.gov/content/hiv-db/SNAP/README.html`, 1998.

[36] B Korber. *HIV signature and sequence variation analysis*, chapter 4, pages 55–72. Kluwer Academic, Dordrecht, Netherlands, 2000.

[37] S Kumar, K Tamura, and M Nei. MEGA: Molecular Evolutionary Genetic Analysis software for microcomputers. *Computer Applications in the Biosciences*, 10(2):189–192, 1994.

[38] S Kumar, K Tamura, and M Nei. MEGA3: Integrated software for Molecular Evolutionary Genetic Analysis and sequence alignment. *Briefings in Bioinformatics*, 2004.

[39] C Lanave, G Preparata, C Saccone, and G Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, (20):86–93, 1984.

[40] M Nakamura, Y Toyha, T Miyazawa, M Mochizuki, HTT Phung, NH Nguyen, LMT Hunyh, LT Nguyen, PN Nguyen, PV Nguyen, NPT Nguyen, and H Akashi. A novel antigenic variant of Canine Parvovirus from a Vietnamese dog. *Archives of Virology*, Jul 2004.

[41] N Nathanson, KA McGann, J Wilesmith, RC Desrosiers, and R Brookmeyer. The evolution of virus diseases: Their emergence, epidemicity and control. *Virus Research*, 29(1):3–20, Jun 1993.

[42] M Nei and T Gojobori. Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3:418–426, 1986.

[43] T Ota and M Nei. Variance and covariances of the numbers of synonymous and non-synonymous substitutions per site. *Molecular Biology and Evolution*, 11(4):613–619, 1994.

[44] LM Palermo, K Hueffer, and CR Parrish. Residues in the apical domain of the feline and canine transferrin receptors control host-specific binding and cell infection of Canine and Feline Parvoviruses. *Journal of Virology*, 77(16):8915–8923, Aug 2003.

[45] JSL Parker, WJ Murphy, D Wand, SJ O'Brian, and CR Parrish. Canine and feline parvoviruses can use human or feline transferrin receptors to bind, enter and infect cells. *Journal of Virology*, 75(8):3896–3902, Apr 2001.

[46] JSL Parker and CR Parrish. Canine Parvovirus host range is determined by the specific conformation of an additional region of the capsid. *Journal of Virology*, 71(2):9214–9222, Dec 1997.

[47] CR Parrish. Mapping specific functions in the capsid structure of canine parvovirus and feline panleukopenia virus using infectious plasmid clones. *Virology*, 183(1):195–205, Jul 1991.

[48] CR Parrish. Host range relationships and the evolution of Canine Parvovirus. *Veterinary Microbiology*, 69(1–2):29–40, Sep 1999.

[49] CR Parrish, CF Aquadro, ML Strassheim, JF Evermann, JY Sgro, and HO Mohammed. Rapid antigenic-type replacement and DNA sequence evolution of Canine Parvovirus. *Journal of Virology*, 65(12):6544–6552, Dec 1991.

[50] CR Parrish and LE Carmichael. Characterization and recombination mapping of an antigenic and host range mutation of canine parvovirus. *Virology*, 148(1):121–132, Jan 1986.

[51] CAD Pereira and EL Durigon. Genetic diversity of the VP1/VP2 gene of Canine Parvovirus type 2b amplified from clinical specimens in Brazil. *Brazilian Journal of Microbiology*, 31(4):312–314, Oct 2000.

[52] D Posada and KA Crandall. Modeltest: Testing the model of DNA substitutions. *Bioinformatics*, 14(9):817–818, 1998.

[53] S Prohaska. Personal communication, 2004.

[54] A Rambaut and NC Grassly. Seq-Gen: An application for the Monte Carlo simulation of dna sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13:235–238, 1997.

[55] AP Reed, EV Jones, and TJ Miller. Nucleotide sequence and genome organization of canine parvovirus. *Journal of Virology*, 62(1):266–76, Jan 1998.

[56] F Rodriguez, JF Oliver, A Marin, and JR Medina. The general stochastic model of nucleotide substitutions. *Journal of Theoretical Biochemistry*, (142):485–501, 1990.

[57] N Saitou and M Nei. The Neighbour-Joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, Jul 1987.

[58] H Sakaoka, K Kurita, Y Iida, S Takada, K Umene, YT Kim, CS Ren, and AJ Nahmias. Quantitative analysis of genomic polymorphism of Herpes Simplex Virus type 1 strains from six countries: Studies of molecular evolution and molecular epidemiology of the virus. *Journal of General Virology*, 75(3):513–527, Mar 1994.

[59] LA Shackelton, CR Parrish, U Truyen, and EC Holmes. High rate of viral evolution associated with emergence of Carnivore Parvovirus. *Proceedings of the National Academy of Sciences*, 2004. in press.

[60] AA Simpson, V Chandrasekar, B Hebert, GM Sullivan, MG Rossman, and CR Parrish. Host range and variability of Calcium binding by surface loops in the capsids of Canine and Feline Parvoviruses. *Journal of Molecular Biology*, 300(3):597–610, Jul 200.

[61] V Sobolev, A Sorokine, J Prilusky, EE Abola, and M Edelman. Automated analysis of inter-atomic contacts in proteins. *Bioinformatics*, 15:327–332, 1999.

[62] AL Spitzer, CR Parrish, and IH Maxwell. Tropic determinant for Canine Parvovirus and Feline Panleukopenia Virus functions through the capsid protein VP2. *Journal of General Virology*, 78(4):925–928, Apr 1997.

[63] SPSS Inc., Chicago IL. *SPSS Base 10.0 for Windows User's Guide*, 1999.

[64] A Steinel, L Munson, M van Vuuren, and U Truyen. Genetic characterization of Feline Parvovirus sequences from various Carnivores. *Journal of General Virology*, 81(2):345–350, Feb 2000.

[65] A Steinel, CR Parrish, ME Boolm, and U Truyen. Parvovirus infections in wild Carnivores. *Journal of Wildlife Diseases*, 37(3):594–607, Jul 2001.

[66] RR Stocsits. *Nucleic acid sequence alignments of partly coding regions*. PhD thesis, University of Vienna, 2003.

[67] J Stoye, D Evers, and F Meyer. ROSE: Generating sequence families. *Bioinformatics*, 14(2):157–163, 1998.

[68] DL Swofford. *PAUP*. Phylogenetic Analyses Using Parsimony (*and other methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts, 1998.

[69] K Tamura and M Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10:512–526, 1993.

[70] JD Thompson, DG Higgins, and TJ Gibson. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[71] U Truyen. Emergence and recent evolution of Canine Parvovirus. *Veterinary Microbiology*, 69(1–2):47–50, Sep 1999.

[72] U Truyen. Personal communication, 2004.

[73] U Truyen, K Geissler, CR Parrish, W Hermanns, and G Siegl. No evidence for a role of modified live virus vaccines in the emergence of canine parvovirus. *Journal of General Virology*, 79(5):1153–1158, May 1998.

[74] U Truyen, A Gruenberg, SF Chang, B Obermaier, P Veijlainen, and CR Parrish. Evolution of the feline subgroup Parvoviruses and the control of canine host range *in vivo*. *Journal of Virology*, 69(8):4702–4710, Aug 1995.

[75] U Truyen, T Müller, R Heidrich, K Tackmann, and LE Carmichael. Survey on viral pathogens in wild Red Foxes (*Vulpes vulpes*) in Germany with emphasis on Parvoviruses and analysis of a DNA sequence from a Red Fox Parvovirus. *Epidemiology and Infection*, 121:433–440, 1998.

[76] U Truyen and CR Parrish. Canine and feline host ranges of Canine Parvovirus and Feline Panleukopenia Virus. *Journal of Virology*, 66(9):5399–5408, Sep 1992.

[77] J Tsao, MS Chapman, M Agbandje, W Keller, K Smith, H Wu, M Luo, TJ Smith, MG Rossmann, and RW Compans. The three-dimensional structure of Canine Parvovirus and its functional implications. *Science*, 251(5000):1456–1464, Mar 1991.

[78] J Tsao, MS Chapman, H Wu, M Agbandje, W Keller, and MG Rossmann. Structure determination of monoclinic Canine Parvovirus. *Acta Crystallographica Biologica*, D48(1):75–88, Feb 1992.

[79] M van Vuuren, A Steinel, T Goosen, E Lane, J van der Lugt, and U Truyen. Felkine panleukopenia virus revisited: Molecular characteristics and pathological lesions associated with three recent isolates. *Journal of the South African Veterinary Society*, 71(3):140–143, Sep 2000.

[80] M Vihinen-Ranta, D Wang, WS Weichert, and CR Parrish. The VP1 N-terminal sequence of Canine Parvovirus affects nuclear transport of capsids and efficient cell infection. *Journal of Virology*, 76(4):1884–1991, Feb 2002.

[81] A von Haeseler and D Liebers. *Molekulare Evolution*. Fischer, Frankfurt, May 2003.

[82] CH Wan, M Söderlund-Venermo, DJ Pintel, and LK Riley. Molecular characterization of three newly recognized Rat Parvoviruses. *Journal of General Virology*, 83:2075–2083, 2002.

[83] D Wang, W Yuan, I Davis, and CR Parrish. Non-Structural Protein 2 and the replication of Canine Parvovirus. *Virology*, 240(2):273–181, Jan 1998.

[84] Q Xie and MS Chapman. Canine Parvovirus capsid structure, analyzed at 2.9å resolution. *Journal of Molecular Biology*, 246:597–520, 1996.

Curriculum Vitae

Jennifer Hetzl

Personal Data

Date of Birth	June 15, 1977
Place of Birth	Villach, Austria
Citizenship	Austrian
Marital Status	Married

Education

2001–present	Diploma student of Genetics, University of Vienna
1999–2001	Diploma student of Biology, University of Graz
1995–1999	Diploma student of Civil Engineering,Technical University Graz
1995	General qualification for university entrance ("Matura") with distinction

Experiences

Apr–Jul 2004	SHK (IFI Leipzig, GER)
Aug 2003	Summer internship (TBI Vienna, AUT)
2000–2001	Internship (Sto AG Austria Headquarter)
1999	Opinion Research (GMK Graz, AUT)
1997–1998	Internship (Heraklith AG Austria Headquarter)
1994	Participation at High School exchange program and intensive language course (FRA)
1994	Participation at Mathematics intensive course ("Mathematikolympiade")

Presentations

09/2004	"The [Still] Unsolved Mysteries of Canine Parvovirus Evolution" IZBI Herbstseminar Chribska, CZ *Biochemistry and Bioinformatics*
02/2004	"The Neighbor-Joining Algorithm" TBI Winterseminar Bled, SLO *Computational Mathematics and Theoretical Biology*
01/2004	"Expression of Mammalian Host Intron Encoded Box C/D snoRNAs"
2003	"A Three-Zone Model for Protein Fold Use"