Molecular Evolution of Overlapping Genes

A Dissertation

Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Niv Sabath

December 2009

Molecular Evolution of Overlapping Genes

_____

Niv Sabath

APPROVED:

_____

Dr. Dan Graur, Chair

_____

Dr. Ricardo Azevedo

_____

Dr. George E. Fox

_____

Dr. Luay Nakhleh

_____

Dean, College of Natural Sciences and Mathematics

Acknowledgments

I thank Dr. Giddy Landan for his advice and discussions over many cups of coffee. I thank Dr. Jeff Morris for precious help in matters statistical, philosophical, and theological. I thank Dr. Eran Elhaik and Nicholas Price for enjoyable collaborations.

I thank Dr. Ricardo Azevedo, Dr. George Fox, and Dr. Luay Nakhleh, my committee members, for their help and advice.

I thank Hoang Hoang and Itala Paz for their tremendous support with numerous cases of computer failure.

Throughout the past five years, many people have listened to oral presentations of my work, critically read my manuscripts, and generously offered their opinions. In particular, I would like to thank Dr. Michael Travisano, Dr. Wendy Puryear, Dr. Maia Larios-Sanz, Lara Appleby, and Melissa Wilson. A special thanks to Debbie Cohen for her help in finding phase bias in the English language and musical analogies for overlapping genes.

Amari usque ad mare[1], researchers point to the fact that "you never really leave work" as one of the hardships of academic life. However, working on my *magnum opus*[2], rarely felt like "coming to work", but rather enlightening since *scientia est potentia*[3]. For that, I

thank my advisor, Dr. Dan Graur, whom I found to be a *rara avis*[4]. Unlike most advisors, who impose their specialized research interests on their students, Dan has encouraged me to explore various subjects and *quaere*[5] for a question that inspired me. Although the quest has been at times frustrating *ad nauseam*[6], it enabled me to *alis volare propriis*[7] and to find a topic that I am passionate about. Finally, *inter alia*[8], I have learned from Dan that *ars longa, vita brevis*[9] and that *quidquid latine dictum sit, altum sonatur*[10].

[1] From sea to sea
[2] A great work
[3] Knowledge is power
[4] Rare bird
[5] Seek
[6] To (the point of) seasickness or disgust
[7] Fly with my own wings
[8] Among other things
[9] Art (=science) is long, life is short
[10] Anything said in Latin sounds profound

Molecular Evolution of Overlapping Genes

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Niv Sabath

December 2009

Abstract

Overlapping genes are defined as a pair of protein-coding genes whose coding regions overlap on either the same strand or on the opposite strand. The sequence interdependence between two overlapping coding regions adds complexity to almost all molecular evolution analyses. Here, I use a comparative-genomic approach aimed at resolving several open questions concerning the evolution of overlapping genes. I demonstrate that estimates of selection intensity that ignore gene overlap are biased and that the magnitude and the direction of this bias is dependant on the type of overlap. I present a new method for the simultaneous estimation of selection intensities in overlapping genes. I show that overlapping genes are mostly subjected to purifying selection, in contradistinction to previous studies, which ignored the interdependence between overlapping reading frames and detected an inordinate prevalence of positive selection. Using simulation and two case studies, I show that this method can be used to distinguish between spurious and functional overlapping genes by using purifying selection as a tell-tale sign of functionality. In the first study, I test for the functionality of a hypothetical overlapping gene, which is central in the "Rosetta stone" hypothesis for the origin of the two aminoacyl tRNA synthetase classes from a pair of overlapping genes. I found no evidence of selection acting on the hypothetical gene, implying that the gene is non-functional, thus rejecting the "Rosetta stone" hypothesis. In the second study, I search for unannotated overlapping genes in viral genomes. I present evidence for the existence of a novel overlapping gene in the genomes of four viruses that infect

Hymenoptera. In another study, I present a method for the detection of selection signatures on hypothetical overlapping genes using population-level data. I apply the method to test whether the hypothetical gene in influenza A is under selection. Finally, I study a previously unexplained difference in the frequencies of overlapping genes of different types. I show that the structure of the genetic code and the abundance of different amino acids in proteins explain this difference between overlap types and lead to a correlation between overlap frequency and genomic composition.

Contents

Chapter One: General introduction

Information in protein-coding genes is contained within nucleotide triplets (codons) that are transcribed into RNA and eventually translated into amino acids, the building blocks of proteins. A DNA sequence can, therefore, be read in three reading-frames on one strand and three reading frames on the complementary strand potentially encoding six different proteins. When two or more proteins are encoded by a single DNA region, they are said to be encoded by overlapping genes. For example, Figure 1.1 shows a region of overlap between the *gag* and *pol* genes in the HIV genome.

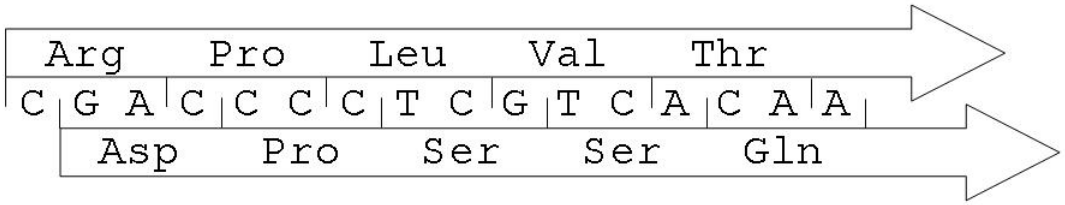

Figure 1.1: An example of gene overlap in the HIV genome, in which two proteins, *gag* (upstream) and *pol* (downstream), are encoded by two different reading frames.

Genes can overlap on the same strand or on opposite strands. In addition, overlaps can be "internal," in which one gene is entirely embedded within the other or "terminal," where both genes extend beyond the overlap region (Figure 1.2). Terminal overlaps on opposite strands can either be "tail-to-tail" or "head-to-head" overlaps (Figure 1.2).

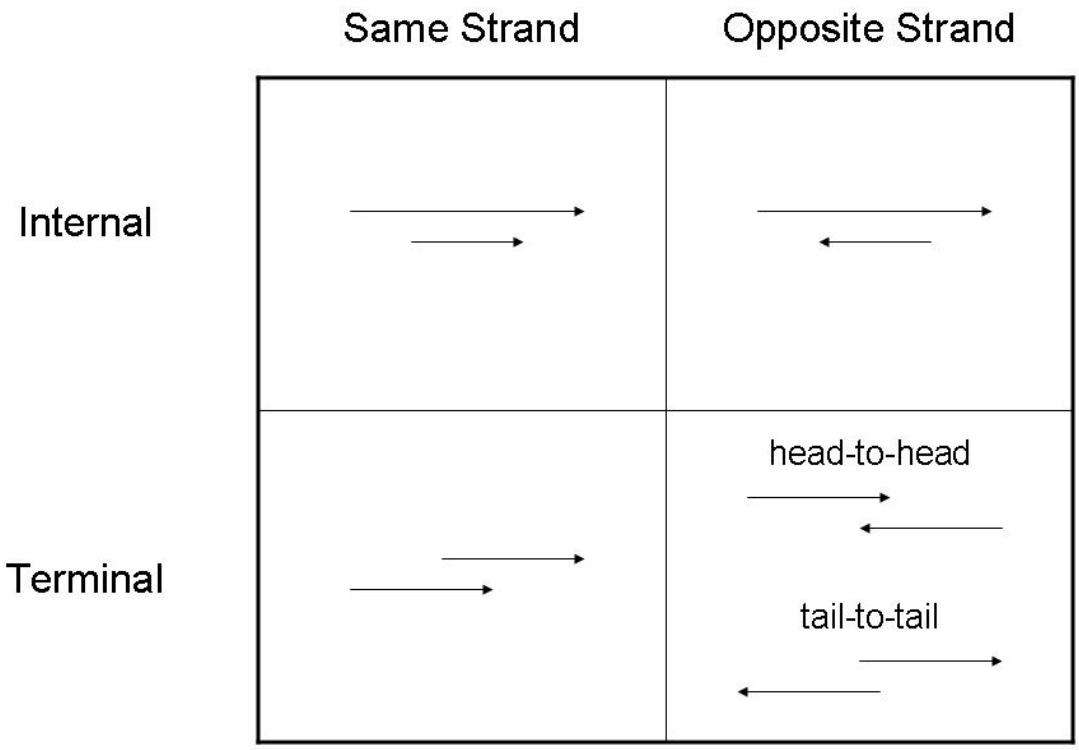

Figure 1.2: Overlap types. Genes are represented by arrows (5' → 3').

If we denote the reading frame of a gene as phase 0, there are five possible overlap phases (Figure 1.3). Same-strand overlaps occur in frameshifts of one nucleotide (phase 1) or two nucleotides (phase 2). Opposite-strand overlaps can be of three phases (0, 1, and 2).

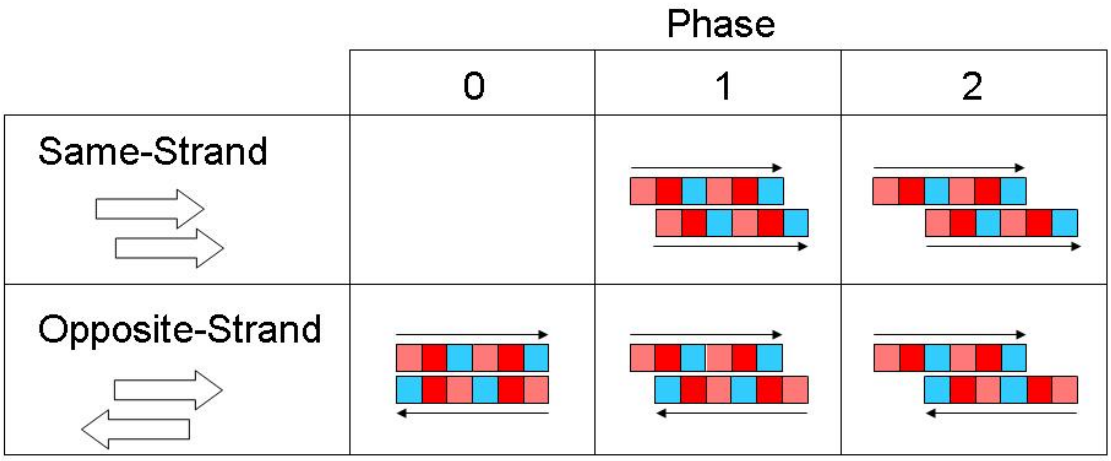

Figure 1.3: Orientations and phases of gene overlap. Genes can overlap on the same strand or on the opposite strand. The reference gene in a pair of overlapping genes is called phase 0. Same-strand overlaps can be of two phases (1 and 2); opposite-strand overlaps can be of three phases (0, 1, and 2). First and second codon positions, in which ~5% and 0% of the changes are synonymous, are marked in light and dark red, respectively. Third codon positions, in which ~70% of the changes are synonymous, are marked in blue.

The existence of overlapping genes was considered a plausible possibility long before such genes were actually discovered (Vandenberg 1967; Parkinson 1968). Overlapping genes were first discovered in viruses (Barrell, Air, and Hutchison 1976) and later in all cellular domains of life (Smith and Parkinson 1980; Montoya, Gaines, and Attardi 1983; Jones et al. 1995). The abundance of overlapping genes in a genome varies across species. In eukaryotes, the percentage of genes that are involved in overlap is 5–14% (Table 1.1) and most of the overlaps are on opposite strands (Chen and Stein 2006; Makalowska, Lin, and Hernandez 2007; Nagalakshmi et al. 2008). Makalowska, Lin,

and Hernandez (2007) have examined the conservation of overlapping genes across several eukaryotic genomes and showed that overlapping genes are often species-specific (Table 1.2).

Table 1.1: Overlapping genes in eukaryotic genomes. The percentage of genes, which are involved in overlap and the percentage of overlaps, which are exon-exon, are given in parenthesis.

Study	Species	Number of genes	Number of genes in overlap (%)	Number of overlaps	Exon-exon (%)
Makalowska et al. 2007	Human	22291	2978 (13.4)	1766	634 (35.9)
	Chimpanzee	21506	2219 (10.3)	1276	479 (37.5)
	Mouse	25383	3456 (13.6)	2053	819 (39.9)
	Rat	22159	1080 (4.9)	607	102 (16.8)
	Chicken	17709	1960 (11.1)	1135	511 (45.0)
	Fugu	20796	993 (4.8)	556	290 (52.2)
	Zebrafish	23524	1625 (6.9)	1026	98 (9.6)
Chen and Stein 2006	*C. elegans*	21188	2380 (11.2)	1190	5 (0.4)
Nagalakshmi et al. 2008	*S. cerevisiae*	4646	550 (11.8)	275	NA

Table 1.2: Conservation of overlapping genes in eukaryotic genomes (adapted from Makalowska, Lin, and Hernandez 2007). Above diagonal shows total number of conserved overlaps, and below diagonal shows numbers of conserved exon/exon overlaps.

	Human	Mouse	Rat	Chicken	Fugu	Zebrafish
Human	-	274	98	64	23	17
Mouse	146	-	141	76	26	16
Rat	11	48	-	45	19	6
Chicken	9	10	2	-	22	13
Fugu	1	0	0	0	-	13
Zebrafish	5	5	0	0	1	-

In bacteria, the number of overlaps is strongly correlated with the number of ORFs (Figure 1.4) (Fukuda, Nakayama, and Tomita 2003; Johnson and Chisholm 2004). Johnson and Chisholm (2004) showed that ~85% of the overlaps in bacteria are shorter than 30 bases and that ~83% of them are on the same strand. Overlaps on the same strand are more abundant because, on average, 70% of the genes in bacterial genomes, are located on one strand (Fukuda, Nakayama, and Tomita 2003).

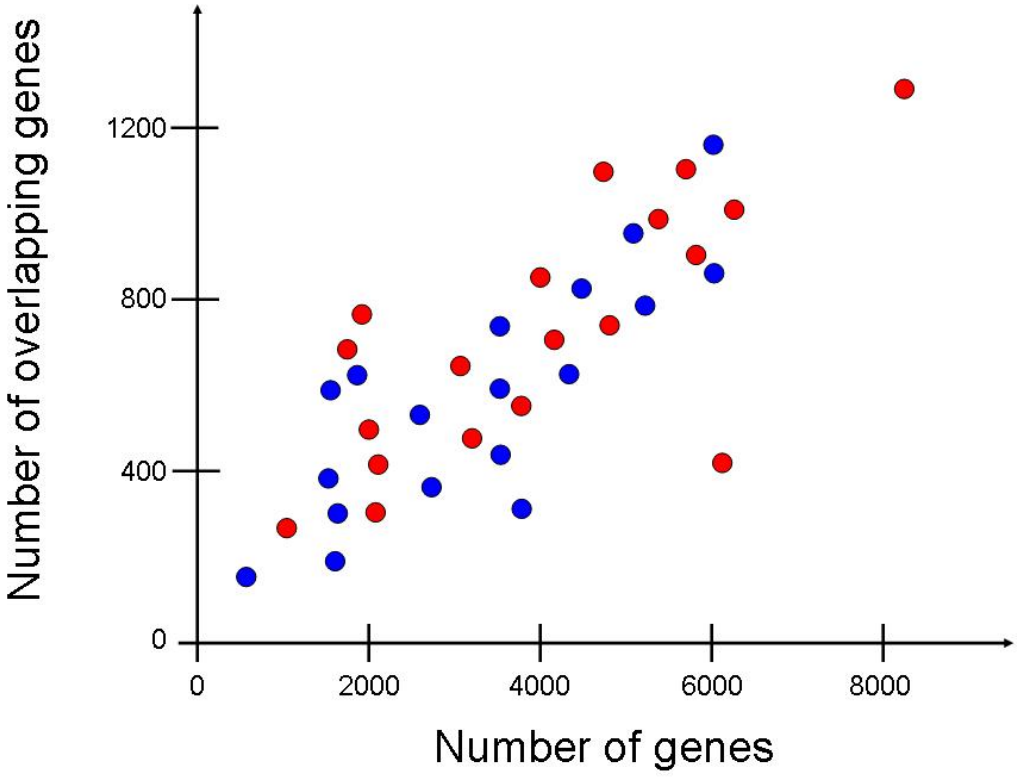

Figure 1.4: Correlation between the total number of genes and the number of overlapping gene pairs. Red: results using all genes; blue: results using only genes with high confidence in annotation (adapted from Fukuda, Nakayama, and Tomita 2003).

Lillo and Krakauer (2007) examined the characteristics of gene overlap in several archaeal and bacterial genomes. They found that Archaea have on average a smaller fraction of same-strand overlapping and non-overlapping consecutive genes (Figure 1.5). They suggested that this difference between Archaea and Bacteria may be related to the reduced frequency of operons in Archaea (Lillo and Krakauer 2007).

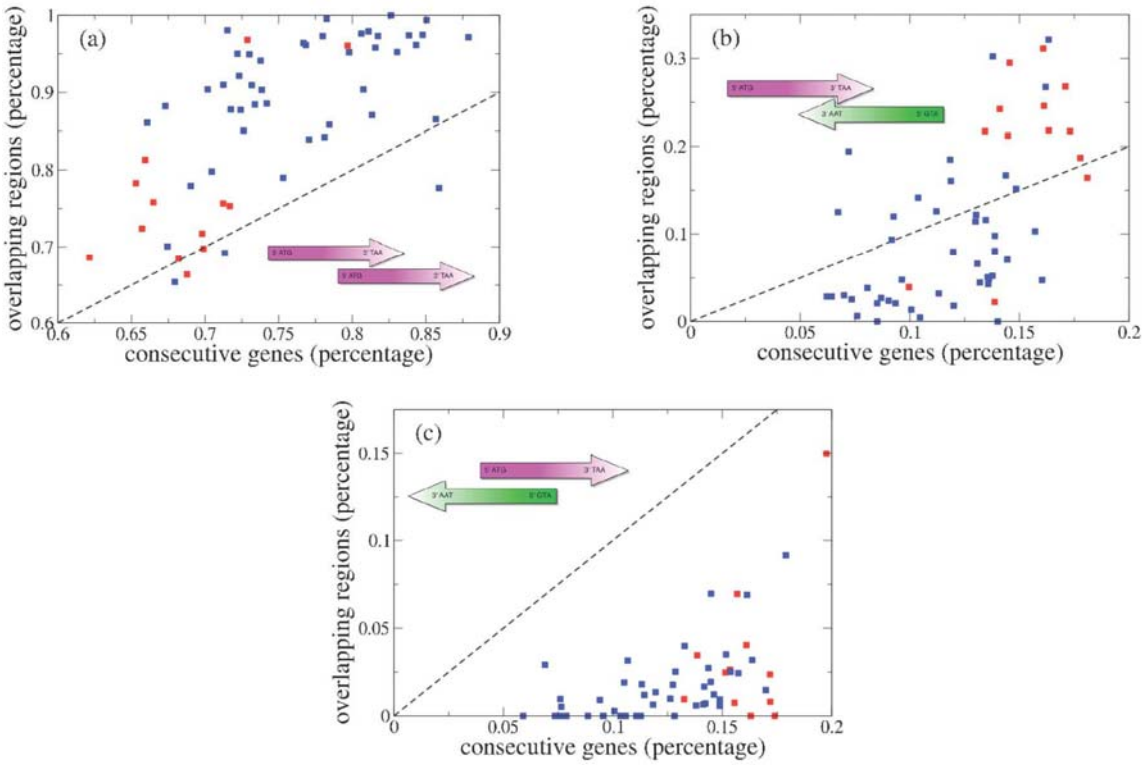

Figure 1.5: Percentage of overlapping genes in a given configuration (same-strand (a), opposite-strand tail-to-tail (b) and head-to-head (c)) versus the percentage of consecutive genes in the same configurations. The dashed lines are the *y* = *x* lines and serves as a guide to the eye for testing the hypothesis that the two percentages are equal. Red, Archaea; blue, Bacteria (Lillo and Krakauer 2007).

Unlike cellular organisms, in viruses the prevalence of overlapping genes is inversely correlated with genome size (Figure 1.6). For example the genome of Hepatitis B virus (Hepadnaviridae family, point 23 in Figure 1.6) contains four genes. Each of the genes overlaps with as least one other gene, leading to overlap proportion of ~12%.

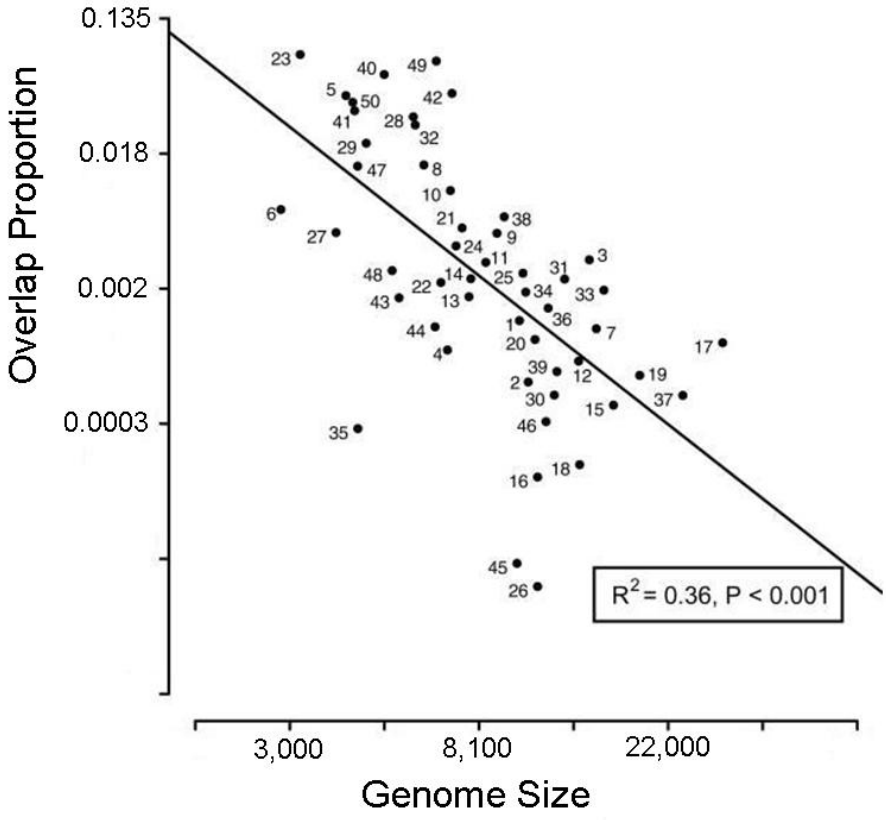

Figure 1.6: Relationship between overlap proportion and genome size, both presented in the scale of natural logarithms (adapted from Belshaw, Pybus, and Rambaut 2007).

Overlapping genes were suggested to have unique roles in numerous processes such as gene regulation (Normark et al. 1983; Boi, Solda, and Tenchini 2004), genome imprinting (Cooper et al. 1998), development of human diseases (Karlin et al. 2002), and

the evolution of the genetic code (Kozlov 2000). In addition, overlapping genes were hypothesized to be a means of genome size reduction (Sakharkar et al. 2005) and increasing complexity (Assis et al. 2008), as well as a mechanism for creating new genes (Keese and Gibbs 1992).

The main purpose of this study is to develop a framework for the evolutionary analysis of overlapping genes. I present a set of comparative-genomic methods aimed at resolving several open questions concerning the evolution of overlapping genes.

Inferring the intensity of negative and positive selection acting on protein-coding genes is a fundamental task in molecular evolution, in particular, since positive Darwinian selection is used to shed light on the process of adaptation. The inference of selection intensity in overlapping genes is complicated by the sequence interdependence between two overlapping coding regions (Miyata and Yasunaga 1978; Smith and Waterman 1981), which vary among overlap types (Krakauer 2000). Several attempts at estimating selection intensity in overlapping genes reported inordinate degrees of positive selection (e.g., Hughes et al. 2001; Li et al. 2004; Campitelli et al. 2006; Obenauer et al. 2006). For example, *PB1-F2*, an overlapping gene in influenza A, was reported to have a rate of nonsynonymous substitutions, which is nine times higher than that of synonymous substitutions (Obenauer et al. 2006). In Chapter Two, I present a new method for the simultaneous estimation of selection intensities in overlapping genes (see also, Sabath, Landan, and Graur 2008). By simulation, I verify the accuracy of the method, test its limitations, and compare the possible outcomes of estimating selection without

accounting for gene overlap across different overlap types. I show that the appearance of positive selection is caused by assuming that selection operates independently on each gene in an overlapping pair, thereby ignoring the unique evolutionary constraints on overlapping coding regions.

Another problem is how to detect functional overlapping genes. I define functional protein-coding gene to be a region in the genome, which is transcribed into RNA and eventually translated into a protein. Because it is fairly common that at least one of the five possible overlapping reading frames of any gene (Figure 1.3) will contain an open reading frame (ORF) of a length that may be suitable to encode a protein, it is extremely difficult to decide whether an intact overlapping ORF is functional or spurious. The main reason for this difficulty is that the sequence of an overlapping gene is (by definition) constrained by the functional and structural requirements of another gene. As a result, numerous annotated overlapping genes were suspected to be spurious (Silke 1997; Palleja, Harrington, and Bork 2008; Williams, Wolfe, and Fares 2009), whereas several unannotated overlapping genes were subsequently identified as bona fide protein-coding genes (Chung et al. 2008; Firth 2008; Firth and Atkins 2008b; Firth and Atkins 2008a; Firth and Atkins 2009). In Chapter Three, I demonstrate how my method for the estimation of selection intensity can be incorporated to distinguish between functional and spurious overlapping genes. Subsequently, I use the method to tackle the sense–antisense hypothesis for the origin of the two aminoacyl tRNA synthetase classes (Rodin and Ohno 1995; Carter and Duax 2002).

In Chapter Four, I use the method to scan viral genomes for overlapping genes that were missed in annotation (see also Sabath, Price, and Graur 2009). I present a discovery of a new overlapping gene in the genomes of Israeli Acute Paralysis Virus (IAPV) and three other viruses. IAPV infects honeybees and is associated with colony collapse disorder, a syndrome characterized by the mass disappearance of bees from hives.

The method presented in Chapters Two, Three, and Four is limited to the analysis of sequences from divergent species. In some cases, the question of functionality is asked for an overlapping gene, which is unique to a population of clinically important viruses and bacteria. One such interesting case is that of influenza A. An ORF in the negative strand of segment eight of influenza A viruses was noted when this segment was first sequenced (Baez et al. 1980). The ORF, which is commonly found in human influenza A viruses, is absent from non-human influenza A viruses (e.g., avian) as well as from influenza B and C viruses. Recently, it was suggested that the ORF codes a functional gene (Zhirnov et al. 2007; Clifford, Twigg, and Upton *in press*). In Chapter Five, I present a method for the detection of selection signatures on hypothetical overlapping genes using population-level data. I test the method on both known and spurious overlapping genes. Finally, I apply the method to test whether the hypothetical gene in influenza A is under selection.

In Chapter Six, I deal with the factors that influence the phase-distribution of overlapping genes. Krakauer (2000) defined the freedom for each gene to evolve independently (protein evolvability) according to the probability for changes, which are

nonsynonymous in one gene and synonymous in the overlapping gene. He showed that overlapping genes in different orientations and phases differ in the freedom for each gene to evolve independently (Figure 1.7). Therefore, he suggested that the variation in protein evolvability would be reflected in the frequency of the overlap phases. For example, in the case of opposite-strand overlaps, phase 1 in which the second codon position of one gene corresponds to the third codon position of the second gene (and vice versa), maximizes the freedom of each gene to evolve independently (Krakauer 2000) (Figure 1.7). In support of this model, Rogozin et al. (2002) found that among opposite-strand overlaps in bacteria, the most evolvable overlap phase (phase 1) was the most abundant. However, this model failed to explain the phase-distribution of same-strand overlaps in bacteria (Johnson and Chisholm 2004; Cock and Whitworth 2007). Cock and Whitworth (2007) attributed the unexpected phase-distribution to either gene location or to an unspecified selective advantage. Using a large set of bacterial genomes, I present a model that explains the phase-distribution of same-strand overlapping genes by compositional factors (i.e., amino-acid frequencies and codon usage) without invoking selection (see also, Sabath, Graur, and Landan 2008).

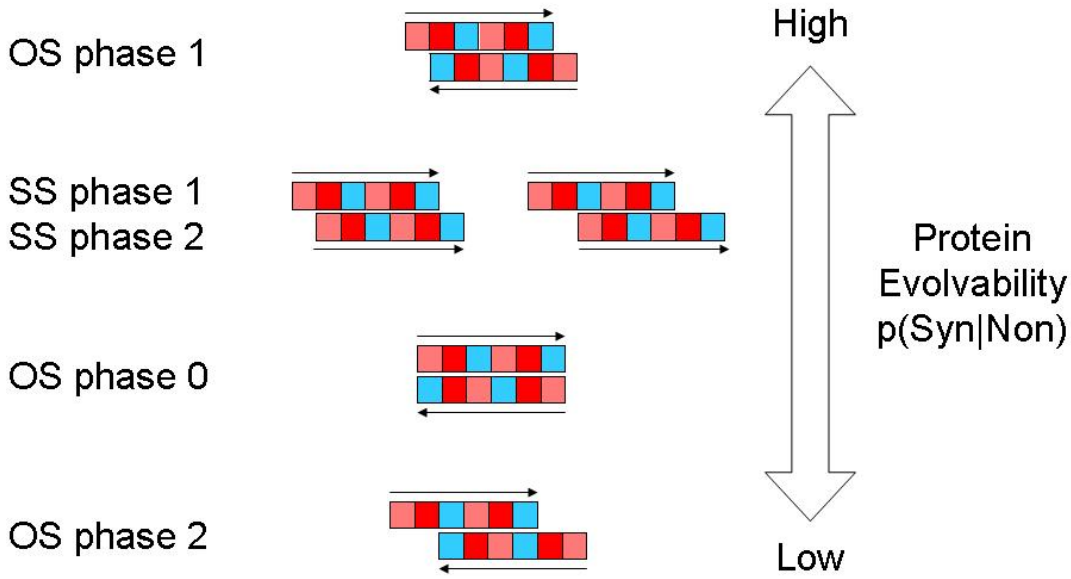

Figure 1.7: Overlapping genes in different orientations and phases differ in the freedom for each gene to evolve independently.

Finally, in Chapter Seven, I discuss my work in the light of our current knowledge of overlapping genes. I list several future lines of research, which I believe, will advance our understanding of the evolution of overlapping genes and, more generally, the evolution of genome architecture.

Chapter Two: A method for the simultaneous estimation of selection intensities in overlapping genes

Abstract

Inferring the intensity of positive selection in protein-coding genes is important since it is used to shed light on the process of adaptation. Recently, it has been reported that overlapping genes, which are ubiquitous in all domains of life, exhibit inordinate degrees of positive selection. Here, I present a new method for the simultaneous estimation of selection intensities in overlapping genes. I show that the appearance of positive selection is caused by the assumption that selection operates independently on each gene in an overlapping pair, thereby ignoring the unique evolutionary constraints on overlapping coding regions. This method uses an exact evolutionary model, thereby voiding the need for approximation or intensive computation. I test the method by simulating the evolution of overlapping genes of different types as well as under diverse evolutionary scenarios. The results indicate that the independent estimation approach leads to the false appearance of positive selection even though the gene is in reality subject to negative selection. Finally, I use the method to estimate selection in two influenza A genes for which positive selection was previously inferred. I find no evidence for positive selection in both cases.

Introduction

The interdependence between two overlapping coding regions results in unique evolutionary constraints (Miyata and Yasunaga 1978; Smith and Waterman 1981), which vary among overlap types (Krakauer 2000). Several attempts at estimating selection intensity in overlapping genes have been made (Hughes et al. 2001; Guyader and Ducray 2002; Li et al. 2004; Hughes and Hughes 2005; Narechania, Terai, and Burk 2005; Campitelli et al. 2006; Holmes et al. 2006; Obenauer et al. 2006; Pavesi 2006; Suzuki 2006; Pavesi 2007; Zaaijer et al. 2007). In some studies, one gene was found to exhibit positive selection while the overlapping gene showed signs of strong purifying selection (e.g., Hughes et al. 2001; Li et al. 2004; Campitelli et al. 2006; Obenauer et al. 2006). Inferences of positive selection in overlapping genes have been questioned (Holmes et al. 2006; Suzuki 2006; Pavesi 2007), mostly because ignoring overlap constraints might bias selection estimates. Rogozin et al. (2002) tried to overcome this problem by focusing on sites in which all changes are synonymous in one gene and nonsynonymous in the overlapping gene. This method, however, is only practical when dealing with one type of overlap.

A model for the nucleotide substitutions in overlapping genes was introduced by Hein and Stovlbaek (1995), who followed approximate models for non-overlapping genes that classify sites according to degeneracy classes (Li, Wu, and Luo 1985; Nei and Gojobori 1986; Pamilo and Bianchi 1993). This model was later incorporated into a method for

annotation of viral genomes (McCauley and Hein 2006; de Groot, Mailund, and Hein 2007; McCauley et al. 2007), and recently used for estimating selection on overlapping genes (de Groot et al. 2008). The main weakness of approximate methods is that it assumes a constant degeneracy class for each site, whereas degeneracy changes over time as substitutions occur. Pedersen and Jensen (2001) suggested a non-stationary substitution model for overlapping reading frames that extended the codon-based model of Goldman and Yang (1994). This model encompasses the evolutionary process more accurately than the approximate model (Hein and Stovlbaek 1995) by accounting for position dependency of each site in an overlap region (Pedersen and Jensen 2001). However, this improvement disallowed the straightforward estimation of parameters and forced the authors to apply a computationally-expensive simulation procedure (Pedersen and Jensen 2001). Surprisingly, these models for nucleotide substitutions in overlapping genes were rarely cited, not to mention used, by the majority of studies estimating selection in overlapping genes. One reason that these methods were seldom used might be the lack of an accessible implementation.

Here, I describe a non-stationary method, similar to that of Pedersen and Jensen (2001). The method simplifies selection estimation and avoids the need for costly simulation procedure. I test the method by simulating the evolution of overlapping genes of different types and under various selective regimes. Further, I describe the nature and magnitude of the error when selection is estimated as if the genes evolve independently. Finally, I use the method to estimate selection in two cases for which independent estimation has previously yielded indications of positive selection.

Methods

A gene can overlap another on the same strand or on the opposite strand. Each overlap orientation has 2 or 3 possible overlap phases (Figure 1.3). To understand the consequences of estimating selection pressures on overlapping genes as if they are independent genes, let us consider a simplified view of the genetic code, in which all changes in first and second codon positions are nonsynonymous and all changes in third codon position are synonymous. In reality, the proportions of changes that are synonymous are ~5%, 0%, and ~70% for the first, second, and third codon positions, respectively. From Figure 1.3 we see that in all overlap types, but one (opposite-strand phase 2), all synonymous changes in one gene are nonsynonymous in the overlapping gene, while half of the nonsynonymous changes are synonymous in the overlapping gene. Since the rate of synonymous substitutions is usually higher than that of nonsynonymous substitutions, ignoring overlap constraints would result in the underestimation of the rate of synonymous substitutions. In the case of opposite-strand phase-2 overlaps, ignoring the overlap would result in the underestimation of nonsynonymous substitutions rate. The bias in the estimation would be correlated with the strength of purifying selection on the overlapping gene. Thus, a false inference of positive selection is likely for genes under relaxed purifying selection when the overlapping gene is under strong purifying selection.

Goldman and Yang's (1994; 2006) method for the estimation of selection intensity in non-overlapping coding sequences

The most commonly used method for estimating selection intensity on protein coding genes fits a Markov model of codon substitution to data of two homologous sequences (Goldman and Yang 1994; Yang 2006). The codon-based model of nucleotide substitution is specified by the substitution-rate matrix, $Q_{codon} = \{q_{ij}\}$, where q_{ij} is the instantaneous rate of change from codon *i* to codon *j*.

(1)

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ k\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega k\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases}$$

Here, k is the transition/transversion rate, ω is the nonsynonymous/synonymous rate ratio (dN/dS), and π_j is the equilibrium frequency of codon *j*, which can be estimated from the sequence data by several models (Fequal, F1x4, F3x4, and F61, reviewed in Yang 2006). Parameters π_j and k characterize the pattern of mutations, whereas ω characterizes selection on nonsynonymous mutations. Q_{codon} is used to calculate the transition-probability matrix

(2) $$P(t) = \{p_{ij}(t)\} = e^{Q_{codon}t},$$

where $p_{ij}(t)$ is a probability that a given codon *i* will become *j* after time *t*. Parameters k, t, and ω are estimated by maximization of the log-likelihood function

(3) $$\ell(t) = \sum_i \sum_j n_{ij} \log\{\pi_i p_{ij}(t)\},$$

where n_{ij} is the number of sites in the alignment consisting of codons *i* and *j*. The estimated parameters are then used to calculate *dN* and *dS* (Yang 2006).

A new method for the simultaneous estimation of selection intensities in overlapping genes

I follow the maximum likelihood approach of Goldman and Yang (1994; 2006) to construct a model that accounts for different selection pressures on the genes in the overlap. I start with the simplest case, that of opposite-strand phase-0 overlaps. The reason this is the simplest case is that each codon overlaps only one codon in the overlapping gene. The substitution of nucleotides in opposite-strand phase-0 overlaps is specified by the substitution-rate matrix, $Q_{codon} = \{q_{ij}\}$, where q_{ij} is the instantaneous rate of change from codon *i* to codon *j*.

(5)

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion in both genes,} \\ k\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition in both genes,} \\ \omega_1\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion in gene A and synonymous in gene B,} \\ \omega_2\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion in gene B and synonymous in gene A,} \\ \omega_1 k\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition in gene A and synonymous in gene B,} \\ \omega_2 k\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition in gene B and synonymous in gene A,} \\ \omega_1\omega_2\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion in both genes,} \\ \omega_1\omega_2 k\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition in both genes.} \end{cases}$$

The main difference between this model and the single-gene model is that here I distinguish between two *dN*/*dS* ratios (ω_1 and ω_2 for gene 1 and gene 2, respectively). Another difference is the estimation of codon-equilibrium frequencies. Since the parameters of codon frequencies characterize processes that are independent of the selection on overlapping regions, I estimate these frequencies using the non-overlapping regions of each gene. The calculation of the transition-probability matrix and the log-likelihood function is done in the same way as in the single-gene model (equations 2 and 3).

The above model is a simple expansion of the single-gene model to account for opposite-strand overlaps in phase 0. However, this model cannot be used in the other four overlap cases, same-strand phase-1 and phase-2 overlaps and opposite-strand phase-

1 and phase-2 overlaps, because in all these cases a codon overlaps two codons of the second gene. Therefore, I set the unit of evolution to be a codon (the reference codon) and its two overlapping codons, which together constitute a sextet (Figure 2.1). The sextet is, therefore, the smallest unit of evolution in overlapping genes. In this model, each gene constitutes a set of sextets and within each sextet, only the reference codon is allowed to evolve. Changes in this codon affect the two overlapping codons. For example, consider the red and blue overlapping genes in Figure 2.1a. A change from G to A in position five (Figure 2.1a, bold) is illustrated in Figure 2.1b for the red gene as a reference and in Figure 2.1c for the blue gene as a reference. Restricting changes to the reference codon only is essential for the model, since changes outside the reference codon will require the consideration of other overlapping codons outside of the sextet, and so *ad infinitum*. In addition, this restriction allows the model to maintain the assumption that each reference codon evolves independently.

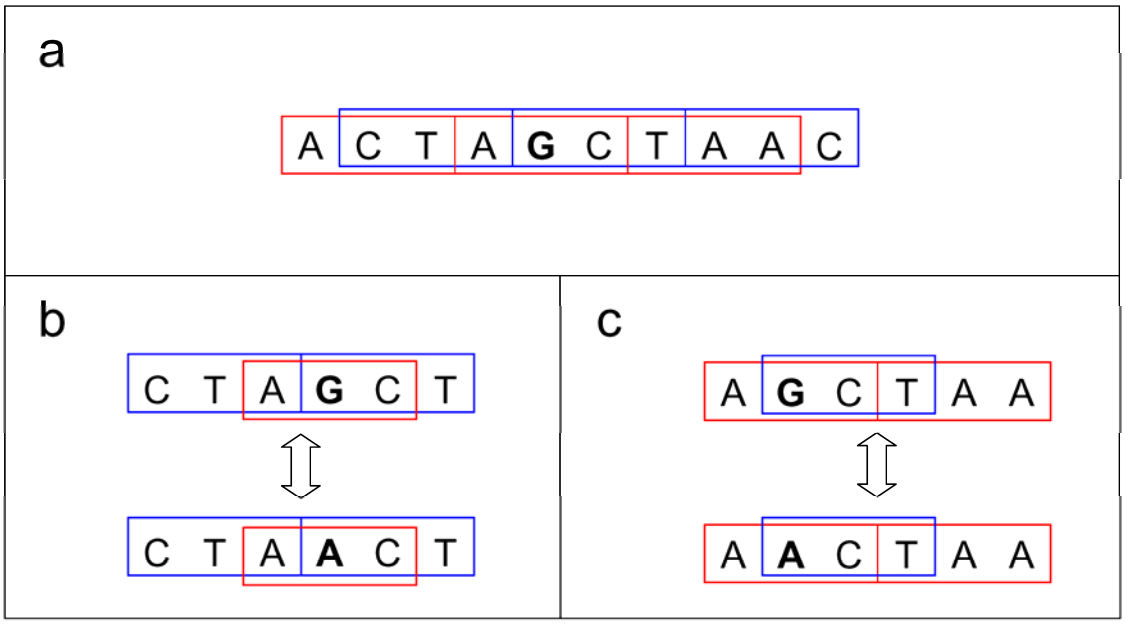

Figure 2.1: a. An overlapping gene pair (red and blue). b. The codon that is allowed to evolve is marked in red. The substitution in the second-codon position affects the overlapping codon in blue. c. The opposite situation in which only the codon marked in blue is allowed to change.

For gene A as the reference gene, I specify the substitution-rate matrix, $Q^{A}_{sextet} = \{q^{A}_{uv}\}$ where q^{A}_{uv} is the instantaneous rate from sextet u to sextet v with the codons of gene A as the reference codons:

(6)

$$q^{A}_{uv} = \begin{cases} 0, & \text{if } u \text{ and } v \text{ differ at two or three codon positions or at a position outside the reference codon,} \\ \pi_v, & \text{if } u \text{ and } v \text{ differ by a synonymous transversion in both genes,} \\ k\pi_v, & \text{if } u \text{ and } v \text{ differ by a synonymous transition in both genes,} \\ \omega_1\pi_v, & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transversion in gene A and synonymous in gene B,} \\ \omega_2\pi_v, & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transversion in gene B and synonymous in gene A,} \\ \omega_1 k\pi_v, & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transition in gene A and synonymous in gene B,} \\ \omega_2 k\pi_v, & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transition in gene B and synonymous in gene A,} \\ \omega_1\omega_2\pi_v, & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transversion in both genes,} \\ \omega_1\omega_2 k\pi_v, & \text{if } u \text{ and } v \text{ differ by a nonsynonymous transition in both genes.} \end{cases}$$

Similarly, I specify the substitution-rate matrix, $Q^{B}_{sextet} = \{q^{B}_{uv}\}$ for gene B as the reference gene, where q^{B}_{uv} is the instantaneous rate from sextet *u* to sextet *v* with gene B codons as the reference codons. These substitution-rate matrixes, Q^{A}_{sextet} and Q^{B}_{sextet}, can be used to calculate transition-probability matrixes (equation 2). However, these transition-probability matrixes cannot be used directly in the maximization of a log-likelihood function (equation 3) because they do not allow changes between any two sextets (as required in a Markov process). For example, the transition probability between sextets AAAAAA and CAAAAA (where the reference codons at positions 3-5 are underlined) would be zero for any given time *t*, because changes at a position outside of the reference codon are not allowed. A similar difficulty led Pedersen and Jensen (2001) to use a complicated, computationally-expensive, simulation procedure to estimate model parameters. Hence, I use Q^{A}_{sextet} and Q^{B}_{sextet} to construct codon-based substitution-rate matrixes $Q^{A}_{codon} = \{ q^{A}_{ij} \}$ and $Q^{B}_{codon} = \{ q^{B}_{ij} \}$ by summing the rates

over all sextets that share the same reference codon. A similar approach was used by Yang et al. (1998) to construct an amino acid substitution-rate matrix from a codon substitution-rate matrix. Let *I* and *J* represent the sets of sextets whose reference codons are *i* and *j*, respectively, than, the substitution rate from codon *i* to codon *j* is

(7) $$q_{ij} = \sum_{u \in I, v \in J} q_{uv} .$$

Q^A_{codon} and Q^B_{codon} are used to calculate a transition-probability matrix for each of the genes as in equation 2.

(8) $$P^A(t) = \{p^A_{ij}(t)\} = e^{Q^A_{codon}\, t} \text{ and } P^B(t) = \{p^B_{ij}(t)\} = e^{Q^B_{codon}\, t} .$$

The new transition-probability matrixes are suitable for a maximization of a log-likelihood function since they allow transition between each two codons. $P^A(t)$ and $P^B(t)$ can be used separately to estimate model parameters in a log-likelihood function for each gene (equation 3). However, in order to use all the information in the data, I combine the two transition-probability matrixes to create the following log-likelihood function:

(9) $$\ell(t) = \sum_i \sum_j n^A_{ij} \log\left\{\pi^A_i p^A_{ij}(t)\right\} + \sum_i \sum_j n^B_{ij} \log\left\{\pi^B_i p^B_{ij}(t)\right\}$$

Here, π^A_i and π^B_i are the equilibrium frequencies of codons in gene A and gene B respectively, estimated from the non-overlapping regions of the genes. n^A_{ij} and n^B_{ij} are the number of sites in the alignment consist of codons *i* and *j* for gene A and gene B, respectively.

The method was implemented in Matlab and is available at http://nsmn1.uh.edu/~dgraur/Software.html. Running time is ~7 seconds for a pair of

aligned sequences of length 1000 codons. Similar to the single-gene model, this method can be extended to deal with multiple sequences in a phylogenetic context and to test hypotheses concerning variable selection pressures among lineages and sites (Nielsen and Yang 1998; Yang and Nielsen 1998; Zhang, Nielsen, and Yang 2005).

Results

Simulation studies

I tested the performance of the new method for simultaneous estimation of selection intensities in comparison to the independent estimation that does not account for gene overlap (as described in equation 1). I examined the effects of the nonsynonymous/synonymous rate ratio in each gene (ω_1 and ω_2), the transition/transversion rate ratio (*k*), and the degree of sequence divergence (*t*). In all of the methods, I used the F3x4 model (Yang 2006) to estimate codon equilibrium frequencies. For each set of parameters, I generated 100 replications of random overlapping gene pairs (each gene was 2000 codons in length with 1000 codons in the overlap) by sampling codons from a uniform distribution of sense codons. To simulate the evolution along a branch of length *t*, I divided the sequence of the overlapping gene pair into three regions: non-overlapping region of gene one, non-overlapping region of gene two, and overlapping region. For the non-overlapping regions, I calculated the transition-probability matrixes based on the non-overlapping model in equation 1. For

the overlapping region, I calculated the transition-probability matrixes (based on the overlapping models in equations 5 and 6). Using the three probability matrixes, I simulated nucleotide substitutions at each codon independently (Yang 2006).

Different selection pressures

To examine the effect of different selection pressures, I initially set $k = 1$ and $t = 0.35$, which resulted in a sequence divergence of ~10%. I set $\omega_1 = 0.2$ and varied ω_2 between 0.2 and 2. In Figure 2.2, I compare the simultaneous estimation of ω_1 and ω_2 (blue line) and the independent estimation (red line) to the true simulated value (X axis, dashed green line) in the five types of overlaps. Each data point is the median of 100 replications. I use the median rather than mean since ratios are not normally distributed. In all overlap types, the estimation of the new method is in near-perfect match to the simulated value (blue and green lines, Figure 2.2) and the bias in the independent estimation of ω_2 is greater than that of ω_1.

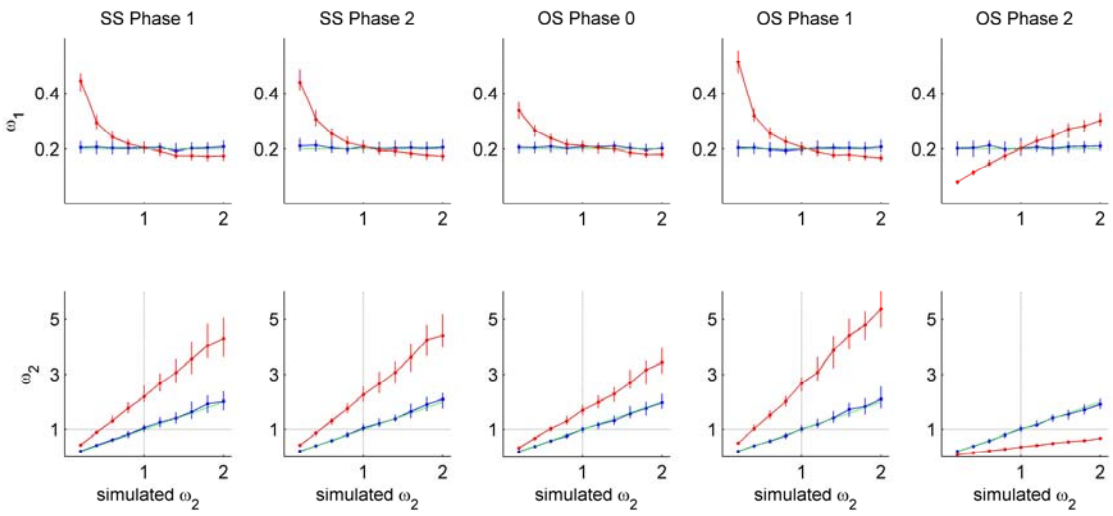

Figure 2.2: Simulation results in same-strand (SS) and opposite-strand (OS) overlaps. Estimations of the ratios of nonsynonymous to synonymous rates in the two genes (ω_1 and ω_2) by simultaneous estimation (blue line) and by independent estimation (red line) are plotted against the true value (X axis, dashed green line) for five types of overlap. The simulated value of ω_1 was set to 0.2 and ω_2 was varied between 0.2 and 2. *k* was set to 1 and *t* was set to 0.35. Each data point is the median of 100 replications. Vertical lines mark the lower and upper quartiles. Top: estimation of ω_1. Bottom: estimation of ω_2. Dotted black lines (X = 1 and Y = 1) illustrate the range of parameters that result in false inference of positive selection by independent estimation, i.e., when simulated $\omega_2 < 1$ and estimated $\omega_2 > 1$.

As expected, I found a similar pattern of bias in all overlap types except opposite-strand phase 2. In all of these overlap types (same-strand phase 1, same-strand phase 2, opposite-strand phase 0, and opposite-strand phase 1), the independent estimation of ω_1

is overestimated for $\omega_2 < 1$ and underestimated for $\omega_2 > 1$. The independent estimation of ω_2 is overestimated throughout the range of the simulation resulting in the false inference of positive selection in gene 2, while in reality this gene is under weak purifying selection. For example, the independent estimation of ω_2 in same-strand phase 1 is greater than one (apparent positive selection) for simulated values of ω_2 between 0.5 and 1.

The bias in opposite-strand phase 2 differs from the other overlap types because this overlap contains positions that are synonymous in both genes (Figure 1.3). Because of this factor, the independent estimation of ω_1 is underestimated for $\omega_2 < 1$ and overestimated for $\omega_2 > 1$. The independent estimation of ω_2 is underestimated throughout the range of the simulation, resulting in inability to detect positive selection in gene 2 for simulated values of $\omega_2 < 2$.

To compare the magnitude of error in the independent estimation of each overlap type, I set $k = 1$, $t = 0.35$, $\omega_1 = 0.2$, and $\omega_2 = 1$. I calculated the mean square error (MSE) for the independent estimation of ω_2 (the parameter whose estimation is most biased) in each overlap type. I use MSE because it measures both the bias and the variance. The most biased type is opposite-strand phase 1 followed by both same-strand phase 1 and phase 2, opposite-strand phase 0, and opposite-strand phase 2 (Table 2.1). As expected, the magnitude of error among overlap types is correlated with the proportion of sites in

each overlap type that are synonymous in one gene and nonsynonymous in the overlapping genes (Table 2.1).

Table 2.1: The mean square error (MSE) of the independent estimation of selection intensity is correlated with the proportion of changes that are synonymous in one gene and nonsynonymous in the overlapping gene (SN changes).

Orientation	Phase	Proportion of SN changes	MSE Independent	MSE Simultaneous
Same-Strand	1	47%	1.83	0.04
	2	47%	1.94	0.05
Opposite-Strand	0	43%	0.64	0.03
	1	63%	3.23	0.06
	2	39%	0.40	0.04

Transition/transversion rate ratio and sequence divergence

I tested the influence of transition/transversion rate ratio (k), and sequence divergence (t) on the performance of the new method for simultaneous estimation. Focusing on same-strand phase 1, I set $\omega_1 = 0.2$, $\omega_2 = 1$ and vary k between 1 and 20, and t between 0.1 and 1.1. I calculated the MSE for the estimation of ω_2. The results of 100 replications suggest that transition/transversion rate ratio does not affect the accuracy of the method, whereas the accuracy of the method is reduced for $t \leq 0.3$ (sequence divergence of ~8% or less, Figure 2.3). I note that although the new method performs well at high degrees of

sequence divergence, the inference of selection can be biased by the reduced quality in alignment of distant sequences.

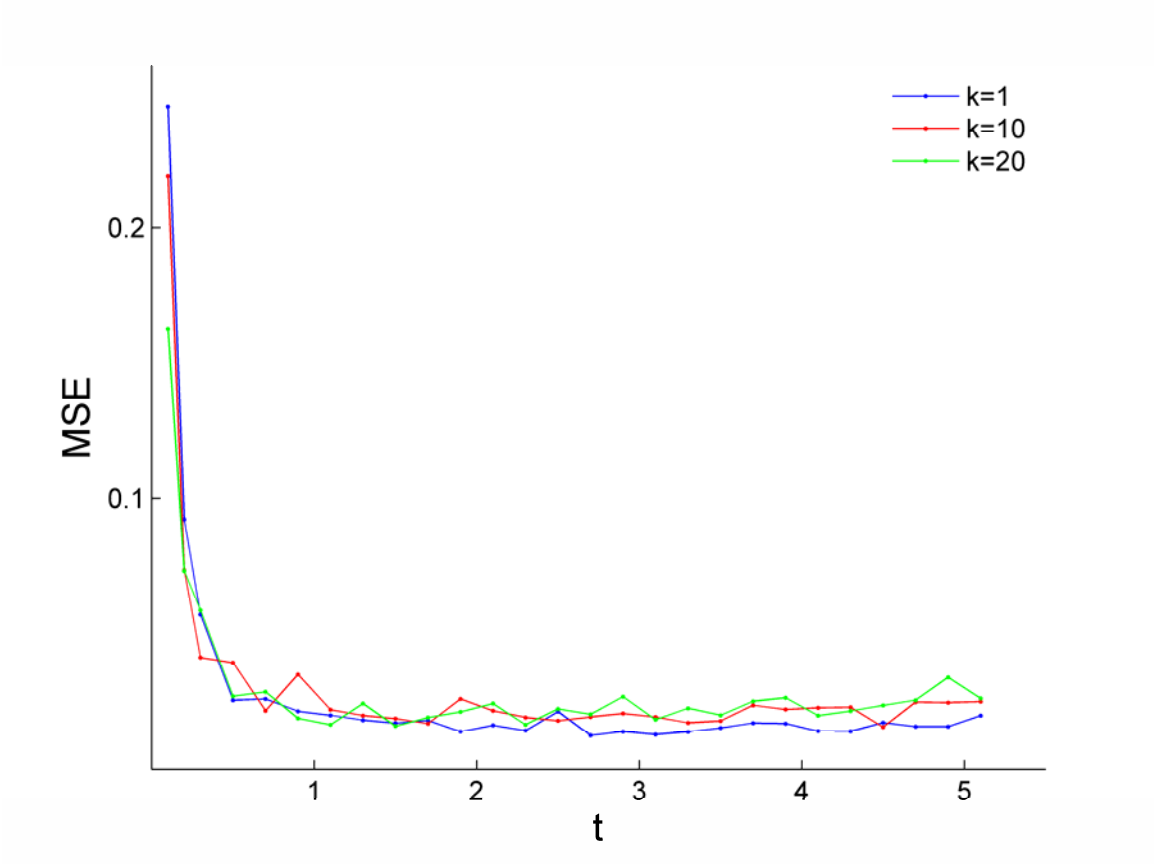

Figure 2.3: The influence of transition/transversion rate ratio (*k*), and sequence divergence (*t*) on the performance of the new method. The mean square error (MSE) is plotted against *t* for *k* = 1, 10, and 20 (blue, red, and green, respectively).

Testing the new estimation method on genes from influenza H5N1 and H9N2 strains

I used the new method to estimate selection pressures in two cases of overlapping genes in avian influenza A. I chose PB1-F2 and NS1 genes (which overlap with PB1 and NS2, respectively), because they were previously reported to exhibit values of dN/dS indicative of positive selection (Li et al. 2004; Campitelli et al. 2006; Obenauer et al. 2006; Pavesi 2007). For each gene, I collected all the annotated gene sequences from the two most sequenced subtypes, H5N1 and H9N2 from the NCBI Influenza Virus Resource (Bao et al. 2008). Within each subtype set, I aligned the overlapping regions of all gene pairs at the amino acid level using the Needleman and Wunsch (1970) algorithm. I used all pairwise alignments with sequence divergence greater than 5% (since estimation is less accurate at low divergence rates) to estimate selection intensities either simultaneously or independently (Table 2.2). Using higher cutoffs for sequence divergence did not affect the results (data not shown). Pairs in which the independent estimation of dS was zero (leading to infinity value for dN/dS) were excluded. In agreement with previous studies, PB1-F2 and NS1 genes appear to be under positive selection when gene overlap is not accounted for. However, by using the new method for simultaneous estimation, these genes seem to be under weak purifying selection. As predicted by the simulation, the bias in the independent estimation is dependent on the degree of purifying selection acting on the overlapping gene, leading to higher bias in PB1-F2 compared to NS1.

Table 2.2: Estimation of selection intensity ($\hat{\omega}$) by independent and simultaneous estimation.

Gene	Subtype[a]	Independent $\hat{\omega}$ [b, c]	Simultaneous $\hat{\omega}$ [b]
NS1	H5N1	**1.25** (0.75 1.93)	0.81 (0.41 1.52)
	H9N2	**1.46** (1.07 2.24)	0.58 (0.38 0.86)
NS2	H5N1	0.34 (0.24 0.52)	0.32 (0.22 0.50)
	H9N2	0.24 (0.15 0.35)	0.23 (0.13 0.36)
PB1-F2	H5N1	**6.75** (5.74 9.88)	0.52 (0.40 0.76)
	H9N2	**6.41** (5.52 7.92)	0.46 (0.34 0.75)
PB1	H5N1	0.03 (0.02 0.05)	0.02 (0.02 0.04)
	H9N2	0.03 (0.02 0.05)	0.02 (0.01 0.04)

[a]Number of pairwise alignments of NS1 – NS2 overlaps is 10,569 and 8,745 for H5N1 and H9N2 subtypes, respectively; Number of pairwise alignments of PB1-F2 and PB1 overlaps is 16,112 and 33,720 for H5N1 and H9N2 subtypes, respectively.

[b]Median of $\hat{\omega}$ over all pairwise comparisons. Lower and upper quartiles are noted in parentheses.

[c]Values of selection intensity in PB1-F2 and NS1 genes that appear as positive selection by independent estimation are bolded.

Discussion

Overlapping genes are widespread in all taxa, but are particularly common in viruses (Belshaw, Pybus, and Rambaut 2007). The sequence interdependence imposed by gene

overlap adds complexity to almost all molecular evolutionary analyses. Here, I presented a new method for the estimation of selection intensities in overlapping genes. By simulation, I verified the accuracy of the method, tested its limitations, and compared the possible outcomes of estimating selection without accounting for gene overlap across different overlap types. I find that estimating selection as if the genes are independent of one another results in the false appearance of positive selection. The new model can be used to identify true functional genes, which are usually under negative or positive selection, from among hypothetical overlapping ORFs, which are mainly spurious.

Chapter Three: Using signature of selection to detect functional overlapping genes

Abstract

As far as protein-coding genes are concerned, there is a non-zero probability that at least one of the five possible overlapping sequences of any gene will contain an open-reading frame (ORF) of a length that may be suitable for coding a functional protein. It is, however, very difficult to determine whether or not such an ORF is functional. In non-overlapping genes, the signature of purifying selection is used as a telltale sign of functionality. Here, I propose an analogous method that predicts functionality of an overlapping ORF if it can be shown that the sequence is subject to selection. Through simulation, I tested the method under several conditions and compared it with an existing method. I applied the method to test the hypothesis that the two aminoacyl tRNA synthetase classes have originated from a pair of opposite-strand overlapping genes. An overlapping ORF on the opposite strand of a heat shock protein 70 coding gene was claimed to be a central component of this hypothesis. I show that there is no signature of purifying selection acting on the overlapping ORF, suggesting that it is not a functional gene. Finally, I discuss the limits of applicability of the method. I conclude that the upper limits of applicability are reached at divergence rates above ~30%.

Introduction

Methods for the detection of protein-coding genes make use of three main properties: (1) the presence of an ORF; (2) expression of mRNA; and (3) conservation of ORFs between species. However, these properties are often uninformative in the case of overlapping genes, because: (1) non-functional ORFs that overlap functional genes are common; (2) non-functional overlapping ORFs are expressed when the overlap is on the same strand and, often, when the overlap is on opposite strands (Lavorgna et al. 2004); and (3) non-functional overlapping ORFs are conserved between species because of their functional overlapping genes.

As a result, annotation programs often fail to correctly predict functional overlapping genes (Delcher et al. 1999), and distinguishing between spurious and functional overlapping genes is of great interest. Silke (1997) showed that the frequency of opposite-strand overlapping genes in vertebrate genomes is highly influenced by genomic GC content and codon usage, suggesting that a large part of these genes may be spurious. Palleja, Harrington, and Bork (2008) examined the conservation of length between overlapping genes in different bacterial species and concluded that many of the long overlapping genes have been misannotated.

An interesting case of overlap is the one between heat shock protein 70 (*HSP70*) and its opposite-strand ORF (OS-ORF), which was reported in several species

(Konstantopoulou et al. 1995; Rother et al. 1997; Silke 1997; Monnerjahn et al. 2000; Carter and Duax 2002). This overlap was described as the "Rosetta stone" for the origin of the aminoacyl tRNA synthetase (aaRS) classes from opposite-strand overlapping genes. This hypothesis is based on proposed similarity between *HSP-70* and OS-ORF to the two aaRS classes (Carter and Duax 2002) (Figure 3.1). Recently, the functionality of OS-ORF was questioned by Williams, Wolfe, and Fares (2009), which presented several lines of evidence, most notably the discontinuous phylogenetic distribution of the gene, suggesting that OS-ORF is spurious.

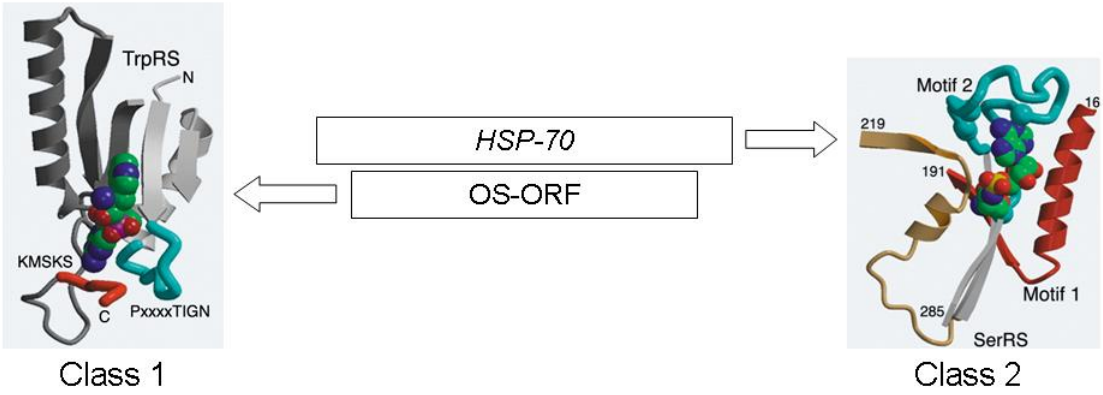

Figure 3.1: The "Rosetta stone" hypothesis. The two aminoacyl tRNA synthetase (aaRS) classes were proposed to originate from opposite-strand overlapping genes, based on a perceived similarity between *OS-ORF* and class 1 aaRS, on the one hand, and between *HSP70* and class 2 aaRS, on the other hand (Rodin and Ohno 1995; Carter and Duax 2002, images from Carter and Duax 2002)

Firth and Brown (2005) were the first to use selection to detect functional overlapping genes. Their method (FB), which is suitable for sequence pairs, calculates several statistics for each particular pairwise sequence alignment and uses a Monte Carlo simulation to determine whether the sequence is single-coding or double-coding. This method was later applied to multiple sequences by choosing only neighboring terminal taxa in the phylogenetic tree (Firth and Brown 2006). With the FB method, possible novel overlapping genes were discovered in Potyviridae (Chung et al. 2008) and other viral clades (Firth 2008; Firth and Atkins 2008b; Firth and Atkins 2008a; Firth and Atkins 2009). Other methods that make use of selection signatures to detect functional overlapping genes were proposed (de Groot, Mailund, and Hein 2007; McCauley et al. 2007; de Groot et al. 2008), but these methods have seldom been used, probably due to the lack of accessible implementation.

Here, I present a new method for the detection of functional overlapping genes. Through simulation, I tested the method under several conditions and compared it with the FB method. Finally, I examine the "Rosetta stone" and use the method to test whether or not the OS-ORF is functional.

Methods

I utilize the method for the estimation of selection intensities in overlapping genes, which I presented in Chapter Two (see also, Sabath, Landan, and Graur 2008). This

method uses a maximum-likelihood framework to fit a Markov model of codon substitution to data from two aligned homologous overlapping sequences. To predict functionality of an ORF that overlaps a known gene, I modified an existing approach for predicting functionality in non-overlapping genes (Nekrutenko, Makova, and Li 2002). Given two aligned orthologous overlapping sequences, I estimate the likelihood of two hierarchical models. In model 1, there is no selection on the ORF. In model 2, the ORF is assumed to be under selection. The likelihood-ratio test is used to test whether model 2 fits the data significantly better than model 1, in which case, the ORF is predicted to be under selection and most probably functional.

Results and Discussion

Simulation

To test the performance of the new method (SG) and compare it to FB (Firth and Brown 2006). I simulated the evolution of overlapping genes (as described in Chapter Two). In each run of the simulation, one gene was designated as known and the second as hypothetical. I examined the effects of the following factors on the ability of the two methods to detect selection in the hypothetical gene: (1) nonsynonymous/synonymous rate ratios in the hypothetical gene and the known gene (ω_h and ω_k, respectively), (2) overlap types (same-strand (SS) phase 1 and 2 and opposite-strand (OS) phase 0, 1, and 2), (3) sequence divergence (t), and (4) sequence length.

I initially set the sequence length to 300 codons and $t = 0.4$, which corresponds to a sequence divergence of ~12%. I set ω_k to 0.2 and varied ω_h between 0.2 (strong purifying selection) and 1 (no selection). For each set of parameters, I generated 100 random pairs of overlapping genes. Sensitivity is defined as the percent of hypothetical genes under selection that were identified correctly by the method. Specificity is defined as the fraction of hypothetical genes that were incorrectly identified to be under selection when ω_h was set to 1 (i.e., no selection). The results are shown in Figure 3.2a. Each square presents the results for SB (solid blue: $p < 0.01$; dashed blue: $p < 0.05$) and FB (red) methods against ω_h (X axis). An ideal detector is exemplified by a dashed green line. Each data point is the percentage of runs in which the methods detected selection. The five overlap types are shown in each column. As expected, the sensitivity of both methods decreased with increase in ω_h. In all overlap types, SG exhibits a higher sensitivity than FB, up to ~80% in same-strand for $\omega_h = 0.4$. As expected, using SG with p-value of 0.05 (rather than 0.01), increase the method's sensitivity at the cost of lower specificity. For opposite-strand phase 2, both methods perform similarly. This phase is unique in that the third codon position of both genes corresponds and, thus, most changes are either nonsynonymous in both genes, or synonymous in both (Figure 1.3). This overlap phase was also reported to generate high rate of false-positive results (Firth and Brown 2006).

In the next three sets, I tested different values of t, ω_k, and sequence length, one parameter at a time. In Figure 3.2b, I present the performance of the methods at high

sequence divergence levels (t = 1, corresponding to a sequence divergence of ~24%). For both methods, the results are similar to those at low sequence divergence. In Figure 3.2c, I present the results for stronger selection level on the known gene (ω_k = 0.1). The performance of SG is similar to that in (a) and (b), whereas the sensitivity of FB is reduced in same-strand phase 1 and 2 and opposite-strand phase 0 and 1. In Figure 3.2d, I present the results for short sequence length (60 codons). Under these conditions, SG and FB perform similarly, with SG showing reduced sensitivity compared to (a), (b), and (c).

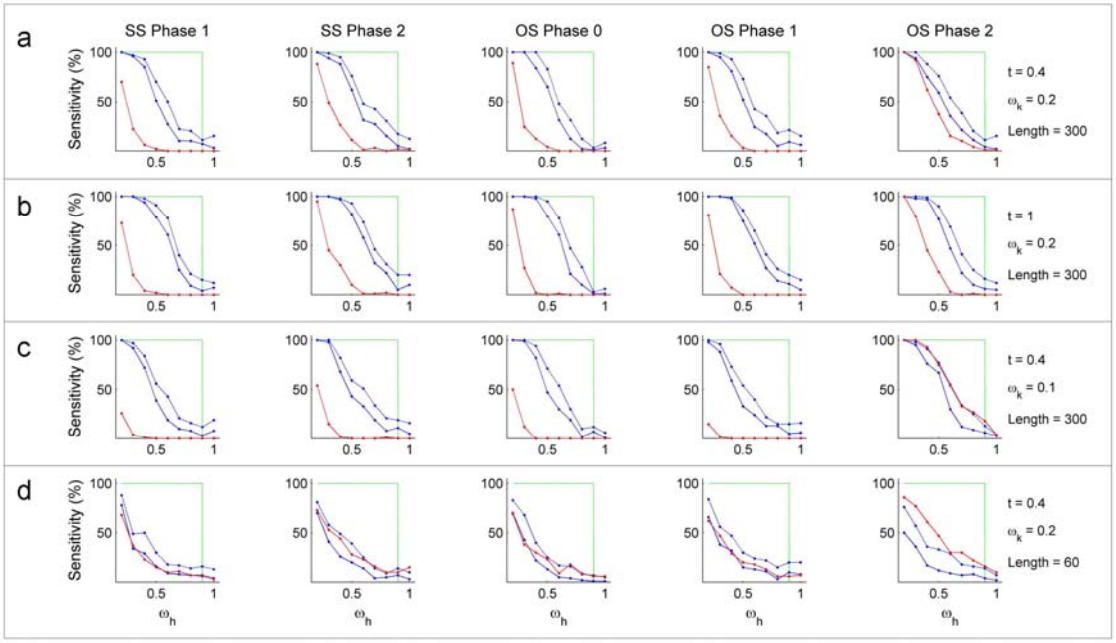

Figure 3.2: Detection of selection by SB (solid blue: p < 0.01; dashed blue: p < 0.05) and FB (red) methods on simulated genes. An ideal detector is illustrated by a dashed green line. Each data point is the percentage of runs for which the methods detected selection. The five overlap types are shown in each column. Four sets of

values for sequence divergence (t), ω_k, and sequence length are shown in each row (see text).

Overall, the simulation demonstrates that, under most conditions, SG performance is as good as FB or higher. The advantage of using SG over FB increases when the known gene is under strong purifying selection, whereas both methods perform alike on short sequences. In addition, SG was found to be more robust among overlap types in comparison to FB, whose performance is more variable, especially in the case of opposite-strand phase-2 overlaps. Similarly to FB, SG can be applied to multiple sequences by choosing only neighboring terminal taxa in the phylogenetic tree (Firth and Brown 2006). This approach, while ingenious, only indirectly addresses the phylogenetic framework and may be biased for trees with non-uniform branch-length distribution. In future studies, it would be beneficial to take full advantage of the maximum-likelihood framework that allows testing hypotheses concerning variable selection pressures among lineages and sites (Nielsen and Yang 1998; Zhang, Nielsen, and Yang 2005). This might be of special significance for overlapping genes because they may exist as a non-functional ORF before they become functional (Keese and Gibbs 1992).

Testing the functionality of functionality of the OS-ORF

The functionality of *HSP70* and its overlapping OS-ORF constitute a central tenet of the hypothesis concerning the origin of the two aaRS classes. A recent study by Williams, Wolfe, and Fares (2009) cast doubt on the functionality of the OS-ORF. I used SG,

which is a method that is different than the approach used by Williams, Wolfe, and Fares (2009), to ascertain whether selection operates on the OS-ORF. I identified 38 bacterial *HSP-70* genes with an intact OS-ORF and tested for selection on the OS-ORFs in all homologous pairs. The results are shown in Figure 3.3. For each pair, the amino-acid sequence divergence of the OS-ORFs was plotted against that of *HSP-70*. Pairs, for which the method did not detect selection, are marked in blue, and pairs, for which a signature of selection was found, are marked in red. The detection of selection signatures only in highly diverged pairs suggests that these are false positive results and that OS-ORF is not a functional gene.

The most likely reason for inaccuracy in high sequence divergence is that the method estimates the probability of one codon to change to another by summing over all possible paths. With the increase in divergence, the number of possible paths rises and, consequently, the power of the method to recover to true path decreases. These results imply that ~30% divergence between sequences should be the upper boundary for using the SG method. This boundary is comparable to the one suggested for exon detection using a single-coding genes analogous method (Nekrutenko, Makova, and Li 2002).

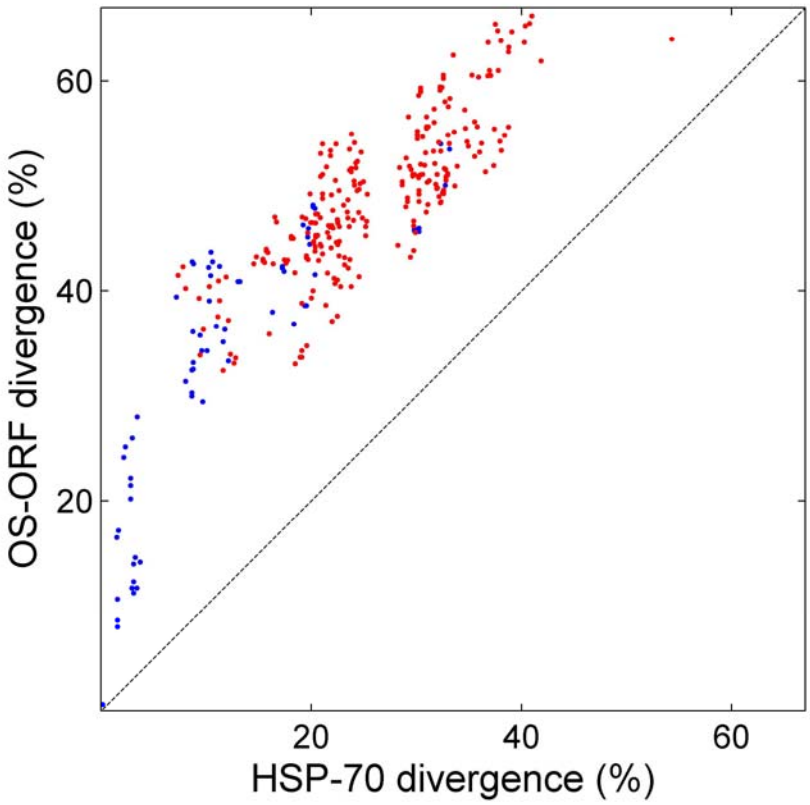

Figure 3.3: Testing for selection on OS-ORF. The amino-acid sequence divergence of the OS-ORF is plotted against that of *HSP-70* for all pairs of homologous sequences. Red: sequence pairs, for which the method detected selection. Blue: sequence pairs, for which the method did not detect selection.

In this chapter, I presented a new method for the detection of functional overlapping genes. By simulation, I compared the method to the FB method, and tested both methods across different overlap types. I found that under most conditions, my method predicts functionality with higher sensitivity while maintaining high specificity. Finally, I

conclude that OS-ORF is most likely not a functional gene and therefore cannot be regarded as the “Rosetta stone” for the overlap origin of the aaRS classes.

Chapter Four: A potentially novel overlapping gene in the genomes of Israeli acute paralysis virus and its relatives

Abstract

The Israeli acute paralysis virus (IAPV) is a bee-infecting virus that was found to be associated with colony collapse disorder. The IAPV genome was previously described to contain only two long open-reading frames encoding a structural and a nonstructural polyprotein. By using the method for the detection of functional overlapping genes, I provide evolutionary evidence for the existence of a third, overlapping gene. The new gene, which I provisionally call *pog* (*p*redicted *o*verlapping *g*ene), is translated in the +1 reading frame of the structural polyprotein gene. Conserved orthologs of this gene were also found in the genomes of a monophyletic clade that includes IAPV, acute bee paralysis virus, Kashmir bee virus, and *Solenopsis invicta* (red imported fire ant) virus 1. The discovery of a new gene may improve our understanding of this virus and its interaction with its host.

Introduction

Colony collapse disorder (CCD) is a syndrome characterized by the mass disappearance of honeybees from hives (Oldroyd 2007). CCD imperils a global resource valued at approximately $200 billion (Gallai et al. 2009). It has been estimated that up to 35% of hives in the US may have been affected (van Engelsdorp et al. 2008). Many culprits have been suggested as causal factors of CCD, among them fungal, bacterial, and protozoan diseases, external and internal parasites, in-hive chemicals, agricultural insecticides, genetically modified crops, climatic factors, changed cultural practices, and the spread of cellular phones (Oldroyd 2007).

The Israeli acute paralysis virus (IAPV), a positive-strand RNA virus belonging to the family Dicistroviridae, was found to be strongly correlated with CCD (Cox-Foster et al. 2007). It was first isolated in Israel (Maori et al. 2007)—hence the name—but was later found to have a worldwide distribution (Cox-Foster et al. 2007; Blanchard et al. 2008; Palacios et al. 2008).

The genome of IAPV contains two ORFs separated by an intergenic region. The 5' ORF encodes a structural polyprotein; the 3' ORF encodes a non-structural polyprotein (Maori et al. 2007). The non-structural polyprotein contains several signature sequences for helicase, protease, and RNA-dependent RNA polymerase. The structural polyprotein,

which is located downstream of the non-structural polyprotein, encodes two (and possibly more) capsid proteins.

Overlapping genes may be missed by annotation, even in genomes of highly studied viruses (Chen et al. 2001). Recently, several overlapping genes were detected using the signature of purifying selection (Chung et al. 2008; Firth 2008; Firth and Atkins 2008b; Firth and Atkins 2008a; Firth and Atkins 2009). Here, I apply the method for the detection of functional overlapping genes, which I described in Chapter Three, to the genome of IAPV and its relatives.

Methods

Sequence Data, Processing, and Analysis

Fourteen completely sequenced dicistrovirid genomes were obtained from NCBI (Table 4.1). Each genome was scanned for the presence of overlapping ORFs. I used BLASTP (Altschul et al. 1990) with the protein sequences of the known genes to identify matches of orthologous overlapping ORFs (E value < 10^{-6}). Matching overlapping ORFs were assigned into clusters. Within each cluster, I aligned the amino-acid orthologs by using the sequences of the known genes as references. If alignment length of the overlapping sequence exceeded 60 amino acids, and if the amino-acid sequence identity among the

hypothetical genes within a cluster was higher than 65%, I tested for signature of purifying selection on the hypothetical gene (as described in Chapter Three).

Table 4.1: A list of completely sequenced dicistroviruses used in this study

Name	Accession number
Israel acute paralysis virus (IAPV)	NC_009025
Acute bee paralysis virus (ABPV)	NC_002548
Kashmir bee virus (KBV)	NC_004807
Solenopsis invicta virus (SINV-1)	NC_006559
Black queen cell virus (BQCV)	NC_003784
Cricket paralysis virus (CrPV)	NC_003924
Homalodisca coagulata virus-1 (HoCV-1)	NC_008029
Drosophila C virus (DCV)	NC_001834
Aphid lethal paralysis virus (ALPV)	NC_004365
Himetobi P virus (HiPV)	NC_003782
Taura syndrome virus (TSV)	NC_003005
Plautia stali intestine virus (PSIV)	NC_003779
Triatoma virus (TrV)	NC_003783
Rhopalosiphum padi virus (RhPV)	NC_001874

I aligned the protein sequences of the two polyproteins with CLUSTAW (Thompson, Gibson, and Higgins 2002) as implemented in the MEGA package (Kumar et al. 2008). Alignment quality was confirmed using HoT (Landan and Graur 2007). I reconstructed

two phylogenetic trees (one for each polyprotein) by applying the neighbor joining method (Saitou and Nei 1987), as implemented in the MEGA package (Kumar et al. 2008). Trees were rooted by the mid-point rooting method (Farris 1972) and confidence of each branch was estimated by bootstrap with 1000 replications.

Motifs

I searched for motifs within the inferred protein sequences encoded by the overlapping ORF by using the motif search server (http://motif.genome.jp/) and the My-Hits server (http://hits.isb-sib.ch/cgi-bin/PFSCAN) with the following motif databases: PRINTS (Attwood et al. 2002), PROSITE (Hulo et al. 2006), and Pfam (Finn et al. 2008). I used PSIPRED (McGuffin, Bryson, and Jones 2000) to predict secondary structure, and MEMSAT (Jones 2007) to predict transmembrane protein topology.

Results and Discussion

In the fourteen completely sequenced dicistrovirus genomes (Table 4.1), I identified 43 overlapping ORFs of lengths equal or greater than 60 codons on the positive strand. Ten overlapping ORFs were found in concordant genomic locations in two or more genomes. The concordant overlapping ORFs were assigned into three orthologous clusters (Table 4.2). The overlapping ORFs in all three clusters are phase-1 overlaps, i.e., shifted by one nucleotide relative to the reading-frames of the known polyprotein genes. Two of the

orthologous clusters (B and C) overlap the gene encoding the nonstructural protein, and one cluster (A) overlaps the reading frame of the structural protein.

Table 4.2: Clusters of orthologous overlapping ORFs on the positive strand

Cluster	Virus	Start of ORF	End of ORF	Length (nucleotides)
A	IAPV	6589	6900	312
	ABPV	6513	6815	303
	KBV	6601	6909	309
	SINV-1	4382	4798	417
B	ABPV	5958	6227	270
	KBV	5974	6243	270
C	CrPV	2396	2614	219
	DCV	2216	2602	387
	HoCV-1	2377	2574	198
	PSIV	2333	2527	195

I identified a strong signature of purifying selection in cluster A that contains overlapping ORFs from four genomes: IAPV, Acute bee paralysis virus (ABPV), Kashmir bee virus (KBV), and *Solenopsis invicta* virus 1 (SINV-1) (Govan et al. 2000; de Miranda et al. 2004; Valles et al. 2004). This ORF overlaps the 5' end of the structural polyprotein gene (Figure 4.1a). The signature of selection was identified in the

three bee viruses (IAPV, ABPV, and KBV). The protein product of the orthologous ORF in SINV-1 could not be tested for selection because the amino acid sequence identity between the ORF from SINV-1 and the ORFs from the three bee viruses (Table 4.3) is lower than the range of sequence identities for which the method can be applied (65-95%).

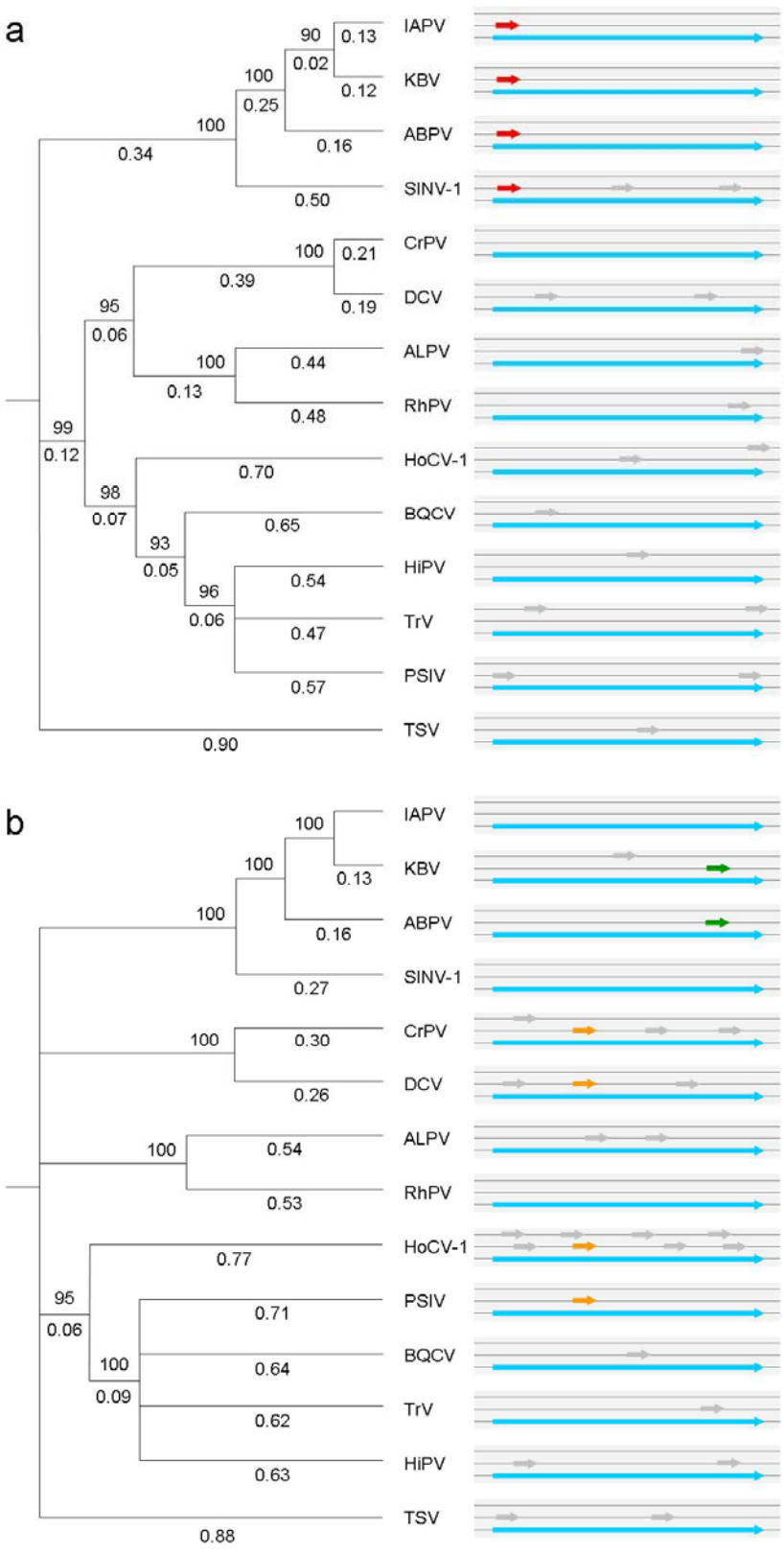
a
IAPV
KBV
ABPV
SINV-1
CrPV
DCV
ALPV
RhPV
HoCV-1
BQCV
HiPV
TrV
PSIV
TSV
b
IAPV
KBV
ABPV
SINV-1
CrPV
DCV
ALPV
RhPV
HoCV-1
PSIV
BQCV
TrV
HiPV
TSV

Figure 4.1: (see previous page) Phylogenetic trees and schematic representation of the dicistrovirid genomes (a. structural polyprotein; b. non-structural polyprotein). Trees were inferred using the neighbor joining method (Saitou and Nei 1987) and rooted by the mid-point rooting method (Farris 1972). Numbers above and below the branches are bootstrap values (1000 replications) and branch lengths (amino-acid substitutions per site), respectively. Phylogenetic analyses were conducted with MEGA (Kumar et al. 2008). The approximate locations and sizes of the known genes (blue), overlapping hypothetical genes (red, green, and orange), and singlet ORFs (gray) are noted in the three reading frames.

An additional indication for selection on these ORFs was obtained by comparing the degrees of conservation of the hypothetical protein sequences of the overlapping ORFs against the protein sequences of the known genes (Table 4.3). The degree of amino-acid conservation and, hence, sequence identity between orthologous protein-coding genes is influenced *ceteris paribus* by the intensity of purifying selection. If both overlapping genes are under similar strengths of selection, the amino-acid sequence identity of one pair of homologous genes would be similar to that of the overlapping pair. On the other hand, if a functional gene overlaps a non-functional ORF, the amino-acid identity between the hypothetical protein sequences of the non-functional ORFs would be much lower than that between the two homologous overlapping functional genes. I found that the degree of amino-acid conservation of the overlapping sequence identity between pairs of overlapping ORFs in cluster A is only slightly lower than that of the known gene

(Table 4.3). In contrast, the amino-acid sequence identity between ORF pairs in clusters B and C is much lower than that between the pairs of known genes (Table 4.3).

Table 4.3: Sequence conservation in comparisons of known orthologous proteins and orthologous products of overlapping ORFs.

Cluster	Genome pair		Identity of known proteins (%)	Identity of hypothetical product of overlapping ORFs (%)
A	IAPV	ABPV	80.2	74.8
	ABPV	KBV	79.3	75.6
	IAPV	KBV	77.4	72.5
	IAPV	SINV-1	42.7	30.3
	ABPV	SINV-1	41.6	32.6
	KBV	SINV-1	36.3	29.4
B	KBV	ABPV	87.7	52.3
C	CrPV	DCV	80.3	36.1
	HoCV-1	PSIV	64.3	40.0
	DCV	HoCV-1	56.4	28.8
	CrPV	HoCV-1	48.0	31.7
	DCV	PSIV	44.2	36.4
	CrPV	PSIV	35.7	25.0

The strong signature of purifying selection on the ORFs in cluster A suggests that they may encode functional proteins. I provisionally term this gene *pog* (*p*redicted *o*verlapping *g*ene). In Figure 4.1, I show that *pog* is found in the genomes of four viruses that constitute a monophyletic clade, but not in any other dicistrovirid genome (Figure 4.1a). Its phylogenetic distribution suggests that *pog* originated before the divergence of SINV-1 from the three bee viruses. The phylogenetic distributions of the ORFs in clusters B and C (Figure 4.1b) are patchy. I interpret this patchiness to indicate that the overlapping ORFs in clusters B and C are spurious, i.e., non-functional.

```
        1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19  20
IAPV    gaa cag ctg tac tgg gca gtt aca gca gtc gta tgg taa cac atg cgg cgt tcc gaa ata
ABPV    gaa cag cta tat tgg gta gtt gta gca gtt gta ttc aaa tga atg cag cgt tcc gaa ata
KBV     aaa ccg cta tat cgg gta gct ata gca gtc gga tag taa tat atc cgg cgt ttc gaa ata
SINV-1  tag cag tca gga tgt cat tct ggc gtt ccg aaa tac cca aac ctg ctc aat caa aca atg

        21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
IAPV    cca tgc ctg gcg att cac aac aag aaa gca ata ctc cca acg tac aca ata cgg aac tcg
ABPV    tca tac ctg ccg atc --- --- aag aaa caa ata ctt cca acg tac ata ata cgc aac tcg
KBV     cca tac ctg ctg ata --- acc aag aaa acg att cta cca atg tac ata aca cga aac tcg
SINV-1  cga ata ctt ttg aga cga aaa cgg caa caa cct ctg ctt ccc acg cac aat cgg aac tta
```

Figure 4.2: Codon alignment of the 5' overlap region between the structural polyprotein and the hypothetical gene. The alignment is shown in the reading frame of the hypothetical gene. The annotated initiation site of the polyproteins is underlined. The first potential initiation site (AUG or CUG) of the hypothetical genes is marked in red. The last stop codon at the +1 reading frames is marked in green.

An examination of the DNA alignment of *pog* (Figure 4.2) reveals a conservation of the first potential start codon (AUG or CUG) in the +1 reading frame in three out of the four viral genomes (IAPV, ABPV, and SINV-1). As seen in Figure 4.3, this conservation cannot be explained by constraints on the overlapping polyprotein, in which the corresponding site is variable and encodes different amino acids (His, Asn, and Pro, in IAPV, ABPV, and SINV-1, respectively). I note, however, that I did not find a conserved Kozak consensus sequence, which are often involved in the initiation of translation (Kozak 1983), upstream of the potential initiation site.

To predict the function of the new gene, I conducted a motif search, which resulted in several matches, all with a weak score. Two patterns were found in all four proteins: (1) a signature of rhodopsin-like GPCRs (G protein-coupled receptors), and (2) a protein kinase C phosphorylation site (Figure 4.3). Prediction of the secondary structures (McGuffin, Bryson, and Jones 2000) suggests that the proteins contain two conserved helix domains, separated by 3–5 residues (except for SINV-1, in which one long domain is predicted), at the C-terminus (Figure 4.3). A search for transmembrane topology (Jones 2007) indicates that the longer helix may be a transmembranal segment (Figure 4.3). Although viruses often use GPCRs to exploit the host immune system through molecular mimicry (Lalani and McFadden 1999; Murphy 2001; Hughes and Friedman 2003; McLysaght, Baldi, and Gaut 2003), the lengths of the proteins encoded by *pog* are shorter than the average virus-encoded GPCR. Therefore, these proteins may have a different function.

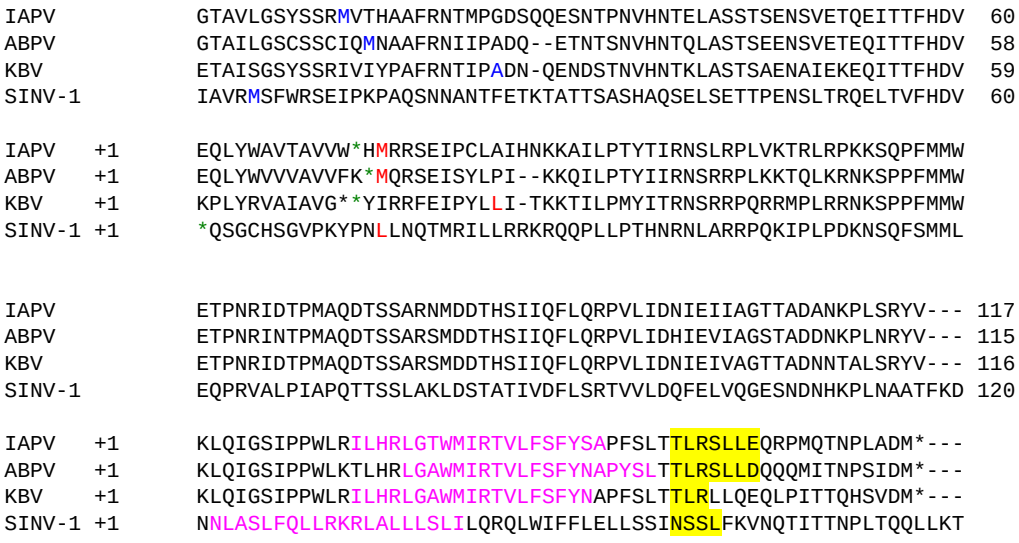

Figure 4.3: The amino-acid alignment of the overlap region between the structural polyprotein and the hypothetical gene (+1 reading frame). The annotated initiation site of the polyproteins is marked in blue. The first potential initiation site (AUG or CUG) of the hypothetical genes is marked in red. The last stop codon at the +1 reading frames is marked in green. Transmembranal helixes predicted by MEMSAT (Jones 2007) are marked in magenta. Conserved protein kinase C phosphorylation sites predicted through My-Hits server (http://hits.isb-sib.ch/cgi-bin/PFSCAN) are marked in yellow.

Overlapping ORFs on the negative strand

I also examined the overlapping ORFs on the negative strand, despite the fact that dicistroviruses are not known to be ambisense, i.e., RNA viruses that encode genes on

both strands (Nguyen and Haenni 2003). In the fourteen completely sequenced dicistroviruse genomes (Table 4.1), I identified 240 overlapping ORFs of length equal or greater than 60 codons on the negative strand. Of the 240 ORFs, 113 were found in concordant genomic locations in two or more genomes. The concordant overlapping ORFs were assigned into 29 clusters (Table 4.4). There are 9, 1, and 19 clusters in phase 0, 1, and 2, respectively. The cluster size ranges from 2 to 9. In only two clusters, 5 and 10, both in phase 2, I detected a weak signature of selection. However, this signature seems to be a false positive, which was driven by the unique structure of opposite-strand phase-2 overlap (as described in Chapter Three). In this structure, codon positions one and two of one gene match codon positions two and one of the overlapping gene. This structure leads to a situation where most changes are either synonymous or nonsynonymous in both overlapping genes and occasionally, to false signal of purifying selection on the overlapping ORF. I therefore conclude that dicistroviruses most probably do not encode proteins on the negative strand.

Table 4.4: Clusters of orthologous overlapping ORFs on the negative strands of dicistrovirid genomes.

Cluster	Virus	Phase	Start	End	Length
1	IAPV	1	913	1131	219
1	ABPV	1	975	1175	201
1	KBV	1	1006	1191	186
2	IAPV	2	902	1087	186
2	ABPV	2	1042	1593	552
2	IAPV	2	1334	1630	297
2	KBV	2	1454	1690	237
3	IAPV	2	1634	1906	273
3	KBV	2	1724	1927	204
3	ABPV	2	1747	2031	285
4	IAPV	2	2507	2755	249
4	ABPV	2	2509	2697	189
4	SINV-1	2	276	581	306
4	HiPV	2	2940	3125	186
4	RhPV	2	2445	2795	351
4	TrV	2	2723	2962	240
5	IAPV	2	4070	4510	441
5	KBV	2	4082	4375	294
5	ABPV	2	3982	4275	294
5	SINV-1	2	1902	2186	285
6	IAPV	2	4694	4948	255
6	KBV	2	4706	4990	285
7	IAPV	2	5696	5941	246

7	SINV-1	2	3570	3761	192
7	CrPV	2	5376	5558	183
8	IAPV	0	1224	1421	198
8	ABPV	0	1127	1318	192
8	KBV	0	1107	1349	243
9	IAPV	0	5859	6086	228
9	RhPV	0	5983	6228	246
9	ALPV	0	6030	6305	276
9	ABPV	0	5957	6307	351
9	IAPV	0	6111	6395	285
9	KBV	0	6102	6305	204
9	SINV-1	0	3811	3996	186
9	HoCV-1	0	5325	5537	213
10	IAPV	2	7601	7792	192
10	KBV	2	7436	7783	348
10	ABPV	2	7426	7761	336
10	SINV-1	2	5805	6104	300
10	TSV	2	7729	7998	270
11	IAPV	2	8471	8677	207
11	ABPV	2	8383	8571	189
12	IAPV	0	7614	8099	486
12	KBV	0	7890	8102	213
12	SINV-1	0	5776	6141	366
13	IAPV	0	8103	8402	300
13	BQCV	0	7094	7405	312
13	TrV	0	7366	7575	210
14	IAPV	0	8406	8600	195

14	KBV	0	8352	8594	243
15	KBV	2	3239	3571	333
15	ABPV	2	3289	3477	189
16	KBV	2	5168	5434	267
16	ABPV	2	5068	5250	183
16	SINV-1	2	3090	3485	396
16	PSIV	2	4716	5006	291
16	TSV	2	5537	5764	228
16	TrV	2	4646	4864	219
16	HiPV	2	5130	5327	198
17	KBV	2	6821	7090	270
17	ABPV	2	6727	6996	270
18	KBV	2	8609	8896	288
18	ABPV	2	8575	8853	279
18	SINV-1	2	6897	7376	480
19	ABPV	2	2035	2340	306
19	SINV-1	2	28	269	242
19	DCV	2	2043	2231	189
19	RhPV	2	2073	2441	369
19	CrPV	2	2004	2258	255
19	ALPV	2	1571	2278	708
19	BQCV	2	1926	2123	198
19	RhPV	2	1617	2051	435
20	ABPV	2	5872	6141	270
20	ALPV	2	6218	6424	207
20	CrPV	2	5562	5786	225
20	TSV	2	6254	6436	183

20	PSIV	2	5643	5825	183
21	ABPV	0	5303	5539	237
21	BQCV	0	4609	4923	315
21	SINV-1	0	3034	3282	249
21	TrV	0	4938	5150	213
21	HoCV-1	0	4596	4877	282
21	ALPV	0	5286	5582	297
21	RhPV	0	5392	5583	192
21	TrV	0	4575	4781	207
21	PSIV	0	5086	5280	195
22	SINV-1	0	169	387	219
22	DCV	0	2152	2397	246
22	CrPV	0	2293	2505	213
22	RhPV	0	2332	2586	255
22	KBV	0	2550	2756	207
22	HiPV	0	2806	3057	252
22	ALPV	0	2142	2456	315
23	CrPV	0	4063	4293	231
23	RhPV	0	4612	4836	225
23	ALPV	0	4692	4877	186
24	CrPV	2	6402	6641	240
24	BQCV	2	5995	6186	192
24	PSIV	2	6291	6578	288
25	DCV	2	7331	7531	201
25	ALPV	2	7806	8033	228
26	DCV	2	7718	7957	240
26	RhPV	2	8456	8752	297

27	DCV	0	6405	6611	207
27	BQCV	0	5834	6226	393
27	HoCV-1	0	6176	6388	213
28	PSIV	2	7674	7910	237
28	BQCV	2	7366	7548	183
28	TrV	2	7701	7940	240
28	HiPV	2	8179	8364	186
29	BQCV	2	4107	4376	270
29	HiPV	2	4746	4988	243

In this chapter, I provided evolutionary evidence (purifying selection) for the existence of a functional overlapping gene, *pog*, in the genomes of IAPV, ABPV, KBV, and SINV-1. To my knowledge, this hypothetical gene, whose coding region overlaps the structural polyprotein, has not been described in the literature before.

Chapter Five: Detection of functional overlapping genes using population-level data

Abstract

Current methods that utilize the signature of purifying selection to detect functional overlapping genes are limited to the analysis of sequences from divergent taxa. Here, I present a method for the detection of selection signatures on overlapping reading frames by using population-level data. I tested the method on both functional and spurious overlapping genes. Finally, I used the method to test whether an overlapping reading frame on the negative strand of segment 8 in influenza A is under selection.

Introduction

It is fairly common for at least one of the five possible overlapping reading frames of any gene to contain an open reading frame (ORF) of a length that may be suitable to encode a protein. Unfortunately, it is extremely difficult to ascertain whether an intact overlapping ORF is functional or spurious. The main reason for this difficulty is that the sequence of an overlapping gene is, by definition, constrained by the functional and structural requirements of another gene. As a result, many putative overlapping genes have been identified as functional (Chung et al. 2008; Firth 2008; Firth and Atkins 2008b; Firth and Atkins 2008a; Firth and Atkins 2009), while at the same time numerous annotated overlapping genes have been deemed upon reexamination to be spurious (Silke 1997; Palleja, Harrington, and Bork 2008; Williams, Wolfe, and Fares 2009). The

common way to detect functional overlapping genes is to identify a signature of purifying selection, which is interpreted as a sign of functionality (Firth and Brown 2005; Firth and Brown 2006; Sabath, Landan, and Graur 2008). In Chapters Three and Four, I showed how my method for the estimation of selection intensity (Chapter Two; Sabath, Landan, and Graur 2008) can be utilized to distinguish between spurious and functional overlapping genes. However, this method, as well as other methods in the literature, are inaccurate when the compared sequences show high sequence similarity (Firth and Brown 2006; Sabath, Landan, and Graur 2008, Chapter Two). In addition, although pairwise methods can be applied to multiple sequences in a phylogenetic tree (Firth and Brown 2006), the estimation is computationally impractical if hundreds of sequences, such as populations of clinically important viruses and bacteria, need to be considered.

One interesting case is that of influenza A, where viral sequences belonging to the same subtype are highly similar to one another. An overlapping ORF in the negative strand of segment 8 of influenza A viruses (Figure 5.1) was noted when this segment was first sequenced (Baez et al. 1980). The ORF is found intact in several human influenza A viruses, but is absent from non-human influenza A viruses, such as avian viruses, and is also absent from influenza B and C viruses. Recently, it was suggested that this overlapping ORF codes for a functional gene (Zhirnov et al. 2007; Clifford, Twigg, and Upton *in press*). Two main indications that this hypothetical gene, called *NEG8*, may be functional were given: (1) The ORF has been conserved in human influenza A viruses for almost a century (Zhirnov et al. 2007; Clifford, Twigg, and Upton *in press*); and (2)

An epitope (a short peptide) encoded by this ORF was reported to induce an immune system response through cytotoxic T cells isolated from mice infected with this virus (Zhong et al. 2003; Clifford, Twigg, and Upton *in press*)

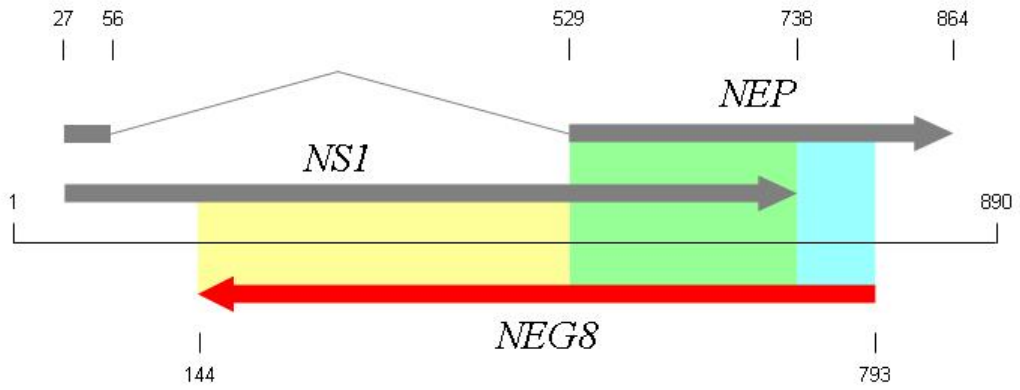

Figure 5.1: Schematic representation of segment 8 in human influenza A viruses. The *NS1-NEG8* overlap is marked in yellow. The *NS1-NEG8-NEP* triple overlap is marked in green. The short *NEG8-NEP* overlap is marked in blue.

Functional gene identification is the *sine qua non* of functional genetics. In viruses, the impact of identifying a *bona fide* novel gene is even greater because their gene number is usually very small. In fact, the eleventh gene in the influenza A genome, *PB1-F2* (which overlaps *PB1*), was discovered 20 years after the annotation of its genome (Chen et al. 2001), thereby increasing the proteome by 10%. Here, I present a method for the detection of purifying selection on hypothetical overlapping reading frames using population-level data. I test the method on both known and spurious overlapping genes.

Finally, I used the method to test whether an overlapping reading frame on the negative strand of segment 8 in influenza A is under selection.

Methods

To detect the signature of purifying selection acting on a hypothetical gene, I employed the principle that nonsynonymous mutations are generally more deleterious than synonymous mutations. If a hypothetical gene is under selection, a mutation, which is nonsynonymous in both genes, is expected to be more deleterious than one that is nonsynonymous in one gene and synonymous in the other.

Mutations are classified into transitions and transversions, which usually occur at different rates. Mutations that become fixed in the population are called substitutions. Substitutions in a known gene (k) that overlaps a hypothetical gene (h) can be classified into four categories: nonsynonymous in both genes (N_kN_h), nonsynonymous in the known gene and synonymous in the hypothetical gene (N_kS_h), synonymous in the known gene and nonsynonymous in the hypothetical gene (S_kN_h), and synonymous in both genes (S_kS_h). Taken together, I was able to define eight categories of substitutions for each pair of overlapping sequences (Table 5.1).

Table 5.1: Notation of the test variables

	Transitions				Transversions			
Categories	N_kN_h	N_kS_h	S_kN_h	S_kS_h	N_kN_h	N_kS_h	S_kN_h	S_kS_h
Possible Substitutions	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Observed Substitutions	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
Expected	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8

Throughout this chapter, I will use the term "category pair" to denote a pair of substitutional categories that differ only in the hypothetical gene (i.e., N_kN_h versus N_kS_h, and S_kN_h versus S_kS_h). The four category pairs are set apart in shaded cells in Table 5.1. For example, being nonsynonymous vs. synonymous in the hypothetical gene is the only difference between a transitional mutation in the N_kN_h category vs. a transitional mutation in the N_kS_h category. In the absence of selection on the hypothetical gene, the rates of the two substitutional categories in a pair should be equal to each other. If, on the other hand, the hypothetical gene is functional and under purifying selection, the rate of substitution in the N_kN_h category should be lower than that in N_kS_h, because a nonsynonymous change in both the known and the hypothetical genes will affect two gene products rather than one. Similarly, the rate of change in the S_kN_h category should be lower than that in S_kS_h.

For example, Figure 5.2 illustrates three possible substitutions at site 4 in the phase-2 opposite-strand overlapping sequence. If the hypothetical gene is under selection, the change A/T→C/G (S_kN_h category) is expected to be more deleterious than A/T→T/A (S_kS_h category), which does not change the amino acid of the hypothetical gene. I note that there is no assumption about the intensity of selection on the known gene and, hence, the method can be used even when the known gene is in fact under no selection (evolving neutrally).

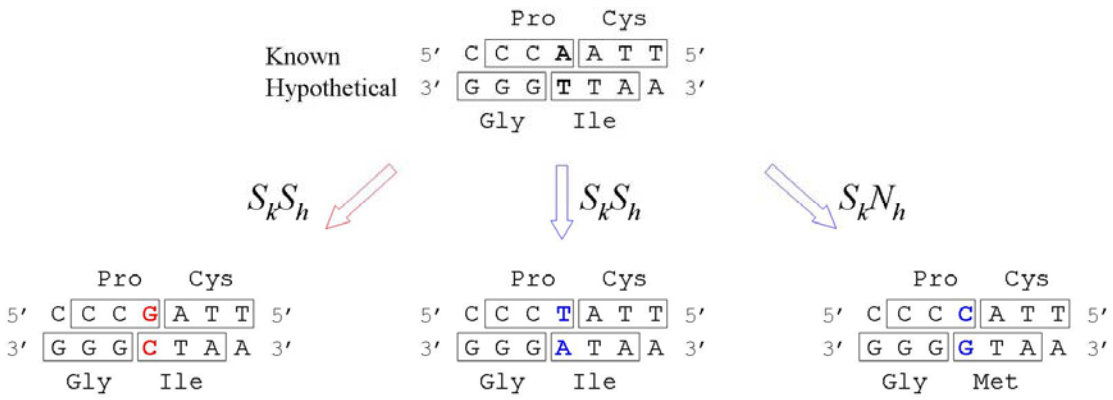

Figure 5.2: Three possible substitutions of site 4 (marked in bold) in the phase-2 opposite-strand overlapping sequence between a known gene and a hypothetical gene. The transition is marked in red and the transversions are marked in blue. The substitutional category of each change is noted.

Given a multiple alignment of closely related DNA sequences, the method includes four steps:

(1) Construction of an unrooted phylogenetic tree.

(2) Reconstruction of ancestral sequences.

(3) Classification of the changes along the tree into the eight substitutional categories.

(4) Testing for signature of purifying selection through comparisons between category pairs that differ at the hypothetical gene only.

I used PAUP (Swofford 2003) to construct a neighbor-joining tree (Saitou and Nei 1987) of each data set and assigned the ancestral character states of the internal nodes using the parsimony criteria (Fitch 1971). Using the reconstructed sequences, I counted the number of unique observed substitutions (*O*) in each category along all the branches. I used the unique number of substitutions rather than the total number of substitutions to minimize the possible biases from non-uniform sampling (e.g., in industrial countries where more isolates are collected) and from highly constrained variable sites, in which only a few character states are permissible (Delport, Scheffler, and Seoighe 2008). For any given sequence of length *n*, there are *3n* possible substitutions that can be classified into these eight categories (Figure 5.2). For any given set, I calculated the number of possible substitutions (*P*) in each category as the average across the sequences in all nodes of the tree.

I use the ratio $\frac{O_i}{P_i}$ as a measure of the rate of substitutions in category *i*. If the hypothetical gene is not under selection, I expect no difference between $\frac{O_i}{P_i}$ and $\frac{O_j}{P_j}$,

where *i* and *j* are two categories that differ only at the hypothetical gene:

$< i, j > \in \{< 1,2 >, < 3,4 >, < 5,6 >, < 7,8 >\}$.

The null hypothesis of no selection on the hypothetical gene is defined as:

(1) $\frac{O_i}{P_i} = \frac{O_j}{P_j} = \frac{(O_i + O_j)}{(P_i + P_j)}$.

Under this null hypothesis, I estimated the expected values of O_i and O_j to be

(2) $E_i = P_i \frac{(O_i + O_j)}{(P_i + P_j)}$ and $E_j = P_j \frac{(O_i + O_j)}{(P_i + P_j)}$.

I, then, constructed a contingency table for each category pair

(3) $\begin{pmatrix} O_i & O_j \\ E_i & E_j \end{pmatrix}, < i, j > \in \{< 1,2 >, < 3,4 >, < 5,6 >, < 7,8 >\}$.

This contingency table is used to test the null hypothesis. For example, O_1 and O_2, which are the observed number of substitutions in the transitional N_kN_h and N_kS_h categories (Table 5.1), differ only by being nonsynonymous or synonymous in the hypothetical gene. E_1 and E_2, which are the expected values of O_1 and O_2, are estimated based on the null hypothesis in which the rate of substitutions in the two categories is equal. If the hypothetical gene is subjected to selection, any change in the N_kN_h category would affect both genes and O_1 is expected to be lower than E_1, whereas O_2 is expected to be higher than E_2.

I used a one-tailed Fisher's exact test (Fisher 1925) to determine significance of the negative association, in which the observations tend to lie in the lower left and upper

right of the table. For example, in the contingency table $\begin{pmatrix} O_1 & O_2 \\ E_1 & E_2 \end{pmatrix}$, the alternative, in which category 1 is under stronger purifying selection, requires that O_1 is lower than E_1, and that O_2 is higher than E_2.

Because the test requires exact numbers, the expected values were rounded. Finally, I combined the four *p*-values into a single test statistic using Fisher's method (Fisher 1925).

In the case of overlap types other than phase-2 opposite-strand overlap (the overlap type of *NEG8*), the number of possible S_kS_h substitutions is very small (0 – 3, Table 5.2). This makes the S_kS_h categories, and consequently the N_kS_h – S_kS_h pair, uninformative. Therefore, I focused on the two category pairs of N_kN_h and N_kS_h (shaded in Table 5.1) in the test.

Sequence data

The data were taken from the NCBI Influenza Virus Resource (Bao et al. 2008). The data consists of (1) influenza A, H3N2 and H1N1 subtypes, in which the *NEG8* ORF is intact, (2) influenza A, H5N1 subtype and influenza B, in which the *NEG8* ORF is disrupted, and (3) four other known overlapping genes (*PB1 – PB1-F2* from influenza A, H3N2 subtype; *PB1 – PB1-F2* and *NS1 – NEP* from influenza A, H5N1 subtype; and *NA - NB* from influenza B). For each set, I obtained the multiple alignments of all full-length

sequences excluding sequences with insertions and/or deletions. Because ancestral sequence reconstruction in inaccurate for diverged sequences (Zhang and Nei 1997), I also excluded sequences from early isolates (before 1990) that form a distant clade (e.g., the H1N1 subtype contains 17 sequences from 1918–1945 and only two sequences between 1950 and 1990). For the sets of *NEG8*, I analyzed the region of *NEG8* that overlaps with *NS1* solely (382 bases) and excluded the regions of triple overlap and the short region of *NEP-NEG8* overlap. For all data sets, the frequencies of the possible and the observed number of substitutions in each category are listed in Table 5.2.

Finally, I used the complete genomes of 768 RNA non-ambisense (viruses that utilize only one strand to code for proteins) viruses to evaluate the influence of genome composition on the probability of having an overlapping ORF. Genomes were obtained from NCBI. Stop codon frequencies in the five possible reading frames (on the same strand in phase 1 and 2, and on the opposite strand in phase 0, 1, and 2) were calculated from the coding sequences of each genome.

Table 5.2: Possible and observed number of substitutions in each category

					Nonsynonymous substitutions in gene 2		Synonymous substitutions in gene 2	
	Gene 1	Gene 2			*NN*	*NS*	*SN*	*SS*
Influenza A: H3N2	*NEG8*	*NS1*	Ts	*P*	237.8	16.9	22.9	104.4
				O	73	10	16	68
			Tv	*P*	554.9	78.6	65.8	64.8
				O	30	10	16	7
Influenza A: H1N1	*NEG8*	*NS1*	Ts	*P*	237.8	18.0	18.4	107.8
				O	53	4	12	61
			Tv	*P*	547.1	81.9	70.3	64.7
				O	25	4	13	7
Influenza A: H5N1	*NEG8*	*NS1*	Ts	*P*	237.2	19.8	17.0	108.1
				O	111	12	14	93
			Tv	*P*	539.7	84.4	71.0	69.0
				O	61	13	13	21
Influenza B	*NEG8*	*NS1*	Ts	*P*	346.5	32.3	21.3	167.8
				O	104	11	12	114
			Tv	*P*	829.9	141.1	92.4	72.7
				O	66	12	18	10
Influenza A: H3N2	*PB1-F2*	*PB1*	Ts	*P*	97.9	83.1	86.0	3.0
				O	21	13	76	3
			Tv	*P*	377.0	73.0	89.0	1.0
				O	16	4	19	1
Influenza A: H5N1	*PB1-F2*	*PB1*	Ts	*P*	97.9	83.1	86.0	3.0

				O	15	15	81	4
			Tv	*P*	382.4	70.5	86.0	1.0
				O	20	4	23	1
Influenza A: H5N1	*NS1*	*NEP*	Ts	*P*	62.6	53.1	55.2	0.1
				O	29	25	36	0
			Tv	*P*	236.8	57.9	44.2	3.0
				O	22	7	9	0
Influenza B	*NB*	*NA*	Ts	*P*	92.6	93.3	103.2	2.0
				O	29	33	62	1
			Tv	*P*	358.0	99.8	121.4	2.9
				O	25	7	13	0

Results

As a control, I applied the method on all sets twice (reciprocally), to test for selection on each gene while using its overlapping open reading frame as the known gene (Table 5.3). Using a spurious gene as the known gene in the test (see Methods) allows the evaluation of the method in more cases.

NEG8 – NS1 sets

I found significant signatures of selection in three out of the four known *NS1* genes (the p-value of the fourth one is relatively low, 0.086), demonstrating the ability of the method to detect selection in known functional genes. I used the two sets in which no

NEG8 ORF exists (H5N1 and influenza B), to verify that the method does not yield false positive inferences. In both cases, no signature of selection was identified on the *NEG8* ORF as expected. Finally, I applied the method to test for selection on the hypothetical *NEG8* ORF in the H1N1 and H3N2 sets. I did not find a significant signature of selection on the *NEG8* ORF in either case.

Known same-strand overlapping genes

I used four sets of known overlapping genes in influenza to test the performance of the method in same-strand overlapping genes (Table 5.3). The overlaps of these genes result in very small number of possible substitutions, which are synonymous in both genes (Table 5.2). Therefore, the test is applied only to two category pairs rather than four (see Methods section). In three sets, I identified significant signatures of selection on one gene while no selection was identified on the other. For the forth set, there were no significant signatures detected on any of the two genes.

Table 5.3: sets of sequences in the study

	Virus	Number of sequences	Gene 1	*p*	Gene 2	*p*
Hypothetical *NEG8*	Influenza A: H3N2	410	*NEG8*	0.151	*NS1*	*
gene	Influenza A: H1N1	217	*NEG8*	0.667	*NS1*	**
No *NEG8* gene	Influenza A: H5N1	581	*NEG8*	0.359	*NS1*	0.086
	Influenza B	229	*NEG8*	0.604	*NS1*	*
Known same-	Influenza A: H3N2	999	*PB1-F2*	0.446	*PB1*	***
strand overlapping	Influenza A: H5N1	522	*PB1-F2*	0.690	*PB1*	***
genes	Influenza A: H5N1	581	*NS1*	0.647	*NEP*	0.151
	Influenza B	165	*NB*	0.617	*NA*	*

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

Discussion

I presented a new method for the detection of functional overlapping genes utilizing the signature of selection for population-level data. The method detects selection-signature based on the principle that nonsynonymous mutations are generally more deleterious than synonymous mutations. As far as overlapping genes are concerned, this principle translated into the following expectation: a mutation that is nonsynonymous in both genes is expected to be more deleterious than a mutation that is nonsynonymous in one gene and synonymous in the other.

Variation in selection pressures among sites may affect this method's performance. For example, a mutation of the N_kS_h category at a constrained site of the known protein may be more deleterious than a mutation of the N_kN_h category at less constrained sites of both genes. As a control for site variation, I used data sets of orthologous sequences that share constrained sites and in which the hypothetical *NEG8* ORF is disrupted. In future studies, it may be beneficial to incorporate information of the known protein's constrained sites into the model.

Difference in the intensities of selection acting on the two overlapping genes may also affect the performance of the method, especially when the hypothetical gene is under weaker purifying selection than the known overlapping gene. In an overlapping gene pair, the newer gene is expected to be under weaker purifying selection (Liang and

Landweber 2006), because it has evolved for less time than its overlapping genes as well as under the constraints of its overlapping gene. The hypothetical overlapping gene would usually be the newer gene. Therefore, detection of new overlapping genes by signature of purifying selection is difficult. Indeed, *PB1-F2*, the novel human influenza A gene (Chen et al. 2001) was not detected by the method (Table 5.3). This gene was shown to be under selective pressure that is weaker by an order of magnitude than that on the older overlapping gene, *PB1* (Sabath, Landan, and Graur 2008).

Given these considerations, it is difficult to determine if the lack of selection signature in the *NEG8* ORF is due to the intact reading frame being spurious, or the inability of the method to detect the signal because the gene is too new. There are two additional factors that may have contributed to the conservation of the *NEG8* ORF even in the absence of selection. First, low frequency of stop codons in that reading frame, which result from the codon usage and amino-acid composition of the genome, can lead to spurious overlapping ORFs (Silke 1997; Sabath, Graur, and Landan 2008). Indeed, the specific overlap type between *NEG8* and *NS1* (opposite strand phase 2), which encompasses ~90% of the *NEG8* ORF was found to have the lowest frequency of stop codons (Figure 5.3). Moreover, influenza genomes have a below-average frequency of stop codons in this phase (Figure 5.3b, black dots) increasing the probability of spurious ORFs. Second, the triple overlap between *NEG8* and both *NS1* and *NEP* may also increase the conservation of *NEG8* because any change in this region is likely to be nonsynonymous in either *NS1* or *NEP* or both.

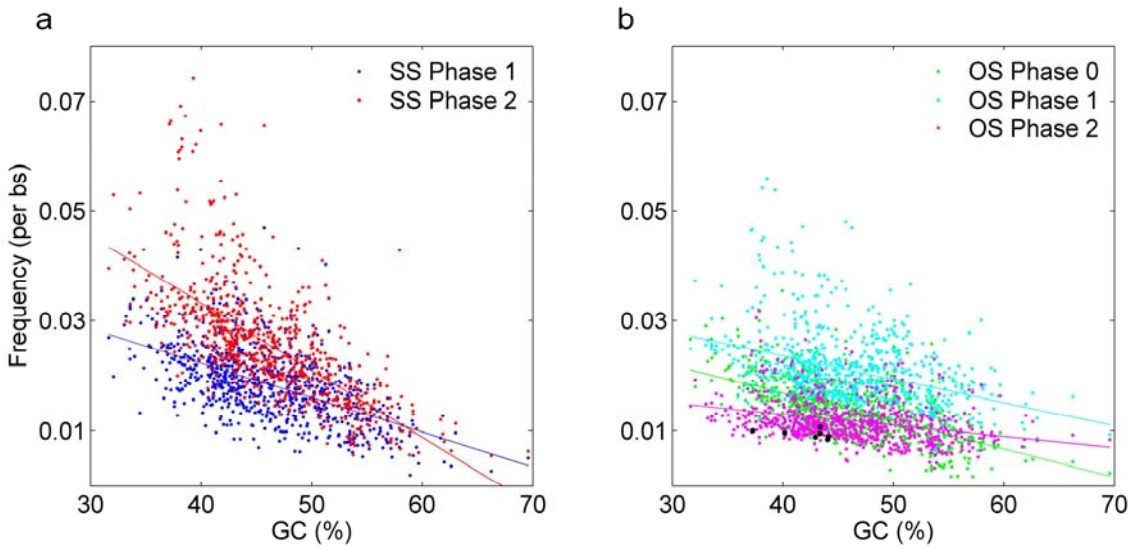

Figure 5.3: Frequencies of stop codon of 768 RNA viruses in the five possible reading frames plotted against genomic GC content. (a) Stop codon frequencies on the same strand (SS) in phase 1 (blue) and phase 2 (red). (b) Stop codon frequencies on the opposite strand (OS) in phase 0 (green), phase 1 (cyan), and phase 2 (magenta). Stop codons frequencies on the opposite strand in phase 2 (*NEG8* phase) of influenza genomes are marked in black.

The method presented in this chapter belongs to a group of approximate methods, in which sites are classified by degeneracy classes. Approximate methods are useful for analyses of large data sets, but are less accurate than maximum-likelihood methods

(Yang and Nielsen 2000). Hence, more powerful methods may be developed within the maximum-likelihood framework. An additional parameter that could be incorporated in future methods is the time of origin of each substitution, which could be estimated by using the sampling dates as calibration (Drummond and Rambaut 2007). Deleterious substitutions are expected to be more prevalent among new substitutions (Pybus et al. 2007) because they had lower chance to be eliminated from the population by selection. Therefore, it would be beneficial to account for the substitution's age when selection is estimated at the population level.

All in all, since none of the existing methods in the literature is applicable to population-level data, I believe that my method is a first step in the right direction.

Chapter Six: Phase bias in same-strand overlapping genes

Abstract

Same-strand overlapping genes may occur in frameshifts of one (phase 1) or two nucleotides (phase 2). In previous studies of bacterial genomes, long phase-1 overlaps were found to be more numerous than long phase-2 overlaps. This bias was explained by either genomic location or an unspecified selection advantage. A Model that focused on the ability of the two genes to evolve independently did not predict this phase bias. Same-strand overlapping genes may arise through either a mutation in the termination codon of the upstream gene or a mutation at the initiation codon of the downstream gene. I hypothesized that given these two scenarios, the frequencies of initiation and termination codons in the two phases may determine the number for overlapping genes. I examined the frequencies of initiation- and termination-codons in the two phases, and found that termination codons do not significantly differ between the two phases, whereas initiation codons are more abundant in phase 1. I found that the primary factors explaining the phase inequality are the frequencies of amino acids whose codons may combine to form start codons in the two phases. I show that the frequencies of start codons in each of the two phases, and, hence, the potential for the creation of overlapping genes, are determined by the abundance of amino acids in proteins and by species-specific codon usage, leading to a correlation between long phase-1 overlaps and genomic GC content. My model explains the phase bias in same-strand overlapping genes by compositional factors without invoking selection. Therefore, it can be used as a

null model of neutral evolution to test selection hypotheses concerning the evolution of overlapping genes.

Introduction

In bacteria, overlaps on the same strand are by far the most abundant (Fukuda, Washio, and Tomita 1999; Johnson and Chisholm 2004), most likely because, on average, 70% of the genes in bacterial genomes, are located on one strand (Fukuda, Nakayama, and Tomita 2003). Same-strand overlaps occur in frameshifts of one nucleotide (phase 1) or two nucleotides (phase 2, Figure 1.3). Overlaps in the same frame (phase 0) are rare (Johnson and Chisholm 2004), and since the reading frame is unaffected, they may be thought of as genes with alternative initiation or termination sites rather than overlapping genes. Phase-0 overlaps are not dealt with here. Several studies have shown that there are significant differences between the frequencies of phase-1 and phase-2 overlapping genes (Johnson and Chisholm 2004; Cock and Whitworth 2007; Lillo and Krakauer 2007) (Figure 6.1). Overlapping-gene pairs, in which the overlap sequence is of length one to five bases (short overlaps), are abundant in phase 2, but rare in phase 1. This difference is dictated by the sequence of termination codons of the upstream gene (Cock and Whitworth 2007). Since none of the stop codons (TGA, TAG, and TAA) ends with AT, GT, or TT (needed to create an initiation codon ATG, GTG or TTG in phase-1 two-nucleotide overlap) or starts with G (needed to create an initiation codon in phase-1 five-nucleotide overlap), short phase-1 overlaps can only use alternative initiation codons. In

contrast, as far as long overlaps (seven nucleotides or longer) are concerned, phase-1overlapping gene pairs are more frequent than those of phase 2 (Johnson and Chisholm 2004; Cock and Whitworth 2007). Cock and Whitworth (2007) suggested that the phase bias in long overlaps is due to some unspecified selective advantage of phase-1 over phase-2 overlapping genes. They also hypothesized that since the bias was found to be universal and independent of gene function, it might be a property of the gene location.

Krakauer (2000) introduced a model in which the frequencies of overlapping genes in different phases are determined by their degree by which the two overlapping proteins can evolve independently, which is defined by the probability for changes, which are nonsynonymous in one gene and synonymous in the overlapping gene (Figure 1.7). That model assumes an adaptive advantage for overlapping genes in evolvable phases (Krakauer 2000). For example, in the case of opposite-strand overlaps, phase 1 in which the second codon position of one gene corresponds to the third codon position of the second gene (and vice versa), maximizes the freedom of each gene to evolve independently (Krakauer 2000) (Figure 1.7). In support of this model, Rogozin et al. (2002) found that among opposite-strand overlaps in bacteria, the most evolvable overlap phase (phase 1) was the most abundant. In contrast, Kingsford et al. (2007) explained this phase distribution in opposite-strand overlapping genes by the frequency of reverse-complementary stop codons in coding sequences. For same-strand overlaps, phase-1 and phase-2 overlaps have equal protein evolvability and are predicted by that model, to occur in equal frequencies (Krakauer 2000).

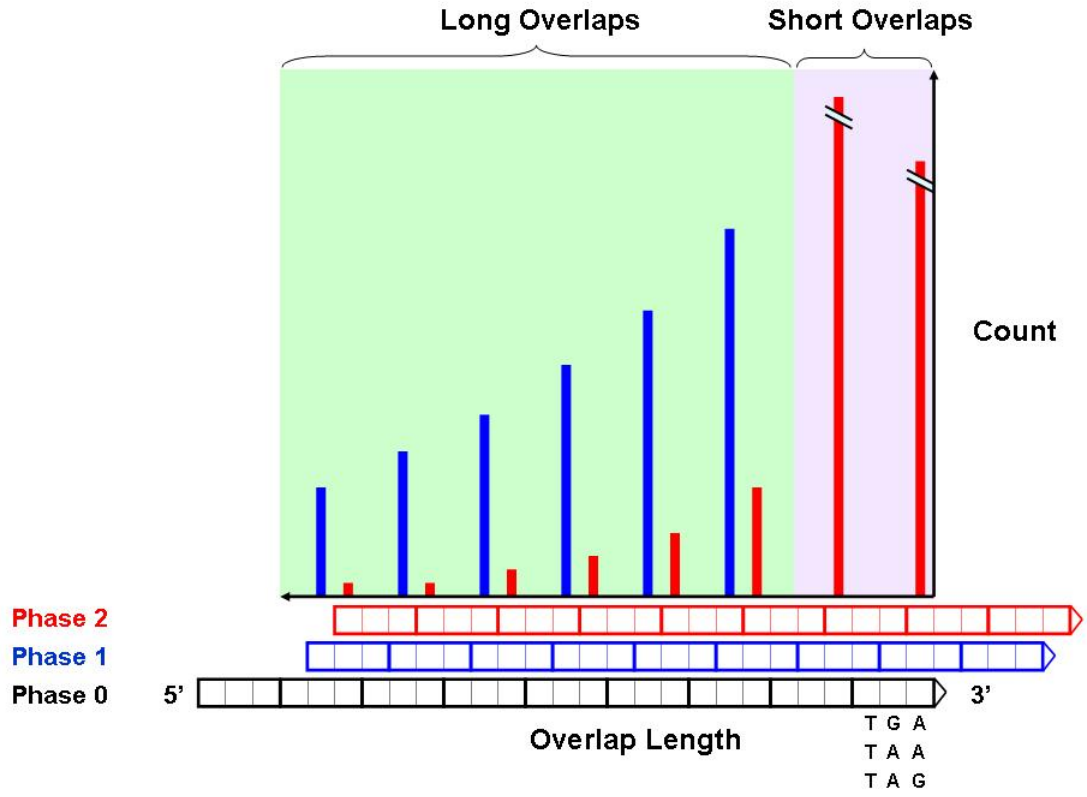

Figure 6.1: Illustration of the phase-distribution for same-strand overlapping genes in bacterial genomes as observed by previous studies. Given an upstream gene (phase 0), genes can overlap in phase 1 (dark blue) or phase 2 (red). The frequency of overlaps (Y axis) is plotted against overlap length (X axis). Short overlaps, in which the start codon overlaps the stop codon of the upstream gene, are marked in purple. Long overlaps are marked in green.

Previous studies (Fukuda, Nakayama, and Tomita 2003; Johnson and Chisholm 2004) have found that the number of overlapping genes in bacterial genomes is positively correlated with the number of genes, implying that gene overlap may be mainly the

result of accidental or random "trespassing" of one gene into another. There can be two scenarios for the creation of same-strand overlapping genes from pre-existing neighboring genes (Figure 6.2a): (1) a mutation in the termination codon of the upstream gene, resulting in an extension of the gene downstream to the first in-frame termination codon and (Figure 6.2b) (2) a mutation in the initiation codon of the downstream gene, resulting in an extension of the gene upstream to the first in-frame functional initiation codon (Fukuda, Nakayama, and Tomita 2003) (Figure 6.2c). As in point mutations, where the effect of nonsynonymous mutation is expected to be stronger than that of synonymous ones, the impact of mutations that cause extension is expected to vary according to the length of the extension. Since most mutations are deleterious, long extensions of genes are expected to be under stronger purifying selection than short ones (Kingsford, Delcher, and Salzberg 2007) and the frequency of initiation and termination codons in a certain phase is an upper-limit constraint to the possible number of overlapping genes in that phase.

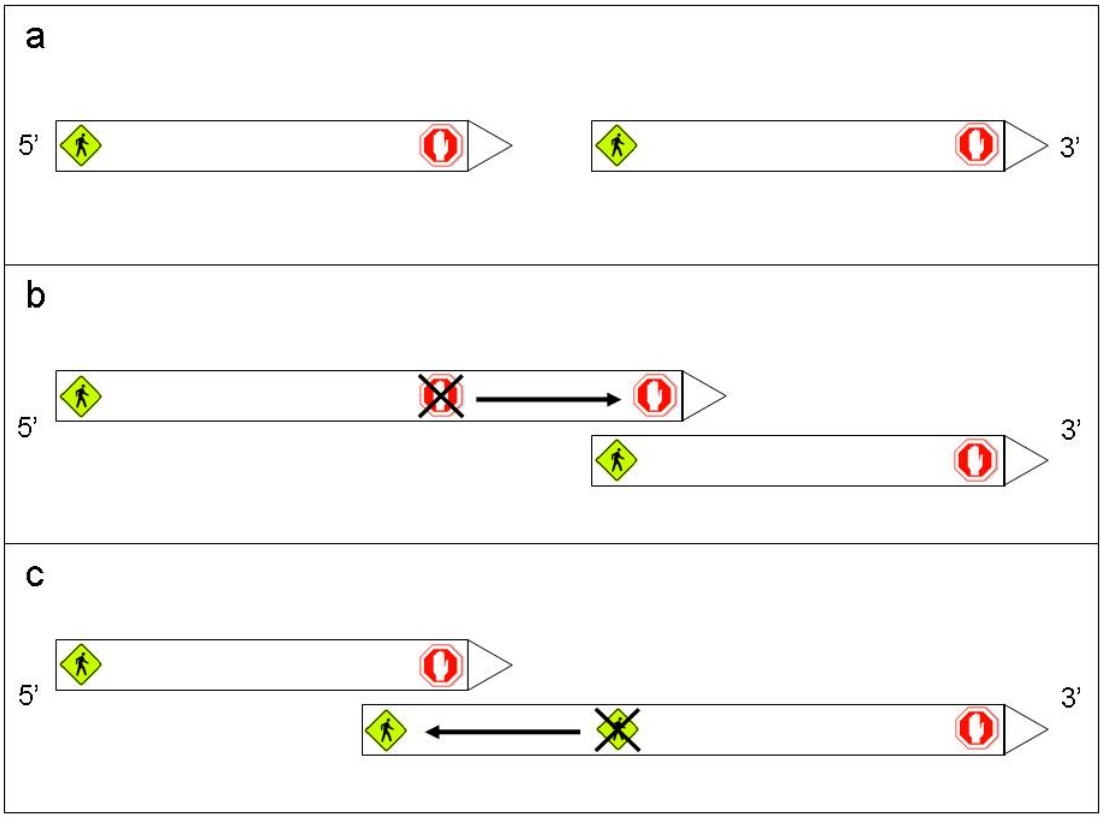

Figure 6.2: An adjacent gene pair on the same strand (a) can evolve into an overlapping gene pair through a mutation in the termination codon of the upstream gene (b), or a mutation in the initiation codon of the downstream gene (c).

Here, I tested the influence of initiation- and termination-codon frequencies as well as genomic GC-content on the number of overlapping genes in the two phases.

Methods

Data of overlapping genes from 167 bacterial genomes that employ the universal genetic code were acquired from the BPhyOG overlapping-genes database (Luo et al. 2007). Same-strand overlapping genes in each genome were classified according to phase and the length of the intersecting segment. I defined overlap frequency as the number of same-strand overlapping genes divided by the number of same-strand neighboring gene pairs (i.e., adjacent genes, which are located on the same strand and in between them there are no genes on the opposite strand) in the genome. In the analysis, I explicitly ignored recombination and therefore I used the number of same-strand neighboring gene pairs, rather than the number of genes, because a neighboring gene pair located on opposite strands cannot become overlapping on the same strand as a result of point mutation. Short overlaps (two and five bases in phase 1 and one and four bases in phase 2) were dealt separately from long overlaps of seven bases or longer.

The coding sequences of the studied genomes were downloaded from NCBI. Codon and amino-acid frequencies, as well as initiation and termination codon frequencies in phase 1 and phase 2, were calculated from the coding sequences of each genome. I denote the frequency of a codon or a group of codons with a superscript for the codon's phase and a subscript for the codon. For example, f^1_{ATG} denotes the frequency of ATG in phase 1 and f^0_{NAT} denotes the frequencies of codons in phase 0 that end in AT, where N denotes any of the four nucleotides. The expected frequencies of each start and stop codons are

calculated as the products of the frequencies of the codons that combine them, i.e., $f^0_{NAT} \times f^0_{GNN}$ and $f^0_{NNA} \times f^0_{TGN}$ for ATG in phase 1 and phase 2, respectively. If the codons frequencies in phase 1 and phase 2 are primarily determined by the frequencies of the codons in phase 0 that combine them, the expected frequencies would match the observed frequencies.

Results

I identified 71,210 same-strand overlapping gene pairs (Table 6.1). Short overlaps (of length two or five bases) are rare in phase 1. In this sample, I found only 18 phase-1 short overlaps (0.08%, Table 6.1). In contrast, the majority of phase-2 overlaps are of length one or four bases (20% and 65%, respectively).

Table 6.1: Number of same-strand overlapping genes.

	Short overlaps (1-5 bases)	Long overlaps (7 bases or more)	Total
Phase 1	18	21,550	21,568
Phase 2	42,177	7,465	49,642
Total	42,195	29,015	71,210

The frequency of long phase-1 overlaps exceeds that of long phase-2 overlaps by a factor of almost 3 (Table 6.1, Figure 6.3, paired Student t-test, $p < 0.001$). The frequency of long phase-1 overlaps is negatively correlated with genomic GC content (Figure 6.3 *r*

= –0.39, $p < 0.001$). In contrast, the correlation between the frequency of long phase-2 overlaps and GC content is not significant ($p = 0.4$). The frequencies of start and stop codons in phase 1 and phase 2 in the coding regions of the genomes are presented in Figure 6.4. Pooling together phase 1 and phase 2, the frequency of stop codons (average of 13.16%) is significantly higher than that of start codons (average of 9.36%, paired Student t-test, $p < 0.001$).

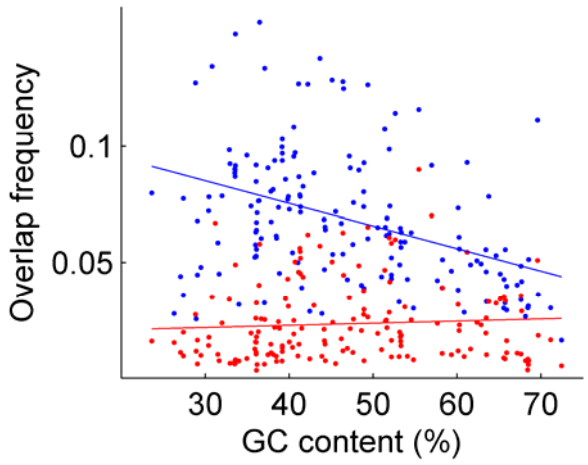

Figure 6.3: Frequency of overlapping genes in 167 bacterial genomes plotted against genomic GC content. Long phase-1 overlaps are marked in blue. Long phase-2 overlaps are marked in red.

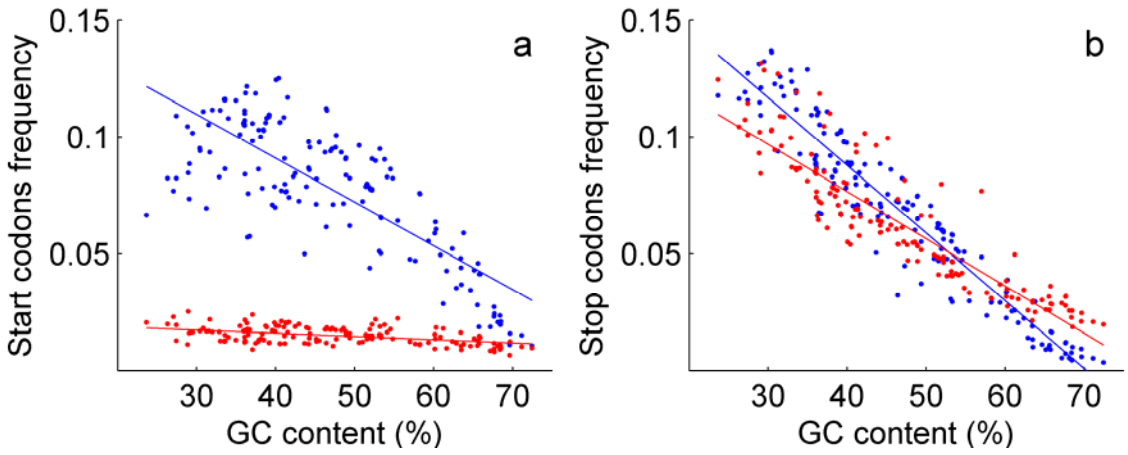

Figure 6.4: a. Start codon frequencies in phase-1 (blue) and phase-2 (red) reading frames plotted against genomic GC content. b. Stop codon frequencies in phase-1 (blue) and phase-2 (red) reading frames plotted against genomic GC content.

I found that the frequency of start codons in phase 1 is significantly higher than that in phase 2 by a factor of 5.2 on average (Figure 6.4a, paired Student t-test, $p < 0.001$). There is no significant difference between the frequencies of stop codons in the two phases (Figure 6.4b, paired Student t-test, $p = 0.13$). These results suggest that the difference between the number of long overlaps in phase 1 and phase 2 is primarily influenced by the frequencies of start codons in the two reading frames. The difference in start codon frequencies between phase 1 and phase 2 can be explained by the codons in phase 0 that may potentially lend a dinucleotide to a start codon (ATG, GTG, and TTG) in each of the phases. In phase 2, all start codons consist of phase-0 TGN codons, which may lend TG to form a phase-2 start codon. One of these codons, TGA, is a stop codon that cannot be a part of long overlap. The remaining three codons (TGT, TGC, TGG) encode for two amino acids (cysteine and tryptophan), which are among the rarest

in protein-coding genes, with a mean frequency of ~1% (Table 6.2). In contrast, in phase 1, the amino acids coded by NAT, NGT, and NTT codons that may lend a dinucleotide to one of the start codons (ATG, GTG, and TTG, respectively), are found in moderate to high frequencies in proteins (Table 6.2). Interestingly, the abundance of NAT-, NGT-, and NTT-encoded amino acids is inversely correlated with the frequency of start codons (Table 6.2). Moreover, amino acids encoded by NAT codons which can form the most common start codon, ATG, appear in lower frequencies than amino acids encoded by NGT- and NTT-encoded amino acids. For all bacteria and for all GC contents the frequencies of amino acids coded by TGN codons are lower than each of the amino acid groups encoded by NAT, NGT, and NTT (Figure 6.5, all pairwise paired Student t-tests, $p < 0.001$).

Table 6.2: Codons in phase 0 that may lend a dinucleotide to form a start codon in phase 1 and phase 2. The usage of each start codon in (a) all genes; (b) the downstream gene of long phase-1 overlaps; and (c) the downstream gene of long phase-2 overlaps, is noted.

Start Codon (usage in: all genes, phase1, phase 2)	Phase	Codon Group	Amino Acids	Mean amino acid frequency
ATG ([a]77%, [b]73%, [c]64%)	1	NAT	Tyr, His, Asn, Asp	3.67%
	2	TGN	Cys, Trp	1.06%
GTG ([a]14%, [b]15%, [c]23%)	1	NGT	Cys, Arg, Ser, Gly	4.87%
	2	TGN	Cys, Trp	1.06%
TTG ([a]9%, [b]12%, [c]14%)	1	NTT	Phe, Leu, Ile, Val	7.12%
	2	TGN	Cys, Trp	1.06%

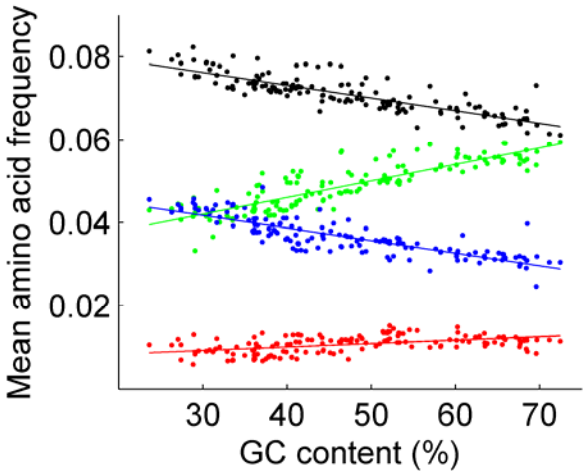

Figure 6.5: Mean frequencies of groups of amino acids in the 167 bacterial genomes plotted against genomic GC content. Mean frequency of amino acids, which are encoded by TGN, NAT, NGT, or NTT codons, are marked in red, blue, green, and black, respectively. NAT, NGT, and NTT codons may lend a dinucleotide to one of the start codons in phase 1. TGN codons may lend a dinucleotide to one of the start codons in phase 2.

Thus, consideration of the number of amino acids and their frequencies alone will lead us to expect start codons to occur much more frequently in phase 1 than in phase 2. However, the difference in amino acids usage does not provide a very good fit to the observed frequencies. This can be achieved by a more detailed compositional argument, one that is based on codon frequencies. Such a model will accommodate differences in GC content and codon usage among the bacteria under study. I found that the frequencies of the codons that combine to form start and stop codons (e.g., $f^0_{NAT} \times f^0_{GNN}$

and $f^0_{NNA} \times f^0_{TGN}$ for ATG), are strongly correlated with the frequencies of start and stop codons in both phases, as well as with genomic GC content (Table 6.3).

Table 6.3: The correlation between the frequency of frame-shift start and stop codons and (a) their expected frequencies; and (b) the genomic GC content. All correlations are significant at the $p < 0.001$ level (sample size is 167).

Frame-Shift Codon		Phase	Combining Codons	[a]Correlation Observed-Expected	[b]Correlation Observed-GC%
Start	ATG	1	N**AT,G**NN	0.96	–0.84
		2	NN**A,TG**N	0.89	–0.76
	GTG	1	N**GT,G**NN	0.94	–0.34
		2	NN**G,TG**N	0.86	0.80
	TTG	1	N**TT,G**NN	0.96	–0.80
		2	NN**T,TG**N	0.87	–0.70
Stop	TAA	1	N**TA,A**NN	0.98	–0.87
		2	NN**T,AA**N	0.97	–0.93
	TAG	1	N**TA,G**NN	0.96	–0.89
		2	NN**T,AG**N	0.90	–0.84
	TGA	1	N**TG,A**NN	0.86	0.51
		2	NN**T,GA**N	0.92	–0.84

To control for potential annotation errors, I used a subset of overlapping genes that were not annotated as "hypothetical," "putative" or "pseudogene" in the NCBI genome data. This subset of overlapping genes, which I assume to be more accurately annotated, contains 31,767 gene pairs (45% of the complete set). As in the complete set, the

frequency of long phase-1 overlaps exceeds the frequency of long phase-2 overlaps by a factor of 3.1 and the frequency of long phase-1 overlaps is negatively correlated with genomic GC content ($r = –0.28$, $p < 0.001$), whereas the frequency of long phase-2 overlaps is not ($p = 0.6$). Therefore, the influence of misannotation seems not to be significant.

Discussion

Understanding the distribution of overlapping genes in different phases is a key step towards distinguishing between the effects of selection and mutation on the evolution of overlapping genes. Krakauer (2000) showed that overlapping genes in different orientations and phases differ in the freedom for each gene to evolve independently. Therefore, he suggested that the variation in protein evolvability would be reflected in the frequency of the overlap phases. In the case of same-strand overlapping genes, his model predicted no difference between the frequency of phase-1 and phase-2 overlaps (Krakauer 2000). However, in agreement with previous studies (Johnson and Chisholm 2004; Cock and Whitworth 2007; Lillo and Krakauer 2007), my results indicate a preponderance of long phase-1 overlaps over long phase-2 overlaps. Cock and Whitworth (2007) attributed the difference between the number of long overlaps in the two phases to either gene location or to an unspecified selective advantage.

Considering the two scenarios for the creation of same-strand overlapping genes (Figure 6.2), I showed that the phase bias in long overlaps might be attributed to a great extant to overlaps created by 5'-end mutation of the downstream gene. Since there is purifying selection against long overlaps, the frequency of start codons in phase 2 constrains the number of overlap that can be created in that phase and leads to the phase bias. In addition, I showed that the difference in start codon frequencies between phase 1 and phase 2 is dictated by the frequencies of amino acids whose codons may combine to form start codons in the two phases. Finally, the dependency of frame-shift start and stop codons on species-specific codon usage result in a correlation between long phase-1 overlap frequency and genomic GC content.

Although my model explains the phase bias in overlap frequency, I do not have a full explanation for the absence of correlation between GC content and long phase-2 overlaps as expected from the frequency of frame-shift start and stop codons. This correlation is expected to have lower statistical significance than that of phase-1 overlaps because of the smaller sample size, but it is also possible that other factors affect the potential for overlap as well. A more complex compositional model for overlapping genes frequency, might include the length distribution of overlaps, the frequencies of regulatory elements (e.g., Shine-Delgarno sequences) and the strand-specific composition bias, since bacterial genomes have an asymmetrical chirochoric base composition (Lobry 1996b; Lobry 1996a; Frank and Lobry 1999).

The wide abundance of overlapping genes and the straightforward definition of phase evolvability make the phase distribution of overlapping genes an interesting case study. If evolvability is selected for, the expectation is for a positive correlation to exist between the frequency of an overlap phase and its evolvability. Evolvability considerations predict phase-1 and phase-2 overlaps to occur at equal frequencies (Krakauer 2000). Therefore, my data does not support a role for evolvability in the evolution of same-strand overlapping genes.

Fukuda et al. (2003) examined homologous overlapping genes in related bacterial species and found that the rate of accumulation and degradation of overlapping pairs is higher for overlaps caused by mutation at the 3'-end of the upstream gene compared to overlaps caused by mutation at the 5'-end of the downstream gene. The difference in rates was suggested to be a result of an evolutionary constraint imposed on the 5'-end of genes (Fukuda, Nakayama, and Tomita 2003). Our model predicts a difference in these rates simply because of the higher frequency of frame-shift stop codons compared to the frequency of frame-shift start codons. It would be interesting to test whether the rate difference of accumulation and degradation of overlapping gene pairs in the two scenarios holds even when accounting for the difference in frequency of frame-shift stop codons compared to frame-shift start codons.

The high frequency of frame-shift stop codons was previously suggested to be under positive selection for minimization of frame-shift translation errors (Seligmann and Pollock 2004; Itzkovitz and Alon 2007). I found that the frequency of frame-shift stop

codons is strongly correlated with genomic GC content leading to AT-rich genomes having five times more frame-shift stop codons than GC-rich genomes. Therefore, it seems that the mutation pattern is a major player in determining frame-shift stop-codon frequencies, while selection does not seem to play a major role.

Viral genomes also exhibit high frequencies of overlapping genes. In a study of RNA viruses, Belshaw et al. (2007) distinguished between internal overlaps, in which one gene is embedded within the other, and terminal overlaps. For internal overlaps, it was found that, similar to bacteria, there is a predominance of phase-1 overlaps (Belshaw, Pybus, and Rambaut 2007). In the case of terminal overlaps, Belshaw et al. (2007) reported no frequency difference between phase 1 and phase 2. However, Belshaw et al. (2007) did not distinguish between short overlaps, in which phase-1 overlaps are extremely rare, and long overlaps. I showed that at least as far as bacteria are concerned, pooling short and long overlaps together results in obscuring the pattern for long overlaps (Table 6.1). Therefore, the similar frequencies of over all overlaps in phase 1 and phase 2 in RNA viruses (2007), suggests that the phase bias in long overlaps was most likely unnoticed.

In this chapter, I have shown that the phase-distribution of same-strand overlapping genes in bacteria is determined by the frame-shift frequencies of start and stop codons in protein-coding genes. The predominance of long phase-1 overlaps results from a lower frequency of start codons in phase 2 that limits the potential overlaps created by an upstream extension of the downstream gene. The difference in the frequency of start

codons is dictated by the abundance of those amino acids that are encoded by codons that combine to form start codons in phase 1 and phase 2. This difference is conserved among all the bacterial genomes in the study. The variability of codon usage across bacterial genomes leads to a correlation between long phase-1 overlaps and genomic GC content. My model explains the phase bias in same-strand overlapping genes by compositional factors without invoking selection. Therefore, it can be used as a null model of neutral evolution for testing selection hypotheses affecting the evolution of overlapping genes.

Chapter Seven: Summary

Overlapping genes were first discovered in viruses and for many years were considered limited to small genomes (Barrell, Air, and Hutchison 1976; Szekely 1978). Later, it became clear that gene overlap is present in all domains of life. As a result, several studies have suggested unique roles for gene overlap in multiple regulatory processes (Normark et al. 1983; Cooper et al. 1998; Boi, Solda, and Tenchini 2004; Johnson and Chisholm 2004) and in the evolution of genome architecture (Keese and Gibbs 1992; Makalowska, Lin, and Hernandez 2007; Assis et al. 2008). Recently, studies have begun to employ systematic and comparative-genomic approaches to elucidate the evolutionary dynamics of overlapping genes.

Shortly after the discovery of overlapping genes, Miyata and Yasunaga (1978) noted that because of the overlap, the evolutionary rates of the two genes are interdependent. Still, this sequence interdependence remained a challenge in molecular evolutionary analyses and was, consequently, ignored by many studies. In Chapter Two, I demonstrated that estimates of selection intensity that ignore gene overlap are biased and that this bias differs among overlap types. I presented a new method for the simultaneous estimation of selection intensities in overlapping genes that accounts for the sequence interdependence and allows for an accurate estimation of selection intensities. With the new method, I showed that overlapping genes are mostly subjected to purifying selection, in contradistinction to previous studies, which detected an inordinate prevalence of positive selection. In future studies, it would be valuable to extend this method to deal with multiple sequences in a phylogenetic framework and to incorporate models of variable selection pressures among lineages and among sites, in analogy to the

methodology employed with non-overlapping sequences (Nielsen and Yang 1998; Yang and Nielsen 1998; Zhang, Nielsen, and Yang 2005).

In some cases, overlap occurs between genes in which one (or both) is an RNA gene (Sleutels, Zwart, and Barlow 2002; Das 2009). Therefore, another important extension would be to enable the estimation of selection in overlaps between two RNA genes and between an RNA gene and a protein-coding gene. Similar to protein-coding genes, models of nucleotide substitution in RNA genes (e.g., Rzhetsky 1995; Yu and Thorne 2006) could be incorporated to account for the sequence interdependence of gene overlap and allow an accurate estimation of selection intensity in these cases.

In Chapter Three, I used the new method to estimate selection intensity, thereby distinguishing between spurious and functional overlapping genes. I examined the "Rosetta stone" hypothesis for the origin of the two aminoacyl tRNA synthetase classes from a pair of overlapping genes (Carter and Duax 2002). This fascinating hypothesis, whose implications on other questions (such as the origin of the genetic code) are wide-ranging (Delarue 2007; Rodin and Rodin 2008; Schimmel 2008), was recently questioned (Williams, Wolfe, and Fares 2009). I used my method, which is independent to the approach of Williams, Wolfe, and Fares (2009), to show that there is no signature of purifying selection acting on the overlapping ORF. This result implies that the gene is non-functional, thus rejecting the "Rosetta stone" hypothesis.

Although false-positive predictions of overlapping genes are problematic, it is the false-negatives, i.e., genes that were missed in the annotation, that pose an even greater problem, especially in viruses with small genomes. In Chapter Four, I presented evidence for the existence of a novel overlapping gene in the genomes of the Israeli acute paralysis virus (IAPV) and three related viruses. IAPV was found to be associated with colony collapse disorder (Cox-Foster et al. 2007), a syndrome characterized by the mass disappearance of honeybees from hives (Oldroyd 2007). I hope that the discovery of this new gene will improve our understanding of this virus and its interaction with its host. The belated discovery of new overlapping genes in well-studied viruses (e.g., in influenza A, Chen et al. 2001), suggests that there are many more unidentified overlapping genes waiting to be found.

My method for detecting functional overlapping genes, which I discussed in Chapters Three and Four, as well as other methods (e.g., Firth and Brown 2006) are limited to the analysis of orthologous sequences within a range of divergence of about 5-30%. In Chapter Five I presented a method for the detection of selection signatures on hypothetical overlapping genes using population-level data (less than 5% divergence). Although, as I showed, the method is not ideal, it is the first attempt to tackle this problem and, thus, can be considered as a starting point for the development of more advanced methods.

Future studies should also focus on development of a statistical framework for predicting the functionality of a hypothetical overlapping gene when sequence divergence is high

(above 30%). Such a framework could model the factors that affect the conservation of a non-functional overlapping ORF throughout evolution, including (1) time, (2) mutation rate, (3) mutation pattern, (4) ORF length, and (5) selection intensity on the overlapping functional gene. Given that the current genomic data is only a small fraction of the world-wide genetic pool, I believe that use of methods for the detection of functional overlapping genes will continue to unveil new genes for many years to come.

Although the sequence interdependence imposed by gene overlap adds complexity to many molecular evolutionary analyses, it also provides us with the opportunity to study a couple of fundamental questions in evolution. One such question is the evolvability of biological entities, which has been a subject of great interest in the recent years (reviewed in Pigliucci 2008). A biological system is evolvable if it can acquire novel functions through genetic change. However, the quantification of evolvability has been a difficult task. In the case of overlapping genes, evolvability is defined simply as the degree to which each of the overlapping genes can evolve independently (Krakauer 2000). Several studies suggested that the phase distribution of overlapping genes is shaped by positive selection for genes in evolvable phases (Krakauer 2000; Rogozin et al. 2002). In contrast, I have shown in Chapter Six that the phase distribution of overlapping genes (at least in bacteria) can be explained by the frequencies of start and stop codons in the different phases (see also Kingsford, Delcher, and Salzberg 2007; Sabath, Graur, and Landan 2008). These frequencies, which specify the potential for the creation of overlapping genes, are determined by the abundance of amino acids in proteins and by species-specific codon usage. Another interesting result that came out of

this study reflects on the evolution of the genetic code, which was previously suggested to be under selection for minimization of frame-shift translation errors (Seligmann and Pollock 2004; Itzkovitz and Alon 2007). I have found that AT-rich genomes have five times more frame-shift stop codons than GC-rich genomes. Therefore, it seems that the impact of selection on frame-shift stop codon frequency should be small compared to the impact of the mutation pattern that affects genome composition.

The variation of evolutionary rates among proteins is another topic that has been under extensive investigation (Graur and Li 2000). Among several possible factors that influence this variation, the age of a gene was found to be inversely correlated with evolutionary rate (Alba and Castresana 2005; Toll-Riera et al. 2009; Wolf et al. 2009). This inverse relationship was suggested to result from stronger purifying selection on old genes than on new ones. In a simple simulation study (Elhaik, Sabath, and Graur 2006), I demonstrated a bias in the common methodology for age classification of genes that employs homology searches by Blast (Altschul et al. 1990). Since genetic distance increases with time of divergence and rate of evolution, it is difficult to identify homologs of fast-evolving genes in distantly related taxa. Thus, fast-evolving genes could be misclassified as new (Elhaik, Sabath, and Graur 2006). In the case of overlapping genes, the age of the younger gene within an overlapping pair can be assessed while its overlapping gene serves as a control for homology detection. Hence, overlapping genes could facilitate the study of the relationship between gene age and the rate of evolution.

Finally, many questions regarding overlapping genes remain open. As first predicted by Barrell, Air, and Hutchison (1976) and later confirmed by Belshaw, Pybus, and Rambaut (2007), gene overlap is more common in small genomes. However, we do not know yet whether the benefit of the overlap is caused by a lower genomic (contrary to per-base) mutation rate, a faster replication rate, a physical compactness of the genome, or other factors. The regulatory roles of overlapping genes, which have been documented by individual examples (Cooper et al. 1998; Yu et al. 2007; Herrera et al. 2008; Wadhawan, Dickins, and Nekrutenko 2008), is another topic that would benefit from more comprehensive study. Specifically, the abundant data on gene expression, protein abundance, and protein interactions, may be used to test the implications of overlap on gene regulation and protein-protein interactions on a large scale. I hope that the tools I have developed will prove useful in addressing these challenges and, as phrased by Boi et al. (2004), will help "shed light on the dark side of the genome."

References

Alba, M. M., and J. Castresana. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol **22**:598-606.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol **215**:403-410.

Assis, R., A. S. Kondrashov, E. V. Koonin, and F. A. Kondrashov. 2008. Nested genes and increasing organizational complexity of metazoan genomes. Trends Genet **24**:475-478.

Attwood, T. K., M. J. Blythe, D. R. Flower, A. Gaulton, J. E. Mabey, N. Maudling, L. McGregor, A. L. Mitchell, G. Moulton, K. Paine, and P. Scordis. 2002. PRINTS and PRINTS-S shed light on protein ancestry. Nucleic Acids Res **30**:239-241.

Baez, M., R. Taussig, J. J. Zazra, J. F. Young, P. Palese, A. Reisfeld, and A. M. Skalka. 1980. Complete nucleotide sequence of the influenza A/PR/8/34 virus NS gene and comparison with the NS genes of the A/Udorn/72 and A/FPV/Rostock/34 strains. Nucleic Acids Res **8**:5845-5858.

Bao, Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. 2008. The influenza virus resource at the National Center for Biotechnology Information. J Virol **82**:596-601.

Barrell, B. G., G. M. Air, and C. A. Hutchison, 3rd. 1976. Overlapping genes in bacteriophage phiX174. Nature **264**:34-41.

Belshaw, R., O. G. Pybus, and A. Rambaut. 2007. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res **17**:1496-1504.

Blanchard, P., F. Schurr, O. Celle, N. Cougoule, P. Drajnudel, R. Thiery, J. P. Faucon, and M. Ribiere. 2008. First detection of Israeli acute paralysis virus (IAPV) in France, a dicistrovirus affecting honeybees (Apis mellifera). J Invertebr Pathol **99**:348-350.

Boi, S., G. Solda, and M. L. Tenchini. 2004. Shedding light on the dark side of the genome: overlapping genes in higher eukaryotes Current Genomics **5**:509-524.

Campitelli, L., M. Ciccozzi, M. Salemi, F. Taglia, S. Boros, I. Donatelli, and G. Rezza. 2006. H5N1 influenza virus evolution: a comparison of different epidemics in birds and humans (1997-2004). J Gen Virol **87**:955-960.

Carter, C. W., and W. L. Duax. 2002. Did tRNA synthetase classes arise on opposite strands of the same gene? Mol Cell **10**:705-708.

Chen, N., and L. D. Stein. 2006. Conservation and functional significance of gene topology in the genome of *Caenorhabditis elegans*. Genome Res **16**:606-617.

Chen, W., P. A. Calvo, D. Malide, J. Gibbs, U. Schubert, I. Bacik, S. Basta, R. O'Neill, J. Schickli, P. Palese, P. Henklein, J. R. Bennink, and J. W. Yewdell. 2001. A novel influenza A virus mitochondrial protein that induces cell death. Nat Med **7**:1306-1312.

Chung, B. Y., W. A. Miller, J. F. Atkins, and A. E. Firth. 2008. An overlapping essential gene in the Potyviridae. Proc Natl Acad Sci U S A **105**:5897-5902.
Clifford, M., J. Twigg, and C. Upton. *in press*. Evidence for a novel gene associated with human influenza A viruses.
Cock, P. J., and D. E. Whitworth. 2007. Evolution of gene overlaps: relative reading frame bias in prokaryotic two-component system genes. J Mol Evol **64**:457-462.
Cooper, P. R., N. J. Smilinich, C. D. Day, N. J. Nowak, L. H. Reid, R. S. Pearsall, M. Reece, D. Prawitt, J. Landers, D. E. Housman, A. Winterpacht, B. U. Zabel, J. Pelletier, B. E. Weissman, T. B. Shows, and M. J. Higgins. 1998. Divergently transcribed overlapping genes expressed in liver and kidney and located in the 11p15.5 imprinted domain. Genomics **49**:38-51.
Cox-Foster, D. L., S. Conlan, E. C. Holmes, G. Palacios, J. D. Evans, N. A. Moran, P. L. Quan, T. Briese, M. Hornig, D. M. Geiser, V. Martinson, D. vanEngelsdorp, A. L. Kalkstein, A. Drysdale, J. Hui, J. Zhai, L. Cui, S. K. Hutchison, J. F. Simons, M. Egholm, J. S. Pettis, and W. I. Lipkin. 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. Science **318**:283-287.
Das, S. 2009. Evolutionary origin and genomic organization of micro-RNA genes in immunoglobulin lambda variable region gene family. Mol Biol Evol **26**:1179-1189.
de Groot, S., T. Mailund, and J. Hein. 2007. Comparative annotation of viral genomes with non-conserved gene structure. Bioinformatics **23**:1080-1089.
de Groot, S., T. Mailund, G. Lunter, and J. Hein. 2008. Investigating selection on viruses: a statistical alignment approach. BMC Bioinformatics **9**:304.
de Miranda, J. R., M. Drebot, S. Tyler, M. Shen, C. E. Cameron, D. B. Stoltz, and S. M. Camazine. 2004. Complete nucleotide sequence of Kashmir bee virus and comparison with acute bee paralysis virus. J Gen Virol **85**:2263-2270.
Delarue, M. 2007. An asymmetric underlying rule in the assignment of codons: possible clue to a quick early evolution of the genetic code via successive binary choices. RNA **13**:161-169.
Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res **27**:4636-4641.
Delport, W., K. Scheffler, and C. Seoighe. 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLoS Pathog **4**:e1000242.
Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol **7**:214.
Elhaik, E., N. Sabath, and D. Graur. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol **23**:1-3.
Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. Am. Nat. **106**:645-668.
Finn, R. D., J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. 2008. The Pfam protein families database. Nucleic Acids Res **36**:D281-288.
Firth, A. E. 2008. Bioinformatic analysis suggests that the Orbivirus VP6 cistron encodes an overlapping gene. Virol J **5**:48.

Firth, A. E., and J. F. Atkins. 2009. Analysis of the coding potential of the partially overlapping 3' ORF in segment 5 of the plant fijiviruses. Virol J **6**:32.
Firth, A. E., and J. F. Atkins. 2008a. Bioinformatic analysis suggests that a conserved ORF in the waikaviruses encodes an overlapping gene. Arch Virol **153**:1379-1383.
Firth, A. E., and J. F. Atkins. 2008b. Bioinformatic analysis suggests that the Cypovirus 1 major core protein cistron harbours an overlapping gene. Virol J **5**:62.
Firth, A. E., and C. M. Brown. 2005. Detecting overlapping coding sequences with pairwise alignments. Bioinformatics **21**:282-292.
Firth, A. E., and C. M. Brown. 2006. Detecting overlapping coding sequences in virus genomes. BMC Bioinformatics **7**:75.
Fisher, R. 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology **20**:406-416.
Frank, A. C., and J. R. Lobry. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238**:65-77.
Fukuda, Y., Y. Nakayama, and M. Tomita. 2003. On dynamics of overlapping genes in bacterial genomes. Gene **323**:181-187.
Fukuda, Y., T. Washio, and M. Tomita. 1999. Comparative study of overlapping genes in the genomes of Mycoplasma genitalium and Mycoplasma pneumoniae. Nucleic Acids Res **27**:1847-1853.
Gallai, N., J.-M. Salles, J. Settele, and B. E. Vaissière. 2009. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Ecological Economics **68**:810-821.
Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11**:725-736.
Govan, V. A., N. Leat, M. Allsopp, and S. Davison. 2000. Analysis of the complete genome sequence of acute bee paralysis virus shows that it belongs to the novel group of insect-infecting RNA viruses. Virology **277**:457-463.
Graur, D., and W.-H. Li. 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, Mass.
Guyader, S., and D. G. Ducray. 2002. Sequence analysis of Potato leafroll virus isolates reveals genetic stability, major evolutionary events and differential selection pressure between overlapping reading frame products. J Gen Virol **83**:1799-1807.
Hein, J., and J. Stovlbaek. 1995. A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. J Mol Evol **40**:181-189.
Herrera, V. L., P. Bagamasbad, T. Didishvili, J. L. Decano, and N. Ruiz-Opazo. 2008. Overlapping genes in Nalp6/PYPAF5 locus encode two V2-type vasopressin isoreceptors: angiotensin-vasopressin receptor (AVR) and non-AVR. Physiol Genomics **34**:65-77.
Holmes, E. C., D. J. Lipman, D. Zamarin, and J. W. Yewdell. 2006. Comment on "Large-scale sequence analysis of avian influenza isolates". Science **313**:1573; author reply 1573.

Hughes, A. L., and R. Friedman. 2003. Genome-wide survey for genes horizontally transferred from cellular organisms to baculoviruses. Mol Biol Evol **20**:979-987.
Hughes, A. L., and M. A. Hughes. 2005. Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. Virus Res **113**:81-88.
Hughes, A. L., K. Westover, J. da Silva, D. H. O'Connor, and D. I. Watkins. 2001. Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. J Virol **75**:7966-7972.
Hulo, N., A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. Sigrist. 2006. The PROSITE database. Nucleic Acids Res **34**:D227-230.
Itzkovitz, S., and U. Alon. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res **17**:405-412.
Johnson, Z. I., and S. W. Chisholm. 2004. Properties of overlapping genes are conserved across microbial genomes. Genome Res **14**:2268-2272.
Jones, C. E., T. M. Fleming, D. A. Cowan, J. A. Littlechild, and P. W. Piper. 1995. The phosphoglycerate kinase and glyceraldehyde-3-phosphate dehydrogenase genes from the thermophilic archaeon *Sulfolobus solfataricus* overlap by 8-bp. Isolation, sequencing of the genes and expression in *Escherichia coli*. Eur J Biochem **233**:800-808.
Jones, D. T. 2007. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics **23**:538-544.
Karlin, S., C. Chen, A. J. Gentles, and M. Cleary. 2002. Associations between human disease genes and overlapping gene groups and multiple amino acid runs. Proc Natl Acad Sci U S A **99**:17008-17013.
Keese, P. K., and A. Gibbs. 1992. Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci U S A **89**:9489-9493.
Kingsford, C., A. L. Delcher, and S. L. Salzberg. 2007. A unified model explaining the offsets of overlapping and near-overlapping prokaryotic genes. Mol Biol Evol **24**:2091-2098.
Konstantopoulou, I., C. A. Ouzounis, E. Drosopoulou, M. Yiangou, P. Sideras, C. Sander, and Z. G. Scouras. 1995. A Drosophila hsp70 gene contains long, antiparallel, coupled open reading frames (LAC ORFs) conserved in homologous loci. J Mol Evol **41**:414-420.
Kozak, M. 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol Rev **47**:1-45.
Kozlov, N. N. 2000. Overlapping genes and variability of the genetic code. Dokl Biol Sci **375**:677-680.
Krakauer, D. C. 2000. Stability and evolution of overlapping genes. Evolution Int J Org Evolution **54**:731-739.
Kumar, S., M. Nei, J. Dudley, and K. Tamura. 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform **9**:299-306.

Lalani, A. S., and G. McFadden. 1999. Evasion and exploitation of chemokines by viruses. Cytokine Growth Factor Rev **10**:219-233.
Landan, G., and D. Graur. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol **24**:1380-1383.
Lavorgna, G., D. Dahary, B. Lehner, R. Sorek, C. M. Sanderson, and G. Casari. 2004. In search of antisense. Trends Biochem Sci **29**:88-94.
Li, K. S., Y. Guan, J. Wang, G. J. Smith, K. M. Xu, L. Duan, A. P. Rahardjo, P. Puthavathana, C. Buranathai, T. D. Nguyen, A. T. Estoepangestie, A. Chaisingh, P. Auewarakul, H. T. Long, N. T. Hanh, R. J. Webby, L. L. Poon, H. Chen, K. F. Shortridge, K. Y. Yuen, R. G. Webster, and J. S. Peiris. 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature **430**:209-213.
Li, W. H., C. I. Wu, and C. C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol **2**:150-174.
Liang, H., and L. F. Landweber. 2006. A genome-wide study of dual coding regions in human alternatively spliced genes. Genome Res **16**:190-196.
Lillo, F., and D. C. Krakauer. 2007. A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. Biol Direct **2**:22.
Lobry, J. R. 1996a. Origin of replication of *Mycoplasma genitalium*. Science **272**:745-746.
Lobry, J. R. 1996b. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol **13**:660-665.
Luo, Y., C. Fu, D. Y. Zhang, and K. Lin. 2007. BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes. BMC Bioinformatics **8**:266.
Makalowska, I., C. F. Lin, and K. Hernandez. 2007. Birth and death of gene overlaps in vertebrates. BMC Evol Biol **7**:193.
Maori, E., S. Lavi, R. Mozes-Koch, Y. Gantman, Y. Peretz, O. Edelbaum, E. Tanne, and I. Sela. 2007. Isolation and characterization of Israeli acute paralysis virus, a dicistrovirus affecting honeybees in Israel: evidence for diversity due to intra- and inter-species recombination. J Gen Virol **88**:3428-3438.
McCauley, S., S. de Groot, T. Mailund, and J. Hein. 2007. Annotation of selection strengths in viral genomes. Bioinformatics **23**:2978-2986.
McCauley, S., and J. Hein. 2006. Using hidden Markov models and observed evolution to annotate viral genomes. Bioinformatics **22**:1308-1316.
McGuffin, L. J., K. Bryson, and D. T. Jones. 2000. The PSIPRED protein structure prediction server. Bioinformatics **16**:404-405.
McLysaght, A., P. F. Baldi, and B. S. Gaut. 2003. Extensive gene gain associated with adaptive evolution of poxviruses. Proc Natl Acad Sci U S A **100**:15655-15660.
Miyata, T., and T. Yasunaga. 1978. Evolution of overlapping genes. Nature **272**:532-535.
Monnerjahn, C., D. Techel, S. A. Mohamed, and L. Rensing. 2000. A non-stop antisense reading frame in the grp78 gene of *Neurospora crassa* is homologous to the

Achlya klebsiana NAD-gdh gene but is not being transcribed. FEMS Microbiol Lett **183**:307-312.
Montoya, J., G. L. Gaines, and G. Attardi. 1983. The pattern of transcription of the human mitochondrial rRNA genes reveals two overlapping transcription units. Cell **34**:151-159.
Murphy, P. M. 2001. Viral exploitation and subversion of the immune system through chemokine mimicry. Nat Immunol **2**:116-122.
Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**:1344-1349.
Narechania, A., M. Terai, and R. D. Burk. 2005. Overlapping reading frames in closely related human papillomaviruses result in modular rates of selection within E2. J Gen Virol **86**:1307-1313.
Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol **48**:443-453.
Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3**:418-426.
Nekrutenko, A., K. D. Makova, and W. H. Li. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res **12**:198-202.
Nguyen, M., and A. L. Haenni. 2003. Expression strategies of ambisense viruses. Virus Res **93**:141-150.
Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929-936.
Normark, S., S. Bergstrom, T. Edlund, T. Grundstrom, B. Jaurin, F. P. Lindberg, and O. Olsson. 1983. Overlapping genes. Annu Rev Genet **17**:499-525.
Obenauer, J. C., J. Denson, P. K. Mehta, X. Su, S. Mukatira, D. B. Finkelstein, X. Xu, J. Wang, J. Ma, Y. Fan, K. M. Rakestraw, R. G. Webster, E. Hoffmann, S. Krauss, J. Zheng, Z. Zhang, and C. W. Naeve. 2006. Large-scale sequence analysis of avian influenza isolates. Science **311**:1576-1580.
Oldroyd, B. P. 2007. What's killing American honey bees? PLoS Biol **5**:e168.
Palacios, G., J. Hui, P. L. Quan, A. Kalkstein, K. S. Honkavuori, A. V. Bussetti, S. Conlan, J. Evans, Y. P. Chen, D. vanEngelsdorp, H. Efrat, J. Pettis, D. Cox-Foster, E. C. Holmes, T. Briese, and W. I. Lipkin. 2008. Genetic analysis of Israel acute paralysis virus: distinct clusters are circulating in the United States. J Virol **82**:6209-6217.
Palleja, A., E. D. Harrington, and P. Bork. 2008. Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics **9**:335.
Pamilo, P., and N. O. Bianchi. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol **10**:271-281.

Parkinson, J. S. 1968. Genetics of the left arm of the chromosome of bacteriophage lambda. Genetics **59**:311-325.
Pavesi, A. 2006. Origin and evolution of overlapping genes in the family Microviridae. J Gen Virol **87**:1013-1017.
Pavesi, A. 2007. Pattern of nucleotide substitution in the overlapping nonstructural genes of influenza A virus and implication for the genetic diversity of the H5N1 subtype. Gene **402**:28-34.
Pedersen, A. M., and J. L. Jensen. 2001. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol **18**:763-776.
Pigliucci, M. 2008. Is evolvability evolvable? Nat Rev Genet **9**:75-82.
Pybus, O. G., A. Rambaut, R. Belshaw, R. P. Freckleton, A. J. Drummond, and E. C. Holmes. 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. Mol Biol Evol **24**:845-852.
Rodin, S. N., and S. Ohno. 1995. Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. Orig Life Evol Biosph **25**:565-589.
Rodin, S. N., and A. S. Rodin. 2008. On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. Heredity **100**:341-355.
Rogozin, I. B., A. N. Spiridonov, A. V. Sorokin, Y. I. Wolf, I. K. Jordan, R. L. Tatusov, and E. V. Koonin. 2002. Purifying and directional selection in overlapping prokaryotic genes. Trends Genet **18**:228-232.
Rother, K. I., O. K. Clay, J. P. Bourquin, J. Silke, and W. Schaffner. 1997. Long non-stop reading frames on the antisense strand of heat shock protein 70 genes and prion protein (PrP) genes are conserved between species. Biol Chem **378**:1521-1530.
Rzhetsky, A. 1995. Estimating substitution rates in ribosomal RNA genes. Genetics **141**:771-783.
Sabath, N., D. Graur, and G. Landan. 2008. Same-strand overlapping genes in bacteria: compositional determinants of phase bias. Biol Direct **3**:36.
Sabath, N., G. Landan, and D. Graur. 2008. A method for the simultaneous estimation of selection intensities in overlapping genes. PLoS ONE **3**:e3996.
Sabath, N., N. Price, and D. Graur. 2009. A potentially novel overlapping gene in the genomes of Israeli acute paralysis virus and its relatives. Virol J **6**:144.
Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4**:406-425.
Sakharkar, K. R., M. K. Sakharkar, C. Verma, and V. T. Chow. 2005. Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. Int J Syst Evol Microbiol **55**:1205-1209.
Schimmel, P. 2008. Development of tRNA synthetases and connection to genetic code and disease. Protein Sci **17**:1643-1652.
Seligmann, H., and D. D. Pollock. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol **23**:701-705.

Silke, J. 1997. The majority of long non-stop reading frames on the antisense strand can be explained by biased codon usage. Gene **194**:143-155.
Sleutels, F., R. Zwart, and D. P. Barlow. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature **415**:810-813.
Smith, R. A., and J. S. Parkinson. 1980. Overlapping genes at the cheA locus of *Escherichia coli*. Proc Natl Acad Sci U S A **77**:5370-5374.
Smith, T. F., and M. S. Waterman. 1981. Overlapping genes and information theory. J Theor Biol **91**:379-380.
Suzuki, Y. 2006. Natural selection on the influenza virus genome. Mol Biol Evol **23**:1902-1911.
Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.
Szekely, M. 1978. Triple overlapping genes. Nature **272**:492.
Thompson, J. D., T. J. Gibson, and D. G. Higgins. 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics **Chapter 2**:Unit 2 3.
Toll-Riera, M., N. Bosch, N. Bellora, R. Castelo, L. Armengol, X. Estivill, and M. M. Alba. 2009. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol **26**:603-612.
Valles, S. M., C. A. Strong, P. M. Dang, W. B. Hunter, R. M. Pereira, D. H. Oi, A. M. Shapiro, and D. F. Williams. 2004. A picorna-like virus from the red imported fire ant, *Solenopsis invicta*: initial discovery, genome sequence, and characterization. Virology **328**:151-157.
van Engelsdorp, D., J. Hayes, Jr., R. M. Underwood, and J. Pettis. 2008. A survey of honey bee colony losses in the U.S., fall 2007 to spring 2008. PLoS ONE **3**:e4071.
Vandenberg, S. 1967. Hereditary factors in psychological variables in man, with special emphasis on cognition. Pp. 99-134 *in* J. Spuhler, ed. Genetic Diversity and Human Behavior. Wenner-Gren, NY.
Wadhawan, S., B. Dickins, and A. Nekrutenko. 2008. Wheels within wheels: clues to the evolution of the Gnas and Gnal loci. Mol Biol Evol **25**:2745-2757.
Williams, T. A., K. H. Wolfe, and M. A. Fares. 2009. No rosetta stone for a sense-antisense origin of aminoacyl tRNA synthetase classes. Mol Biol Evol **26**:445-450.
Wolf, Y. I., P. S. Novichkov, G. P. Karev, E. V. Koonin, and D. J. Lipman. 2009. Inaugural Article: The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci U S A **106**:7273-7280.
Yang, Z. 2006. Computational Molecular Evolution. Oxford Oxfordshire: Oxford University Press.
Yang, Z., and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol **46**:409-418.
Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17**:32-43.
Yang, Z., R. Nielsen, and M. Hasegawa. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol **15**:1600-1611.

Yu, J., and J. L. Thorne. 2006. Dependence among sites in RNA evolution. Mol Biol Evol **23**:1525-1537.

Yu, J. S., R. J. Kokoska, V. Khemici, and D. A. Steege. 2007. In-frame overlapping genes: the challenges for regulating gene expression. Mol Microbiol **63**:1158-1172.

Zaaijer, H. L., F. J. van Hemert, M. H. Koppelman, and V. V. Lukashov. 2007. Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. J Gen Virol **88**:2137-2143.

Zhang, J., and M. Nei. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol **44 Suppl 1**:S139-146.

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol **22**:2472-2479.

Zhirnov, O. P., S. V. Poyarkov, I. V. Vorob'eva, O. A. Safonova, N. A. Malyshev, and H. D. Klenk. 2007. Segment NS of influenza A virus contains an additional gene NSP in positive-sense orientation. Dokl Biochem Biophys **414**:127-133.

Zhong, W., P. A. Reche, C. C. Lai, B. Reinhold, and E. L. Reinherz. 2003. Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire. J Biol Chem **278**:45135-45144.