Identification and Functional Characterization of Highly Conserved DNA Sequences in Poxvirus Genomes

By

Aliya Mehreen Sadeque
B.Sc., Queen's University, 2007

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Biochemistry and Microbiology

Supervisory Committee

Identification and Functional Characterization of Highly Conserved Sequences in Poxvirus Genomes

By

Aliya Mehreen Sadeque
B.Sc., Queen's University, 2007

Supervisory Committee

Dr. Christopher Upton (Department of Biochemistry and Microbiology)
Supervisor

Dr. Caroline Cameron (Department of Biochemistry and Microbiology)
Departmental Member

Dr. Ulrike Stege (Department of Computer Science)
Outside Member

Abstract

Supervisory Committee

Dr. Christopher Upton, (Department of Biochemistry and Microbiology)
Supervisor
Dr. Caroline Cameron, (Department of Biochemistry and Microbiology)
Departmental Member
Dr. Ulrike Stege, (Department of Computer Science)
Outside Member

The focus of this dissertation is the use of bioinformatics in the identification of highly conserved sequences among a set of poxvirus genomes and the subsequent functional analysis of the conserved functions of these sequences. A novel algorithm, Java Pattern Finder, which identifies sequences of a user-specified length that are conserved with a user-specified number of allowed differences, was used to identify near-perfectly conserved sequences among a set of poxvirus genomes. A scoring method was established to quantify the degree of conservation of these sequences and used to show that the 11 most conserved sequences were significantly more conserved than control sequences. Functional analysis showed that explanations such as low codon degeneracy or the presence of conserved promoter elements partially – but not fully – accounted for the conservation observed in these sequences, suggesting that these highly conserved regions may have novel functions in the poxvirus genome that have yet to be uncovered.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AGS program	Aliya's Gene Sequence program
AT	Adenine + Thymine
bp	base pairs
CSE	conserved sequence element
CVA	Chorioallantois *Vaccinia virus* Ankara
Da	Dalton
DNA	Deoxyribonucleic Acid
E/I/L	Early/Intermediate/Late
E-value	expected value
GC	Guanine + Cytosine
GTPV	Goatpox virus
GUI	Graphical user interface
HIV	Human Immunodeficiency Virus
IMV	Intracellular Mature Virus
ITR/TIR	Inverted Terminal Repeat/Terminal Inverted Repeat
JaPaFi	Java Pattern Finder
kb	kilobase pairs
kDa	kiloDalton
LCS	Longest Common Substring
LSDV	Lumpy skin disease virus
Met	Methionine
Morph	Morphogenesis
MP	Membrane Protein
mRNA	messenger Ribonucleic Acid
MVA	Modified Vaccinia Ankara
MYXV	Myxoma virus
NCBI	National Center for Biotechnology Information
nm	nanometer
nt/nts	nucleotide/nucleotides
ORF	Open Reading Frame
PCNA	proliferating cell nuclear antigen
PO4	Phosphorylated
Pol	Polymerase
poly(A)	polyadenylate
RAP94	RNA Polymerase-Associated Protein
rMVA	recombinant Modified Vaccinia Ankara
RNA	Ribonucleic Acid
SPPV	Sheeppox virus

SWPV	Swinepox virus
Tyr/Ser	Tyrosine/Serine
VACV	*Vaccinia virus*
VBRC	Viral Bioinformatics Research Center
VETF	Viral Early Transcription Factor
VGO	Viral Genome Organizer
VLTF	Viral Late Transcription Factor
VOCs	Viral Orthologous Clusters
WHO	World Health Organization
YLDV	Yaba-like disease virus
YMTV	Yaba monkey tumor virus

Acknowledgements

First and foremost I'd like to thank my supervisor, Dr. Chris Upton, whose guidance and support were so integral in my first venture into the science world as a 'big kid' (read: graduate student). I can't express how much I appreciate your tireless hours of helping me revise and edit this dissertation and the eight drafts that preceded it. It has been a privilege and an honour working with you.

To all of the strong and inspiring women in my life who I have always tried to follow by example, please know what a profound impact you've had on me. To my support network – Celeste, Kate, Kat, Qian, Calli, Katie, Laura and Mel –you are truly remarkable women. I am so grateful for having had the chance to learn from the very best just what friendship means. To Melissa, my mentor, big sister and best friend who showed me the ropes on life as a graduate student and always calmed me down when the 'sequences' hit the fan - I could not have asked for a better role model in the early stages of my career, nor could I think of anyone I'd rather spend 40 hours a week with. Thank you for making me a part of your life, little Simon is the apple of his Auntie Aliya's eye. To my friend and colleague Katie Gregg, thank you for all of your advice and support and for being the tiny powerhouse in my corner. To my committee members, Drs. Caroline Cameron and Ulrike Stege, your guidance has been elemental over the last two years. Lastly, my thanks to Dr. Elisabeth Tillier for giving me my first taste of dry-lab work. My time in your lab is what sparked my interest in Bioinformatics and I haven't turned back since.

To my former labmate Gord, whose astounding computer expertise have been a huge asset to me over the years, thank you for all of the tips, the scripts, the chats in the lab, and the innumerable rounds of Scrabulous. To Dan Godlovitch, who wrote a program for my project and christened it with my name, thanks for all the hours of coding you've put in and for teaching me everything I now know – which mind you, isn't much – about ice growth.

To my dear friend Ian Van Toch, who was a brilliant scientist taken from us far too soon, rest in peace.

To the ladies and gent in the department office – John Hall, Deb Penner, Melinda Powell and Sandra Boudewyn – you are the gems of our department. Thank you for keeping the machine running smoothly, you have all been so helpful in innumerable ways over the years.

And lastly, my deepest thanks to my family, whose unwaivering love and support astound me. To Ammu and Abbu, who taught me honesty and integrity and then set me loose on the world, everything I have achieved is by your grace. To Fuzzy, who is, hands down, the best big brother in the history of time, I could not invent a better lifelong partner in crime. Trust me, I tried. Both Googa and Borshun were very disappointing. I love you all with all of my heart. This dissertation is for you.

1. Introduction

1.1. Introduction to the taxonomic family *Poxviridae*

1.1.1. A Brief History of Poxviruses

The taxonomic family Poxviridae contains large double stranded-DNA viruses and is divided into two subfamilies; viruses in the *Chordopoxvirinae* subfamily infect vertebrates and make up 10 genera, whereas viruses in the *Entomopoxvirinae* subfamily infect insects and consist of four genera.

The ranks of the poxvirus family include infamous members of much historical significance to humans and also to a much wider range of hosts. One of the most well-known members is *Variola virus*, the causative agent of the acute contagious human disease smallpox. Although smallpox has been eradicated now for almost 30 years, it is still considered one of the most devastating diseases known to humanity(World Health Organization). With repeated epidemics of smallpox sweeping across entire continents for centuries, smallpox has changed the course of history. With a mortality rate of 30-35% and no effective treatment, smallpox was such a major killer of infants in some ancient cultures that newborns were not named until they had caught the disease and survived. Even today, although smallpox does not seem like a significant threat, research continues in the areas of outbreak prevention and management and further vaccine development as a precautionary measure in case smallpox is reintroduced through bioterrorism (Jacobs *et al*., 2008).

Another member of the poxvirus family of great significance to humans is *Vaccinia virus*, which has been used as the vaccine for smallpox. The smallpox vaccine was the first vaccine ever developed, and its administration through vaccination campaigns during the 19th and 20th centuries led to a dramatic decline in smallpox infection. Between 1950 and 1967, the number of occurrences of smallpox per year dropped from an estimated 50 million to around 10-15 million. In 1966, the World Health Assembly adopted a resolution accepting the need for coordination among the eradication programs of individual countries, which resulted in the Intensified Smallpox Eradication Program being put into effect in 1967(Parrino and Graham, 2006). As part of the Intensified Smallpox Eradication Program a Smallpox Eradication Unit was established to coordinate the eradication effort from WHO headquarters in Geneva(Bhattacharya and Dasgupta, 2009). In 1980, the World Health Assembly announced the global eradication of smallpox, making it the only human infectious disease to date to be completely eradicated(Jacobs *et al*., 2008).

Even after the eradication of smallpox, *Vaccinia virus* has continued to play a significant role in several areas of biochemistry. Due to the highly conserved nature of structural proteins among orthopoxviruses, the smallpox vaccine has also served as a vaccine against infection by other poxviruses such as cowpox and monkeypox(Jacobs *et al*., 2008). Continued antiviral research on *Vaccinia virus* has produced modified vaccines with improved safety profiles. These include highly attenuated third- generation vaccines which have been modified through sequential passage in an alternative host, causing changes in viral properties such as host range, virulence and genome composition(Jacobs *et al*., 2008).. Two examples of third-generation

vaccines include LC16m8, which was passaged over 40 times through primary rabbit kidney epithelial cells and has reduced adverse effects relative to widely-used first generation vaccines (Mesedaet *al.*, 2009), and Modified Vaccinia Ankara (MVA), which was derived by passaging the chorioallantois VACV Ankara (CVA) strain of VACV nearly 600 times in chick embryo fibroblast cells, resulting in a strain that is unable to replicate productively in human cells(Garzaet *al.*,2009).

Current research is also focusing on fourth generation vaccines which have been attenuated through genetic engineering. The development of methods of genetic engineering - Insertions, deletions and interruptions of genes - have allowed for a targeted approach to attenuation while maintaining the immunogenicity of the virus. One of the best characterized examples of a fourth generation vaccine is NYVAC, a VACV strain developed as a vaccine vector by the deletion of a 18 ORFs from the VACV strain Copenhagen genome (Tartagliaet *al.*, 1992). Among the deleted ORFs were key host range genes and in deleting these genes, the virus was left unable to multiply in human cell lines (Ferrier-Rembertet *al.*, 2008). Studies on the short-term efficacy of NYVAC relative to that of the Lister strain vaccine, one of the traditional first generation vaccine strains, have shown that NYVAC induces protection and high levels of VACV-specific neutralizing antibodies and T-lymphocytes, while prime-boost vaccination studies have shown that NYVAC induced complete long term protection from death against infection in mice (Ferrier-Rembert *et al*., 2008).

Outside of antiviral research, *Vaccinia virus* has also served as a useful model for eukaryotic systems. For instance, studies conducted on the *Vaccinia virus* DNA topoisomerase have shown it to be an instructive model system for mechanistic studies of the type IB family of

DNA topoisomerases (Shuman, 1998). *Vaccinia virus* has also been found to be very accommodating of additional genetic material, successfully accepting as much as 25 kb of foreign DNA. The use of re-engineered forms of the virus in expressing foreign genes has led it to be regarded in laboratory practice as a robust vector forrecombinant protein production(Jacobs *et al*., 2008).

This same feature of *Vaccinia virus* has also made it a strong candidate for recombinant vaccine vectors; while the smallpox vaccine already provided cross-protection against a wide range of orthopoxviruses, it is now also being used to produce vaccines for a much wider range of microbial pathogens, such as rabies (Blanton*et al*., 2007) and HIV (Collier*et al.*, 1989). In the case of rabies vaccinations, first generation oral attenuated rabies virus vaccines proved effective in immunizing fox populations in Europe, but had the potential of causing vaccine-induced rabies and had much lower efficacy in a broader spectrum of host species (Blanton*et al*., 2007). A vaccinia-rabies glycoprotein recombinant virus vaccine was therefore developed in the late 1980s and remains the only licensed oral rabies vaccine in the United States to date (Blanton*et al*., 2007). In the case of HIV, many of the most promising vaccines currently in testing or in the pipeline are viral vectors expressing multiple HIV-1 antigens. Among these viral vectors, MVA has proven to be a promising candidate for a number of reasons, including the loss of immune defense genes through large deletions that arose during the passaging of the vaccine in chicken embryo fibroblasts (Earl *et al*., 2009). HIV-1 genes inserted into recombinant MVA (rMVA) have been shown to be genetically stable after repeated passage in cell culture, resulting in strong HIV-specific cellular and humoral immune responses in mice (Earl *et al*., 2009)

Many viruses have shown promise as a platform for exploratory approaches to cancer treatment given their natural ability to infect, replicate within and ultimately lyse host cells (Shen and Nemunaitis, 2005). *Vaccinia virus* in particular exhibits many properties that make it favourable as an oncolytic virus, including efficient infection and gene expression and potent lytic activity (Yu *et al*., 2009). In a recent study, an attenuated, replication-competent *Vaccinia virus*, strain GLV-1h68, has been examined as an oncolytic agent against six human squamous cell carcinoma cell lines and has, in preliminary investigations, demonstrated significant oncolytic efficacy (Yu *et al*., 2009). Myxoma virus has also been a key player in poxvirus-based cancer treatments primarily as a result of two characteristics of the virus. Firstly, it has very narrow species selectivity, making it nonpathogenic for all vertebrate species other than rabbits, and secondly because despite its narrow host range, myxoma virus can productively infect a number of different cell lines, including some human tumor cells, and replicate without causing disease (Lun *et al*., 2005). In a study conducted in 2005 by Lun *et al*., the oncolytic properties of myxoma virus against human tumor cells *in vivo* were shown for the first time, demonstrating that it infects and kills the majority of human glioma cells tested (Lun *et al*., 2005).

Although *Variola virus* and *Vaccinia virusVaccinia virus* are the most renowned members of the poxvirus family, there are many others that have been of significance to humans; such as cowpox, which Jenner identified as the first rudimentary form of a vaccine(Jacobs *et al*., 2008) and was an early example of disease transfer between mammalian species, and monkeypox, which humans contract from monkeys and squirrels, predominantly in Africa (Assarsson *et al*., 2008). In 2003, the first cluster of human monkeypox cases in the United States created a scare among viral epidemiologists (Guarner *et al*., 2004). The human infections were acquired from

infected prairie dogs, which, in turn, had acquired the infection following contact with various exotic African rodents shipped from Ghana to the United States (Guarner *et al*., 2004). However, the outbreak was of a mild variant and was easily contained (Osorio *et al*., 2009).

Collectively, poxviruses infect a very wide range of organisms including insects, birds and over 30 different mammals, making these highly successful pathogens the subject of great interest both in the context of human disease and, more generally, as agents that interact with many types of cellular systems (Upton *et al*., 2003).

1.1.2. Genome and virion structure

The poxvirus genome is a single linear, nonsegmented molecule of double-stranded DNA ranging in size from 150 – 380 kB containing 150-250 genes. This results in a very tightly-packed genome. Genes are transcribed from both DNA strands and thus far have not been shown to overlap by more than a few nucleotides (Da and Upton, 2005). Essential conserved genes, such as those encoding transcriptional, replicative and structural functions, are generally located in the central regions of the genomes, while those responsible for host range and virulence tend to be located in the terminal regions (Upton *et al*., 2003).

At the genome termini, poxviruses have terminal inverted repeat (TIR) regions frequently containing tandem repeat sequences. The TIR regions may be as long as roughly 15 kb and can also encompass transcribed regions. The double stranded DNA genome is cross-linked at both

ends (Wittek *et al*., 1978). Poxviruses are generally considered to be AT-rich, with vaccinia, the prototypal poxvirus, displaying a base composition of 66.6% A+T (Goebel *et al*., 1990). A 2006 study in which 21 poxviruses were analyzed for GC content showed that 16 out of 21 genomes contained an overall AT content of 70-82%, with the exception of 5 species (Myxomavirus, Rabbit fibroma virus, Orf virus, Bovine popular stomatitis virus and Molluscum contagiosum virus) from three different *Chordopoxvirinae* genera which had an overall AT content ranging from 35 – 60% (Barrett *et al*., 2006).

Poxviruses are enveloped viruses, meaning their genomes are packaged into viral capsids which, in turn, are covered in one or more envelopes that contain viral glycoproteins, which serve to identify and bind to receptor sites on the host's cell membranes. While most enveloped viruses form these envelopes by budding from the host cells, poxviruses package their genetic material in membranous spheres that form deep within the infected cell's cytoplasm(Heuser 2005). The resultant virion is around 200 nm in diameter and 300 nm in length, generally brick- or ovoid-shaped, and contains all components for early transcription within the core of the infectious particle. Poxviruses are the only family of DNA viruses that propagate entirely within the cytoplasm of eukaryotic cells and therefore must encode most, if not all, of the specific enzymes and factors needed for transcription, genome replication, virion production and morphogenesis (Moss *et al*., 1991).

1.1.3. Life Cycle

In the poxvirus life cycle, gene transcription is temporally regulated with genes falling under three classes: early, intermediate and late, with some genes expressed at both early and late times. These latter are referred to as "early/late" (Moss *et al*., 1991) . Following entry, the synthesis of early gene products leads to replication, followed by the expression of intermediate and late genes and, finally, assembly and release of the progeny viral particles(Moss *et al*., 1991)

Early genes encode proteins required for replication and the expression of intermediate and late genes, as well as virulence factors that modulate host response. Thus, RNA polymerase subunits, DNA polymerase and transcription factors for intermediate gene transcription are among the translation products of early genes and DNA replication can therefore occur once all early genes have been expressed (Moss *et al*., 1991). By contrast, late genes encode proteins that are involved with DNA packaging, virion morphology and cell entry, as well as early gene transcription factors for inclusion in the progeny particle (Assarsson *et al*., 2008). Intermediate gene protein products have been shown to act as *trans*-acting transcription factors necessary for the transcription of late genes (Vos and Stunnenberg, 1988). Literature searches thus far have not revealed any additional functions for intermediate genes other than *trans*-acting late gene transcription factors.

A 2006 proteomic assay surveying and quantifying the proteins in the infectious *Vaccinia virusVaccinia virus* intracellular mature virus (IMV) particle identified 75 viral proteins, including core proteins, transcription factors and enzymes, such as poly(A) polymerase subunits, capping enzymes, helicases and DNA-dependent RNA polymerase complexes (Chung *et al*., 2006). Thus, all of the components of early transcription are packaged within the core of the infectious viral

particle, allowing early gene transcription to begin immediately after entry into the host cell cytoplasm. Early gene mRNA appear within minutes of entry into the cell and are capped and polyadenylated shortly thereafter by an RNA polymerase holoenzyme that is believed, according to several lines of evidence, to assemble on early promoters during morphogenesis and virion assembly (Broyles, 2003).

DNA in the infecting viral particle only serves as template for early gene expression, not for intermediate or late transcription which require replicated DNA as template. Thus it follows that after the first phase of the poxvirus life cycle – which consists of early gene transcription and DNA replication, the poxvirus life cycle can enter its second phase in which intermediate genes are transcribed (Moss *et al*., 1991). Translation products of intermediate genes include late gene transactivators which allow transcription of late genes to occur in the third phase of the poxvirus life cycle (Baldick, Keck and Moss, 1992). To complete the cycle, late gene expression results in the production of early transcription factors, which then get packaged into progeny particles alongside RNA polymerase and other proteins (Baldick, Keck and Moss, 1992). Progeny particles are assembled and released, and go on to begin the cycle again.

It is worthy of mention that while a termination signal that takes the form of TTTTTNT is observed 20-50 nts upstream of the ends of most early mRNAs, no termination signal has been recognized in late genes. As a result, the 3' ends of late mRNAs are heterogeneous in length (Moss *et al*., 1991).

1.1.4. Poxvirus Promoters

The temporal regulation of the various gene classes is orchestrated by their promoters and the availability of transcription factors specific to each temporal class. Similar to the genes they are associated with, promoters are classified as early, intermediate and late, with early/late genes containing elements of both early and late promoters in the upstream region (Assarsson *et al*., 2008). Promoters tend to extend approximately 30 nts upstream of the transcription initiation site and substantial similarities can be found among promoters of the same temporal class across members of different poxvirus genera (Fick and Viljoen, 1999). On the basis of single nucleotide substitution studies, models of the optimal promoters have been established as follows:

The early promoter is divided into three regions relative to the mRNA start site at +1:

- 15 nt A-rich critical region (-13 to -28) in which substitutions have a major effect
- 11 nt of less critical T-rich sequences
- 7 nt region within which initiation occurs at a purine.

The critical region specifies the distance to the downstream transcription initiation site, not unlike the TATA box of higher eukaryotic RNA polymerase II promoters. Additionally, a strong promoter requires a G residue at -21, T residues at -22 or -23 , and A residues that are critical at some positions and optimal at others within the critical region (Moss *et al*., 1991). The transcription initiation site of early genes is known to be within 10 nts upstream of the translation initiation codon (Coupar, Boyle and Both, 1987).

The late promoter also consists of three regions:

- an essential upstream region of ~20 nts with consecutive T or A residues, in which runs of T residues have a greater activating effect
- 6 nt separator region
- a highly conserved TAAAT element on the coding strand within which transcription initiates, with a G or A residue immediately downstream of TAAAT in strong promoters. The majority of late promoters overlap with the translation initiation codon for the late protein as a result of this TAAAT sequence (Davison and Moss, 1989)

Mutations within the A triplet of the highly conserved TAAAT element have been shown to dramatically decrease transcription, while substitution in the flanking T residues also had a negative effect on transcription but to a varying degree, depending on the upstream sequence (Moss *et al*., 1991).

Intermediate promoters are quite similar to late promoters and are therefore often hard to discern from the latter by DNA sequence composition alone. Poxvirus genomes only have at most five known intermediate genes, making a consensus even more difficult to support. Nonetheless, the generally accepted model of the intermediate promoter consists of:

- 13 nt core element (-26 to -13)
- linker region of ~12 nts, the length of which is crucial, rather than the sequence
- 4 nt initiator element (-1 to +3) that takes the form of TAAA and within which initiation occurs (Baldick, Keck and Moss, 1992)

Given the very tight packing of ORFs in poxvirus genomes, it is not surprising that promoter sequences of divergent transcription units sometimes overlap giving the appearance of bidirectional promoters. The overlap of the critical and upstream regions of early and late promoters in the short (~50 nts) non-coding region between two adjacent genes is variable which can make deciphering the conserved regions difficult (Fick and Viljoen, 1999)

It should be noted that most natural promoters do not have optimal residues in all positions, creating a degree of variability in promoter strength, which is the primary basis for regulating gene expression (Moss *et al*., 1991).

1.2. Introduction to comparative genomics

The nature of this study falls under the realm of comparative genomics, which is the study of the functions of various parts of the genome - such as genes and regulatory regions - by comparing the genomes of different species. A completely sequenced genome does not reveal how the genetic information it contains gets translated into observable traits(Hardison, 2003). Functional regions of genomes must be identified and characterized in order to gain better insights into how these observable traits came to be. Comparative genomics is one way of approaching functional characterization of genes and regulatory regions.

One of the fundamental principles of molecular evolution is that extensive sequence similarity implies conserved function, and the common features of two organisms will be encoded in parts of their DNA that have been conserved since their divergence from a common ancestor (Hardison, 2003). The theory of comparative genomics therefore is based on the assumption that sequence conservation exposes functionally important regions. Furthermore, if a satisfactory degree of similarity can be found between an uncharacterized sequence and a sequence of known function, inferences can be made regarding the function of the uncharacterized sequence, and these can then serve as a platform to base subsequent experiments investigation into the unknown function. With the onset of available bioinformatics software, a recent instance of the application of comparative genomics has been the functional characterization and structure prediction of the G8R protein, a proliferating cell nuclear antigen (PCNA)-like protein in poxviruses. This protein was characterized through sequence-level analysis

and comparison to human and yeast PCNA proteins, all of which contain a sliding clamp-like motif that is also present in the G8R protein (Da Silva and Upton, 2009).

This scheme does not apply solely to coding sequences; regions of non-coding DNA that display particularly high degrees of conservation are regarded as good candidates for regulatory regions (Hardison, 2003). This point is illustrated by the discovery of the Conserved Sequence Element (CSE) in 2003 during the genome sequencing of the Yaba Monkey Tumor Virus, a member of the Yatapoxvirus genus(Brunetti *et al*., 2003). While sequencing the genome, a 42 nt sequence was identified that seemed unusually well conserved; unusual in both its length and the fact that it was almost perfectly conserved between members of four different poxvirus genera.

Although subsequent experiments on the CSE ultimately led to its classification as a promoter element in poxviruses (Eaton, Metcalf and Brunetti, 2008), the CSE is much more complex than other characterized poxvirus promoters. It appears upstream of the YMTV 23.5L gene, a homolog of the VACV gene F8L and the MYXV gene m018L, both of which are driven by early promoters. In VACV, the region upstream of the F8L gene contains both an early and a late promoter, suggesting that the gene driven by the CSE might be an early/late gene (Eaton, Metcalf and Brunetti, 2008). The CSE is deemed unusual primarily because even for a promoter it is remarkably well conserved. Furthermore, it is longer than the average poxvirus promoter and it is unclear which parts of it are required for promoter activity. Poxvirus promoters are normally in the range of ~30 nt, of which not all parts are conserved promoter elements, so the presence of a single promoter does not account for the high conservation observed over the full length of the

CSE. The discovery of the CSE therefore raises several questions; namely what other conserved functions it might have that would result in the high degree of conservation observed, and also whether the degree of conservation observed was in fact unusual at all, or if other regions of comparable length and conservation existed within poxvirus genomes.

1.3. Introduction to Java Pattern Finder

This project arose from the need for a way of identifying short highly conserved sequences, such as the CSE and any others like it. Classically, one way of searching genomes for short, conserved sequences would be to align whole genomes and look at the consensus sequence for highly conserved regions. The problem with this approach is that poxviruses are not completely collinear and genes often appear in a different order from genome to genome, making them hard to align. BLAST can search for sequence matches without needing to align the genomes, however BLAST requires a query sequence and cannot be used to identify unknown sequence matches *de novo.*

The Longest Common Subsequence (LCS) program was a program designed in 2006 by Marina Barsky at the University of Victoria that identifies unknown sequence matches in given sequences (Barsky *et al*., 2006). This algorithm would search for and identify all perfectly matched sequences of a user-specified length that appear in every genome of a user-specified set of genomes. The drawback to this approach is that near-perfectly conserved sequences in biology are also important in investigating conserved functions and the LCS program fails to identify highly conserved sequences that contain a small number of positions that differ

In the next incarnation of the program, named Java Pattern Finder (JaPaFi), a feature was added enabling the program to identify recurring sequences that are *almost* perfectly conserved, or *approximate matches*. In JaPaFi, the user specifies the length of the approximate matches and the maximum number of allowed differences (insertions, deletions, point mutations). The program then identifies all sequences of the specified length that are within the specified edit distance, where *edit distance* refers to the number of operations (insertions, deletions, point mutations) required to transform one sequence to another and can be used interchangeably with *allowed number of differences* in the context of this project(Barsky, 2006).

1.4. Thesis rationale and objectives

The focus of this project was the application of the Java Pattern Finder program to a set of seven poxvirus genomes – the same genomes in which the CSE was identified – in order to identify other highly conserved sequences shared by them and then, using a variety of bioinformatic techniques, make inferences regarding the conserved functions of these sequences. In so doing, our goal was to be able to either support or refute the claim that the CSE is an unusually well conserved sequence depending on whether or not other sequences of comparable length and high degree of conservation were shared between these genomes, and if so, how many. Furthermore, our hope was that the functional characterization of these highly conserved sequences could further our understanding of how these viruses function.

2. Materials and Methods

2.1. The Java Pattern Finder Algorithm (JaPaFi)

JaPaFi is designed to discover relatively small (< 100 nt), highly conserved DNA sequences present in a set of large DNA sequences. It identifies approximate matches, where the term *approximate match* refers to the fact that the sequences there are a few positions that vary in the matches identified and thus they are not perfectly conserved. Rather, these sequences fall within a set edit distance of one another, where edit distance refers to the number of insertions, deletions or point mutations required to transform one sequence into another. An important feature of JaPaFi is that it is alignment independent - genomes need not be aligned in order to identify highly conserved regions - a feature which is useful for poxviruses in particular since aligning their genomes can be problematic, as explained in section 1.3.JaPaFi is designed to identify highly conserved sequences with one or more differences whereas the Longest Common Substring (LCS) program, available through the Viral Genome Organizer software at www.virology.ca, is better suited to identifying perfect matches (Barsky *et al*., 2006). Ultimately, the development of a graphical user interface that integrates both the LCS program and JaPaFi would be ideal for identifying patterns with zero or more differences.

The current version of JaPaFi allows users to select a set of genomes to search for all approximate matches, and then specify the length, *n*, and the maximum number of differences, *k*, allowed between these approximate matches (Barsky, 2006). It identifies approximate

matching sequences by first identifying all matching regions between the first two genomes. It then looks at each length *n* substring of these matching regions as a *pattern* and iterates through the other genomes, identifying every instance of each pattern that is within an edit distance of *k* from the pattern. Because the program iterates through every sequence, the order of the sequences should not affect the program's output, although it may affect the runtime. If a given pattern appears in all of the genomes, it is shown in the output. The raw output of the program is an enumerated list of all of the patterns identified, along with each instance of that pattern. The start positions of every instance of the pattern are shown in the output, along with genome in which it appeared, and its sequence as it appears in that genome.

All approximate matches identified in this project have been identified using JaPaFi, and all perfect matches have been identified using LCS. The set of 7 genomes used in these studies are shown below (Table 2-1).

Genus	Species	GenBank accession	Abbrevi-ation
Capripoxvirus	*Goatpox virus strain G20-LKV*	AY077836	GTPV
Capripoxvirus	*Lumpy skin disease virus strain Neethling 2490*	NC_003027	LSDV
Leporipoxvirus	*Myxoma virus strain Lausanne*	NC_001132	MYXV
Capripoxvirus	*Sheeppox virus strain A*	AY077833	SPPV
Suipoxvirus	*Swinepox virus strain Nebraska 17077-99*	NC_003389	SWPV
Yatapoxvirus	*Yaba-like disease virus strain Davis*	NC_005179	YLDV
Yatapoxvirus	*Yaba monkey tumor virus strain Amano*	NC_002632	YMTV

Table 2-1Genomes used in this study.

2.2. Identification and visualization of highly conserved regions

As outlined in section 2.1, the raw output of the program lists all instances of each pattern identified, which genome that instance appeared in, and the position in that genome. To see where these patterns fell relative to ORFs in the viral genomes they were visualized against an annotated genome map of the MYXV genome, which served as the model species throughout this project, using the Viral Genome Organizer (VGO) (Figure 2-1) (Upton *et al*., 2001). In these visualizations, the patterns appeared as coloured bands in *data tracks* above the genome (Upton *et al*., 2001). The raw JaPaFi output was converted into a VGO-readable format using an in-house script, although one feature of the current version of the JaPaFi GUI is that it converts the raw output to VGO-readable format automatically. VGO import format can be found at http://athena.bioc.uvic.ca/VGO_How_to.

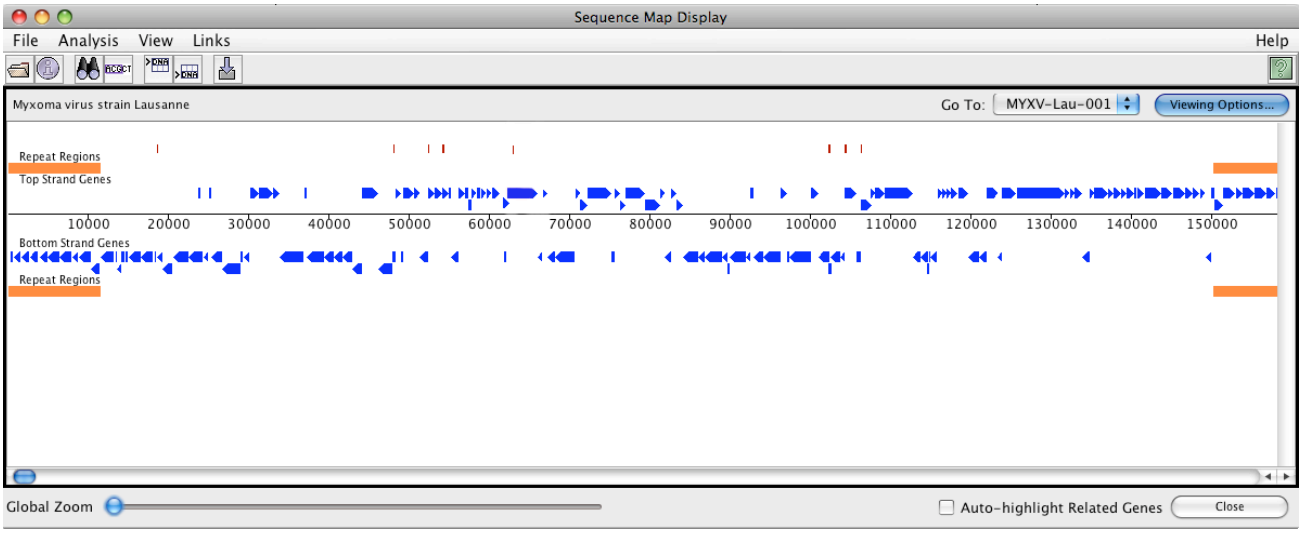


Figure 2-1 MYXV genome map with JaPaFi hits. Blue arrows are MYXV ORFs and red bars above are JaPaFi hits. Orange bars at the right and left extremities are inverted terminal repeat regions.

Upon visualizing the results, it was observed that the patterns identified by JaPaFi were forming clusters of overlapping sequences, thereby highlighting larger contiguous stretches of conservation. This is to be expected considering the algorithm identifies patterns of fixed length *n*. Highly conserved regions that exceed this length will therefore be identified by the program in overlapping length-*n* increments that are shifted over until the whole region is covered, as represented in the diagram below, provided each of these overlapping increments do not exceed the maximum allowed differences (Figure 2-2).

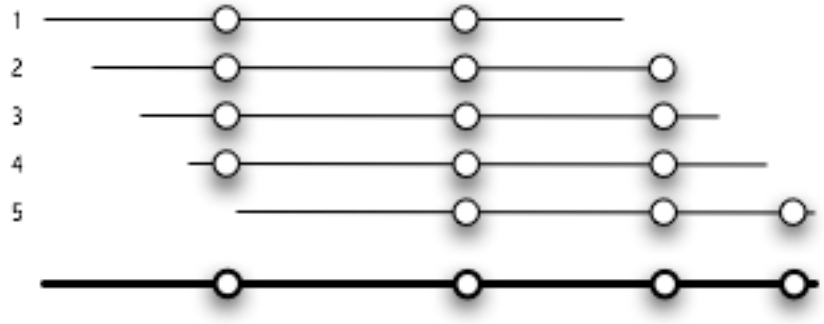


Figure 2-2 Fixed length patterns overlap to highlight longer regions of conservation

These contiguous conserved regions were labeled as "hits" and all subsequent analysis was conducted on these. By this scheme, the number of hits for a given parameter combination was actually less than the number of patterns in the program's raw output, since multiple patterns were combined to form the hits. Therefore, to determine the number of hits observed for a given parameter combination, the output was visualized in VGO where overlapping sequences show up as a single discrete band (hit), and counts were taken based on the number of discrete bands observed.

2.3. Logos

Logos provide useful visual representations of the sequence consensus over short regions in multiple sequence alignments. Essentially, they are histograms in which each bar is a stack of letters (A, T, C and G for a nucleotide sequence logo) representing a position in the sequence. The height of each letter in the stack is proportional to the frequency with which that letter appears at that position in the multiple sequences alignment (Figure 2-3).

Figure 2-3. Sample logo.

The WebLogo program, available at http://weblogo.threeplusone.com/create.cgi, was used to create logos of each of the selected hits (Crooks *et al*., 2004) .

2.4. Functional analysis

1.1.1. Known conserved amino acid sequences

The nucleotide sequences of hits that fell within coding regions were translated into amino acid sequences. The EMBOSS PATMAT motif tool, which compares query protein sequences against the PROSITE database of motifs, was then run on these amino acid sequences (Wallace and Henikoff, 1992). PATMAT was accessed through a web application available at

http://weblab.cbi.pku.edu.cn/program.inputForm.do?program=patmatmotifs(v5.0) which has since become unavailable for public use.

The amino acid sequences for the whole genes in which these hits appeared were queried against the UniProtKB and Swiss-Prot databases using the ScanProsite tool, available at http://ca.expasy.org/tools/scanprosite/(deCastro *et al*., 2006).

2.4.1. Identifying motifs within hits

The hits were searched using two different approaches to see if there were any common motifs that might give hints as to the conserved functions of the hits. For the purpose of this study, the term *motif* refers to short recurring sequences identified within hits. Motifs may include conserved promoter elements, i.e. part of a promoter. Motif is also used in the context of conserved protein domains and the Prosite database, which stores minimal protein motifs required to functionally characterize proteins. The term pattern refers specifically to a conserved sequence identified by JaPaFi.

In the first scheme, promoter elements were identified and marked within the hits according to the known conserved elements of poxvirus promoters corresponding to each temporal class as shown below, with transcription initiating at +1, which falls within the initiator site.

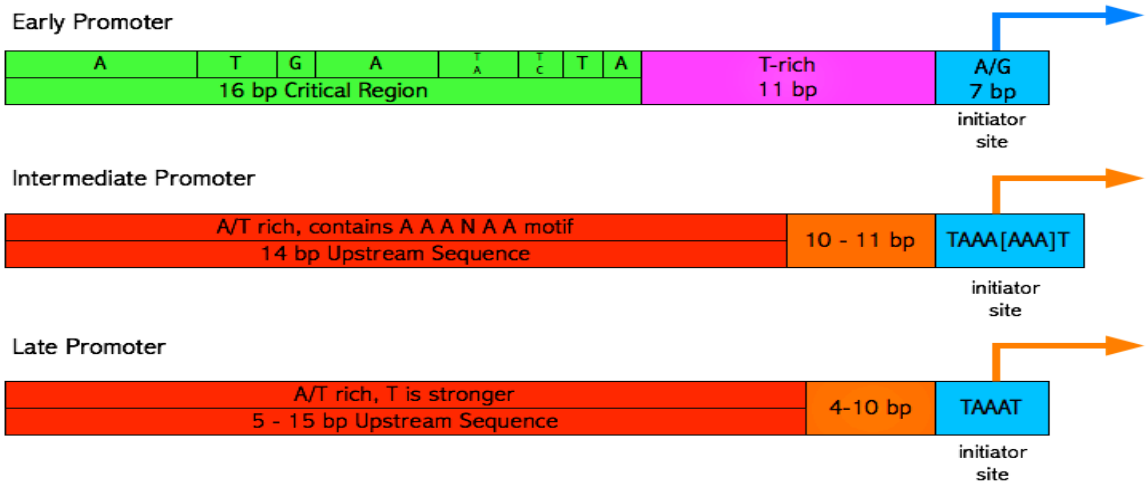


Figure 2-4 Known consensus of conserved poxvirus promoter elements

As a less targeted second approach to determining the functions of promoter and non-promoter hits alike, all hits were searched for smaller recurring motifs within them, in the 3 – 8 nt range. Motifs were identified using MEME/MAST motif finder, available at http://meme.nbcr.net/meme4_1_1/cgi-bin/meme.cgi, which is a web application that analyzes sequences for similarities among them and outputs a list of the motifs it discovers (Bailey *et al*., 2006). MEME 4.1.1 accepts as input a text file containing FASTA formatted sequences to search for motifs within (Bailey *et al*., 2006). Users can then specify an ideal distribution of motifs in the sequences submit, the width of the motifs and the maximum number of motifs to identify. For this study, the search was conducted specifying any number of repetitions of motifs within the sequences submitted, motif widths of 2-8 nts, and only the top 15 highest-scoring motifs were examined. The output displayed each motif identified in the form of a Logo based on every instance of said motif, and a diagram showing the location of these instances in each of the query sequences (Figure 2-5).

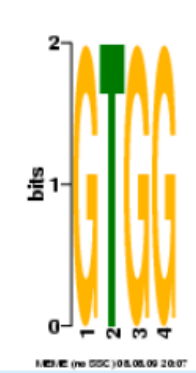
P N **MOTIF 4** width = 4 sites = 2 llr = 15 E-value = 1.6e+003

SEQUENCE LOGO

Information Content
8.0 (bits)

Relative Entropy
10.8 (bits)

Download LOGO
Without SSC:[EPS][PNG]
With SSC:[EPS][PNG]



NAME	STRAND	START	P-VALUE	SITES
09_GTPV.fasta	+	3	5.72e-04	TAA GTGG TTCATAATTA
08_GTPV.fasta	-	63	5.72e-04	TATTATTTAT GTGG TTTATATTCC

Motif 4 block diagrams

Name	Lowest p-value	Motifs
09_GTPV.fasta	5.72e-04	+4
08_GTPV.fasta	5.72e-04	-4
SCALE		1 25 50 75

Figure 2-5. MEME sample output.

3. Results

3.1. Genomes included in this study

The set of 7 genomes in which the CSE had been identified was selected in order to address the question of whether the CSE was in fact unusual in its size and degree of conservation or whether other comparable sequences were present within that set.

All seven of these genomes were from the poxvirus subfamily *Chordopoxvirinae,* which is one of two subfamilies in the poxvirus family and includes all poxviruses affecting vertebrate hosts. Any two genomes within this set of seven were between 56% - 98% identical based on full genome ClustalW alignments (Table 3-1). These were already known to contain at least one 42 nt highly conserved sequence among them – the CSE. At the time that the CSE was identified, during the sequencing and annotation of the Yaba monkey tumor virus genome, these seven were the only sequenced poxviruses in which the CSE was identified.

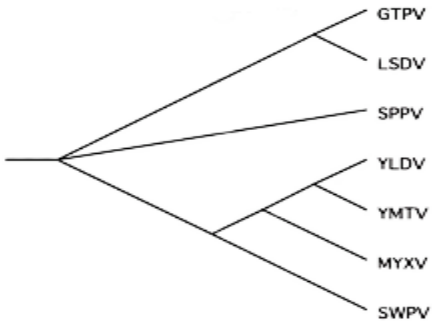


Figure 3-1 a cladogram that was made based on a ClustalW whole genome alignment of the seven.

% ID	GTPV	LSDV	SPPV	YLDV	YMTV	SWPV	MYXV
GTPV	-	97.93	97.06	66.55	65.05	66.44	57.79
LSDV	-	-	97.49	66.36	64.98	66.34	57.78
SPPV	-	-	-	66.59	65.12	66.5	57.75
YLDV	-	-	-	-	79.33	63.59	56.61
YMTV	-	-	-	-	-	62.62	57.39
SWPV	-	-	-	-	-	-	57.49
MYXV	-	-	-	-	-	-	-

Table 3-1 pairwise percent identity values for each pair of genomes (%).

Interestingly, VACV does not contain a close match to the CSE, as revealed by a search of the VACV genome for an approximate match, despite the fact that VACV contains homologs of the two genes between which the CSE appears in these 7 genomes.

3.2. Counting the number of hits for different values of length and edit distance

As outlined in section 2.2, JaPaFi was run on the set of seven genomes for a number of different parameter combinations in order to observe the effects of altering length and allowed differences on the number of hits. JaPaFi's output was visualized against a genome map of the MYXV genome. Overlapping patterns appeared in the visualization as a single band and were

regarded as a single contiguous hit, and hit counts were taken based on visualizations against the MYXV genome.

Hit counts were recorded in a matrix with length (*n*) on the vertical and allowed differences (*k*) on the horizontal (Table 3-2). As explained in section 2.1, perfectly matching hits (0 differences) were identified using the Longest Common Substring program, available through the Viral Genome Organizer software at www.virology.ca, which was designed to identify perfect matches while JaPaFiwas designed to identify approximate matches (Barsky, 2006).

n* \ *k	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
15	16	303						
16	12	115						
17	11	57						
18	10	31	417					
19	9	27	189					
20	6	21	117					
21	5	15	70	423				
22	4	15	55	250				
23	3	13	47	177				
24	2	11	28	111				
25	2	11	25	98				
26	1	10	22	83				
27	1	8	15	50	148	464		
28	1	7	15	45	130	358		
29	1	5	13	37		284		
30	1	4	9	24	76	188		
31	1	4	6	24	65			
32	1	3	6	20	60	148		
33	0	3	5	14	34			
34	0	3	5	12	30	93		
35	0	3	4	10	27		184	
36	0	3	4	9	22	61		
37	0	3	4	8	19		115	
38	0	3	4	8	14	43		
39	0	2	4	4	11		80	
40	0	2	4	3	10	28		
41	0	2	3	3	9	26	*	
42	0	1	3	3	6	16	47	
43	0	1	3	3	6	14	38	
44	0	1	3	3	6	12	35	

45	0	1	2	3	4	6	26	
46	0	1	2	3	3	5	25	
47	0	1	2	3	3	5	23	
48	0	1	2	3	3	5	18	
49	0	1	2	3	3	5	14	
50	0	1	2	3	3	4	12	
51	0	0	2	3	3	3	5	
52	0	0	2	3	3	3	5	18
53	0	0	1	3	3	3	5	18
54	0	0	1	3	3	3	4	11
55	0	0	1	2	3	3	3	9
56	0	0	1	1	2	3	3	8
57	0	0	1	1	2	3	3	5
58	0	0	0	1	2	3	3	5
59	0	0	0	1	1	3	3	5
60	0	0	0	1	1	3	3	5
61	0	0	0	0	1	2	3	3
62	0	0	0	0	1	2	3	3
63	0	0	0	0	1	2	3	3
64	0	0	0	0	1	2	3	3
65	0	0	0	0	1	2	3	3
66	0	0	0	0	1	2	3	3
67	0	0	0	0	0	2	3	3
68	0	0	0	0	0	2	3	3
69	0	0	0	0	0	2	2	3
70	0	0	0	0	0	2	2	3

Table 3-2 Hit counts for varying lengths and allowed differences, as observed by running JaPaFiand Longest Common Substring on a set of genomes consisting of GTPV, LSDV, MYXV, SPPV, SWPV, YLDV and YMTV.

These hit counts were later verified by converting JaPaFi's raw output to spreadsheets of all start and stop positions of matches.

One of the trends observed in the hit count matrix was that hit counts increased across each row as the allowed differences are increased for a constant length. This is to be expected since increasing the allowed differences loosens the search parameters, making it more likely that sequences be found that satisfy the parameters. By a similar principle, hit numbers decreased down each column as the size was increased with differences held constant. This is to be expected since longer pattern matches are less likely to be found than shorter ones. These

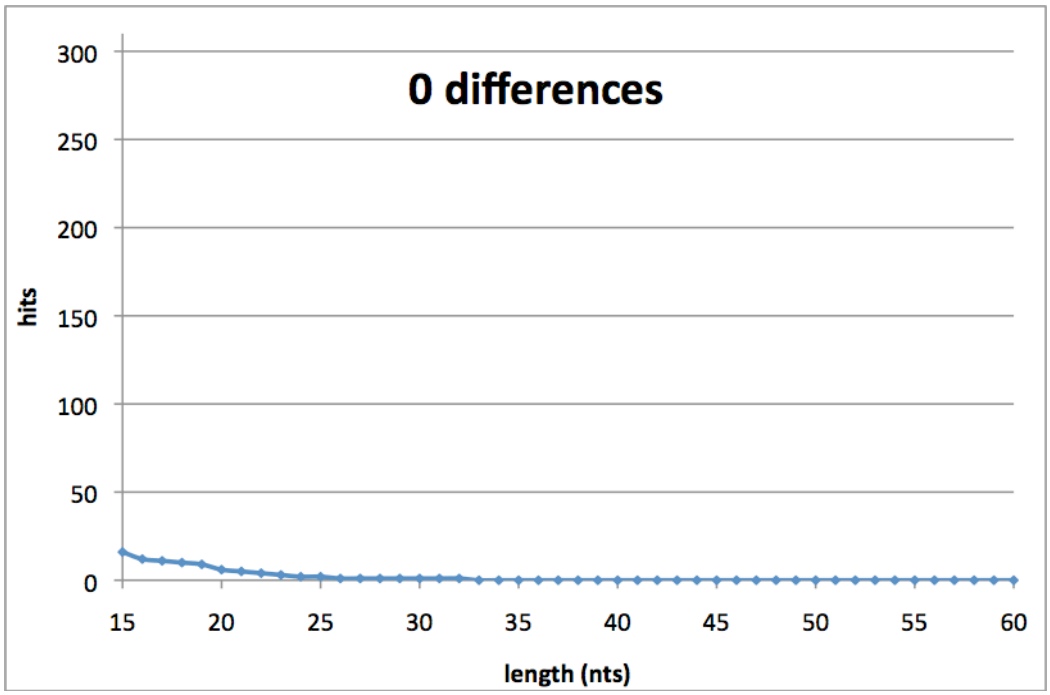
observations support the hypothesis that some parameter combinations – those corresponding to longer hits with fewer differences – identify only a few hits, making these unusually conserved.

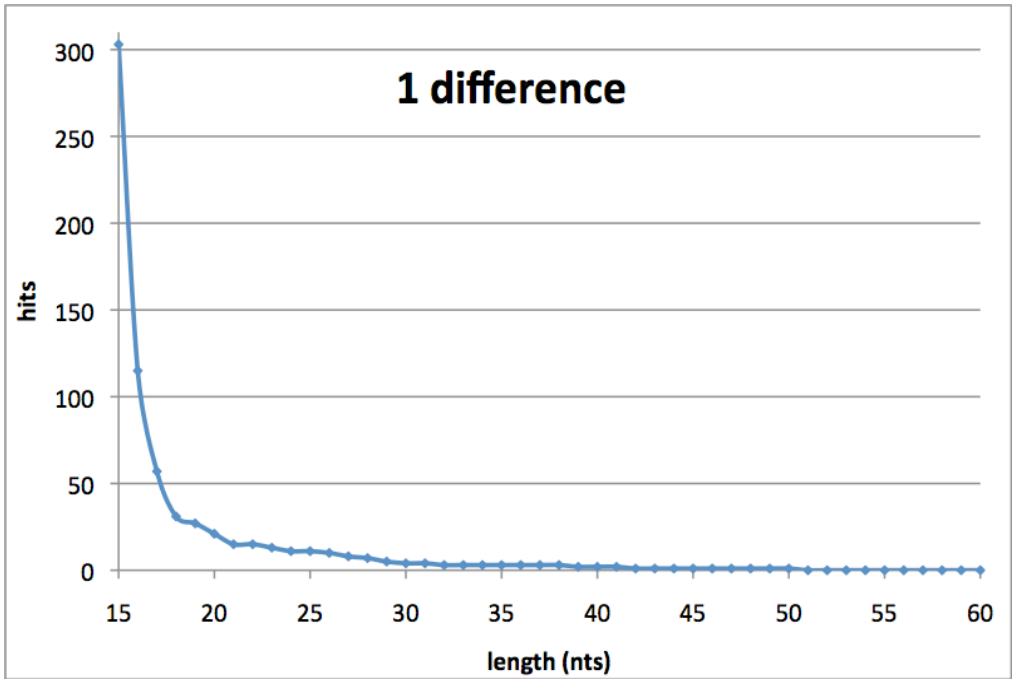
3.3. Signal-to-noise

One question that was raised when altering JaPaFi's search parameters was how to distinguish between signal - which refers to a conserved sequence with a conserved function – and noise – which refers to a pattern that is shared between several genomes as a result of chance and does not reflect a conserved function. From a statistical perspective, this becomes a question of "what are the odds that a particular sequence appearing in one genome also appear in all of the others", and furthermore, given such a sequence, what are the odds that this conserved sequence have a conserved function in all of these genomes.

One feature of conserved sequences that can act as an indicator of signal vs. noise is the location of the hits relative to orthologous genes in the different genomes; hits that fall near or within orthologous genes in every genome support the hypothesis that these sequences have a conserved function. Especially given the high degree of conservation of poxvirus core genes, conserved sequences that reflect a conserved function would be expected to appear in the same position relative to orthologous genes in the various genomes, whereas sequences that are not conserved but rather fit the search criteria for matches by chance would not.

Plotting the number of hits as a function of length illustrates the rapid decrease in the number of hits as parameters are made more stringent. Presumably, this rapid decrease in total hits also marks a decrease in the amount of noise in the results by virtue of the fact that it is less likely that a longer pattern match with fewer differences be identified by chance, as oppose to a short pattern match with numerous differences, however, without a way of distinguishing signal from noise in the hits, this is merely speculation (Figure 3-2).


0 differences
300
250
200
150
100
50
0
hits
15
20
25
30
35
40
45
50
55
60
length (nts)


1 difference
300
250
200
150
100
50
0
hits
15
20
25
30
35
40
45
50
55
60
length (nts)

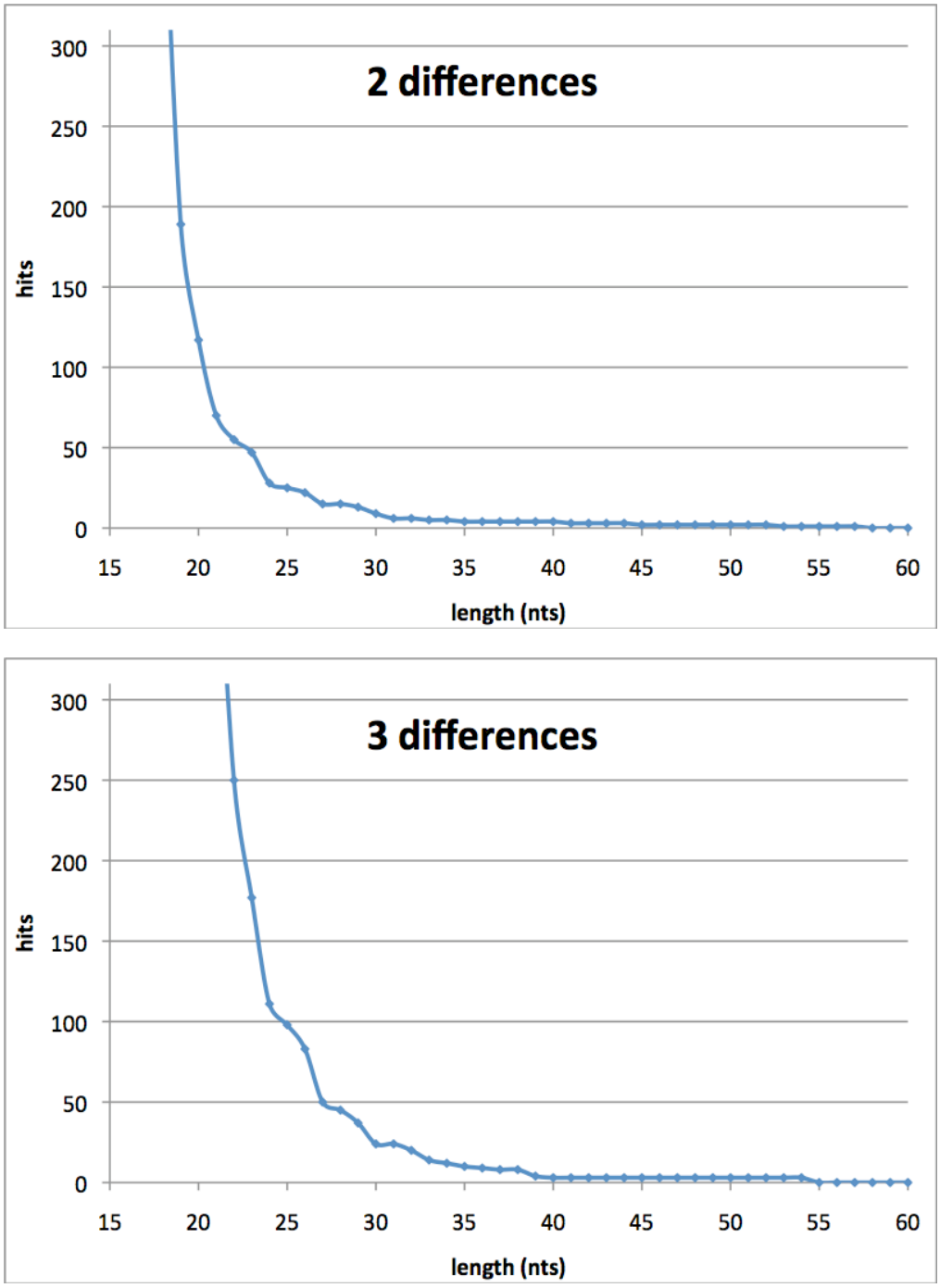


Figure 3-2. Hit counts as a function of length with of a) 0, b) 1, c) 2, d) 3 differences

The rapid decrease in the number of hits occurring as length is increased from 15 to 25 in the 1, 2 and 3 differences figures reflect this speculated increase in noise.

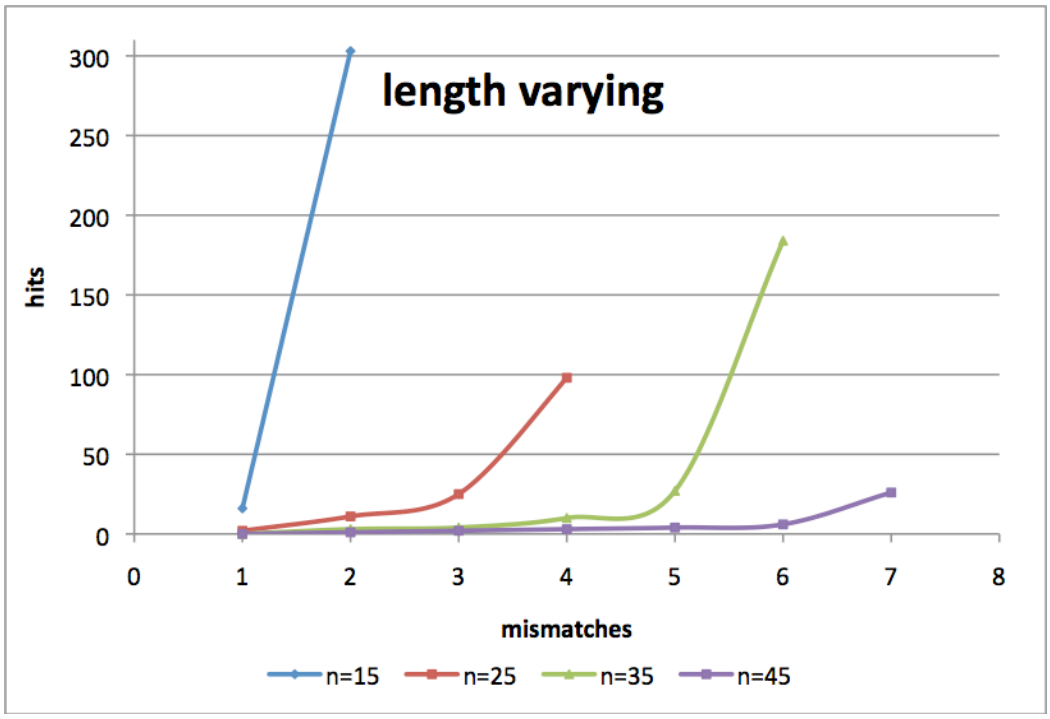


Figure 3-3. Hit counts as a function of differences, shown for 4 different lengths.

Similarly, increasing allowed differences results in a rapid increase in hit counts (Figure 3-3). Thus, although we have not determined strict cut-offs identifying where chance matches exceed weak signals, rapid increases in hit counts can provide rough hints as to where this occurs. A logarithm of the odds (LOD) score that estimates the likelihood that two loci lie near each other in the genome would be useful as a way of establishing these cutoffs.

Further to plotting hit counts, it was speculated that observable trends in the position of the hits within genomes could also aid in identifying signal vs. noise. It was observed upon visualizing a few result sets that hits appeared mostly in promoter regions when search parameters were stricter. To investigate this trend, selected results were visualized against genome maps and the positions of hits relative to ORFs were noted. Fractions of promoter hits

to total hits were recorded in a matrix with length on the vertical and differences on the horizontal (Table 3-3). In these counts, promoter hits were regarded as those that appear within 40 nts upstream of the start of genes. Given the tight packing of genes within poxvirus genomes, often these promoter hits also fell within the coding region of another adjacent gene. Non-promoter hits were those that appeared only in coding regions or in non-coding regions but beyond the 40 nts upstream of gene start sites.

n, k	1	2	3	4	5	6	7
25	1.00		0.21				
26	1.00						
27	1.00	0.87	0.34				
28	1.00	0.87					
29	1.00	0.85					
30	1.00	1.00	0.58	0.34			
31	1.00	1.00					
32	1.00	1.00	0.65				
33	1.00	1.00	0.71	0.44			
34		1.00	0.75	0.47			
35		1.00	0.80	0.48			
36	1.00	1.00	0.89	0.59			
37	1.00	0.88	0.63				
38	1.00	0.88	0.88				
39		1.00	0.73				
40	1.00	1.00	0.70	0.43			
41		1.00	0.78	0.46			
42			1.00	0.63			
43			1.00	0.57			
44				0.67			
45	1.00	1.00	1.00	1.00	1.00	0.46	
46						1.00	0.40
47					1.00	1.00	0.48
48						1.00	0.61
49		1.00		1.00	0.57		
50	1.00	1.00	1.00	1.00	1.00	0.58	
51						1.00	
52				1.00	1.00	1.00	1.00	0.50
53							1.00	0.44
54							1.00	0.73

55	0.00	1.00	1.00	1.00	1.00	1.00	0.67
56							0.63
57							1.00
58							1.00
59							
60					1.00		1.00

Table 3-3 Fractions of promoter hits to total hits for varied parameter combinations.

From this matrix it was observed that hits consistently appeared in promoter regions for stricter parameter combinations, while decreasing length and increasing differences both resulted in the introduction of non-promoter hits (as defined in this thesis) and the subsequent increase in the proportion of hits falling in non-promoter regions. The shaded cells in Table 3-3 show hits going from being all promoter hits to some non-promoter hits.These observations support the hypothesis that the longest hits with the least differences are promoters. This hypothesis is explored further in the remainder of this chapter.

These trends, although interesting, do not resolve the issue of signal vs. noise. Ideally, a statistical measure gauging whether a conserved pattern of a particular length and number of differences can be considered evidence for a conserved function would give some sense of the significance of the hits identified by the program, in a manner similar to the Expect (E) values calculated by the BLAST program. E-values values indicate the probability due to chance that there is another alignment in the database queried with a similarity score greater than that of the query sequence with a particular match sequence match. We had no other techniques at our disposal for distinguishing weak signals from noise in the *de novo* discovery of conserved sequences without carrying out a full functional analysis on every hit identified. This statistical measure is discussed further in the discussion of future directions (section 4.2.2).

3.4. Selecting a set of hits for functional analysis

Manually conducting the functional analysis was a multi-step process that limited the number of hits that could be manageably analyzed to 10-12 of the top hits identified – that is, the longest and most highly conserved. In order to determine which hits were the top hits, a way of ranking the hits was required.

One method of doing so would have been to use the length and number of differences used to identify the hits; selecting those with the greatest lengths and least allowed differences. However, because the hits were made up of overlapping patterns and it was these patterns, rather than the whole hit, that satisfied the search criteria (as explained in section 2.2 and Figure 2-2), the number of differences present in the hits could not be easily determined.

The top hits are the few that lie in the tail end of the graphs in Figure 3-2, where length is highest and number of hits is at a minimum. Intuitively, the hits identified in the 0 differences result set might be considered the most conserved as these are, in fact, perfectly conserved. However, JaPaFi was written in order to identify approximate matches, and it was often observed upon comparing start and stop positions of the hits that a conserved sequence identified in the 0 differences set was a substring of a sequences identified in a 1 or more differences set (Figure 3-4). Allowing a small number of differences therefore identified longer conserved sequences that were still within a small edit distance of one another.

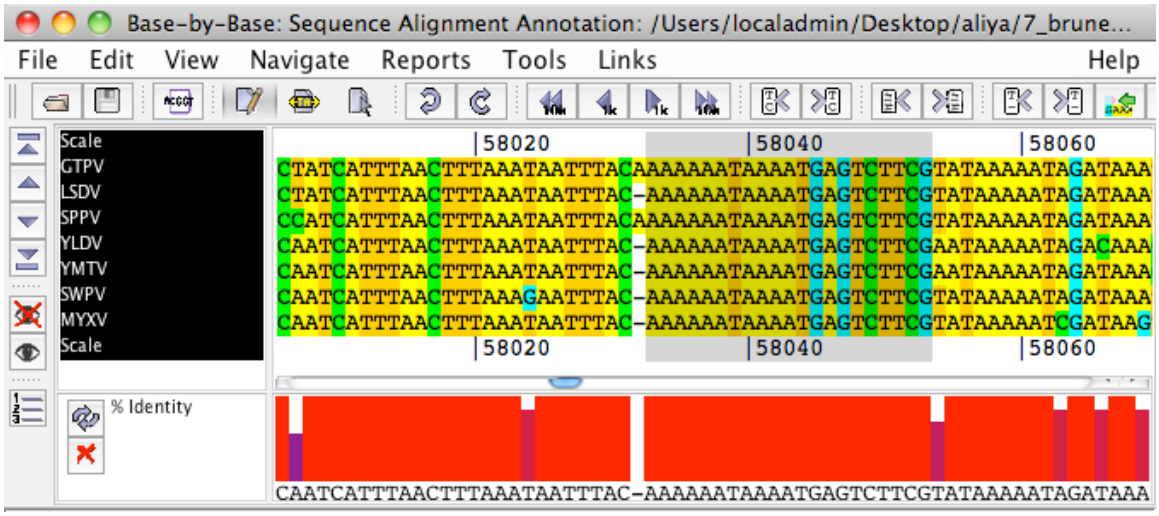


Figure 3-4 Alignment of Brunetti's 7 genomes. This window shows the alignment from 52344 - 52407 of the MYXV genome, which is one of the top hits identified with 2 differences. This region corresponds to 58006 – 58072 in an alignment of all 7 genomes (marked positions in the figure). The highlighted region (MYXV 52370 - 52390) is one of the most conserved hits identified with 0 differences. Red and purple bars on the bottom of the window show the percent identity at each position of the alignment.

As is apparent from the percent identity histogram (Figure 3-4), the whole sequence spanning from 52344 – 52497 in MYXV is highly conserved, despite a small number of differences. It was therefore decided that the final short list of hits would be determined by surveying the hits identified with 0, 1 and 2 differences and, in cases like that shown above where the 0 difference hit is a substring of the 2 difference hit, the longer sequence will be included in the shortlist.

The survey of the top hits was performed by examining the hit-count matrix (Table 3-2). Parameter combinations that only yielded one hit were regarded as the strictest parameter combinations. These parameter combinations included (32, 0), (50, 1), (57, 2) and (60, 3) for 0 to 3 differences. The single hits identified in these sets were dubbed the top hits. In a similar manner, slightly less stringent search parameters revealed 2 hits, the top hit, as identified previously, and the second topmost hit. The parameter combinations in which these second

topmost hits were identified for 0 – 3 differences were (25, 0), (41, 1), (52, 2) and (55, 3). The third, fourth and fifth topmost hits were identified, and so on.

In forming a final set of hits based on the top 5 most conserved hits for 0, 1 and 2 differences, it was observed that the boundaries of these hits varied slightly across the 0, 1 and 2 difference sets, as well as the order in which they were ranked from the most conserved to the least conserved among the top 5.

0 mismatch Top hits

start	stop	length
104295	104326	31
47958	47982	24
100096	100118	22
102559	102580	21
52370	52390	20

1 Mismatch Top hits

start	stop	length
104276	104326	50
102146	102187	41
18531	18572	41
52334	52402	58
106252	106281	29

2 Mismatch Top hits

start	stop	length
52334	52407	63
104274	104330	56
102144	102191	47
18530	18572	42
54155	54190	35

Final set of hits

hit	start	stop	length
01	18531	18573	42
02	52335	52415	80
03	54146	54197	51
04	66601	66645	44
05	80165	80213	48
06	68525	68572	47
07	100085	100138	53
08	102112	102203	91
09	102281	102321	40
10	104268	104341	73
11	106243	106299	56

Figure 3-5. Start and stop positions in MYXV and lengths of top 5 hits from a) 0, b) 1 and c) 2 differences searches, and d) final set of 11 hits.

The variation in the boundaries of the hits is to be expected as the number of allowed differences is varied since the boundaries depend on both the number of allowed differences and their distribution. For instance, in Figure 3-4, the 0 differences hit is bounded by a difference to either side; a gap to the left and a mismatch to the right. If the former had occurred further to the left or the latter had occurred further to the right, this 0 differences hit would have been longer (Figure 3-6).

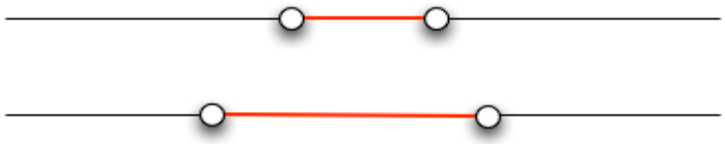


Figure 3-6 Diagram demonstrating how the distribution of differences affects the boundaries of the hit. Black circles represent differences in the sequence (black line). The hit is shown in red.

Similarly, the ranks of the most conserved hits within the top 5 also vary depending on the distribution of the differences within the hits (Figure 3-7). For 1 allowed difference, the sequence on the bottom gives a longer hit and would therefore rank higher. For 2 allowed differences, however, the sequence on the top gives a longer hit and would thereby rank higher while for 3 allowed differences the two hits are the same length.

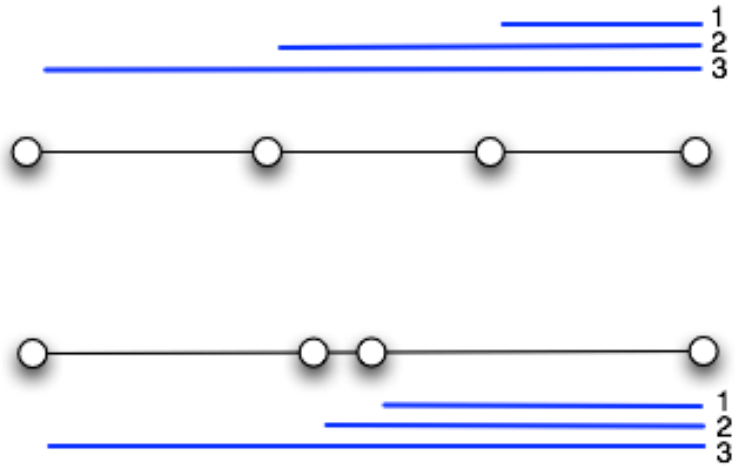


Figure 3-7 Diagram demonstrating how the distribution of differences affects the rank of a hit as the number of differences varies.

Thus by surveying the most conserved hits for a few different parameter values, a shortlist of hits for further analysis was established, despite the fact that the conservation of these hits could not be directly quantified by a numerical measure. A scoring method was established later to quantify the conservation observed in these hits. This method is discussed in detail in section 3.6.

3.5. Description of hits

Sequence logos were made for each of the hits by aligning the hit region from all 7 genomes using WebLogo (Crooks *et al*., 2004) as outlined in section 2.3. The following figures (Figure 3-8 to Figure 3-18) are logos of each of the hits (with a few flanking nucleotides included to either side). Diagrams showing the position of each hit relative to adjacent or overlapping

genes are also shown. These diagrams were obtained directly from visualizations of the hits against the MYXV genome using the Viral Genome Organizer (VGO) software, available at www.virology.ca(Upton *et al*., 2001). Genes are labeled according to the VBRC's ortholog group labeling scheme which takes the form of MYXV-Lau-###, with the numbering starting at the 5' end of the genome and increasing towards the 3' end. Additions are numbered as fractions (a new gene between 015 and 016 would be 015.5). GenBank names for these ORFs are also provided. These are italicized in the description following each figure and take the form of m###(L/R) where ### refers to the assigned gene number and R or L are assigned depending on whether the gene is a top strand gene that gets transcribed towards the right or a bottom strand gene that gets transcribed towards the left, respectively. Right and left are with respect to a set of reference genes in the central region of the linear poxvirus genomes.

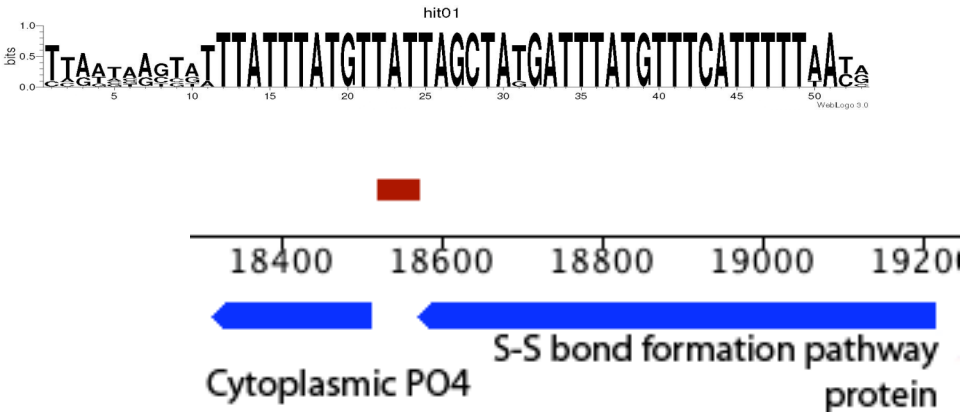


Figure 3-8. Logo and diagrammatic representation of hit 01.

Hit 01 is the conserved sequence element identified by Brunetti et. al. The CSE was identified at approximately 14 700 nts from the left hand side of the YMTV genome, which corresponds to 18531 in the MYXV genome. This hit lies mostly in the non-coding region

between, two bottom strand genes, Cytoplasmic PO4 protein (GenBank name *m018L*) and S-S bond formation pathway protein (GenBank name *m019L*), save for its rightmost end which overlaps with the tail end of the S-S bond formation pathway protein's coding sequence by a short stretch of 4 nucleotides. In the MYXV genome, hit 01 spans from 18521 to 18573, giving it a length of 53 nucleotides. Although it does not contain within it the translation start sites of either of the adjacent genes, it is within close enough proximity of the Cytoplasmic Protein start site, which occurs within 10 nts of the start of the hit, that some promoter elements are detectable. These will be identified and explained in more detail in section 3.7.3.

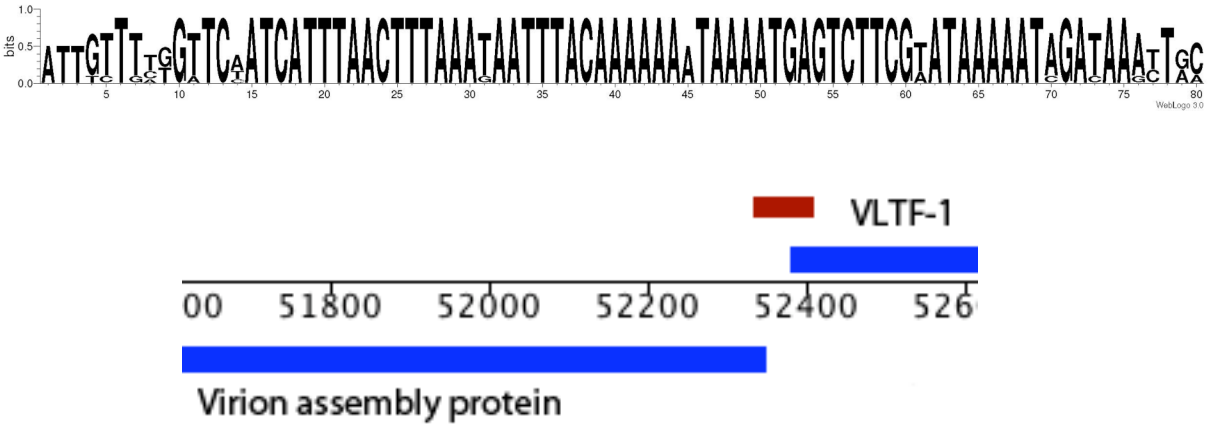


Figure 3-9. Logo and diagrammatic representation of hit 02.

Hit 02 spans the non-coding sequence between the top strand Viral Late Transcription Factor gene, VLTF-1 (*m053R*) and the bottom strand Virion Assembly Protein (*m052L*) gene. This non-coding sequence may act as a bi-directional promoter, containing promoter sequences for both of these opposite-strand genes within the sequence separating the translation start sites for the two genes. Hit 02 also overlaps with the coding sequence for both the Virion Assembly

Protein gene and the VLTF-1 gene by 20 and 30 nts, respectively. In the MYXV genome, hit 02 spans from 52335 to 52415, giving it a length of 80 nts.

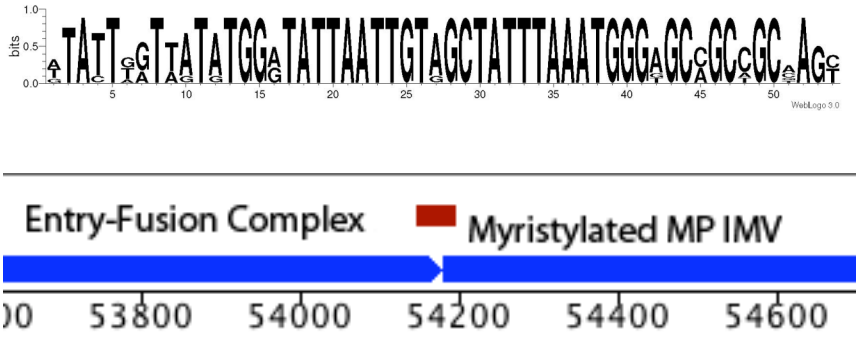


Figure 3-10. Logo and diagrammatic representation of hit 03.

Hit 03 overlaps with the Entry-Fusion Complex Protein (*m054R*) gene and the Myristylated Intracellular Mature Virus (IMV) Membrane Protein (MP) (*m055R*) gene, both of which are both top strand genes with the former ending at the translation start site for the latter. Between the two genes, hit 03 therefore overlaps with coding sequence over its entire length. It also contains the promoter for the Myristylated MP IMV gene. In the MYXV genome, hit 03 spans from 54146 to 54197, giving it a length of 53 nts.

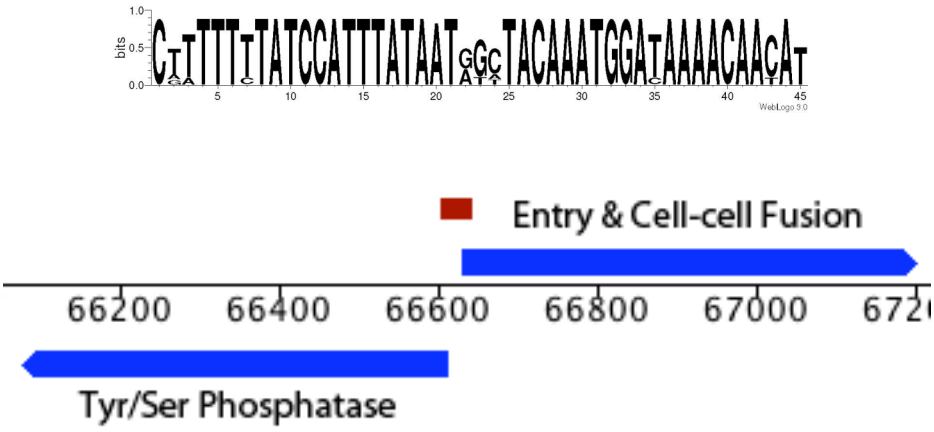


Figure 3-11. Logo and diagrammatic representation of hit 04.

Hit 04 is another instance of a putative bi-directional promoter, with the translation start sites of two non-overlapping opposite strand genes – Tyrosine/Serine Phosphatase (*m069L*) and Entry & Cell-Cell Fusion (*m070R*) – both being contained within the length of the hit. In the MYXV genome, hit 04 spans from 66601 to 66644, giving it a length of 43 nts.

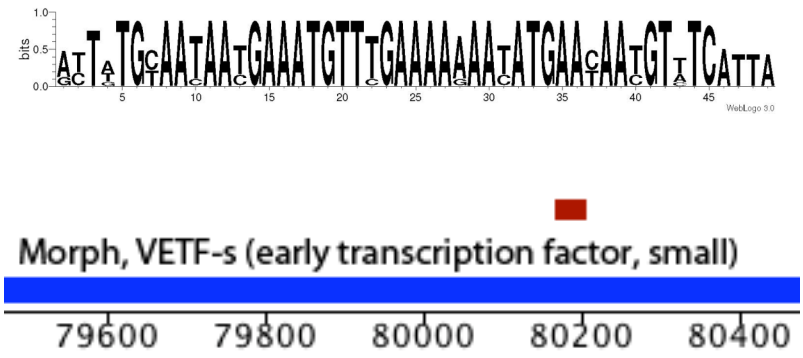


Figure 3-12. Logo and diagrammatic representation of hit 05.

Hit 05 is the first of the selected hits to be entirely contained within the coding sequence of a gene, in this case the top strand Morphogenesis Viral Early Transcription Factor Small Subunit gene (Morph, VETF-s) (*m081R*). It does not overlap with any other genes or predicted ORFs or their promoters. In MYXV it spans from 80165 to 80205, giving it a length of 40 nts.

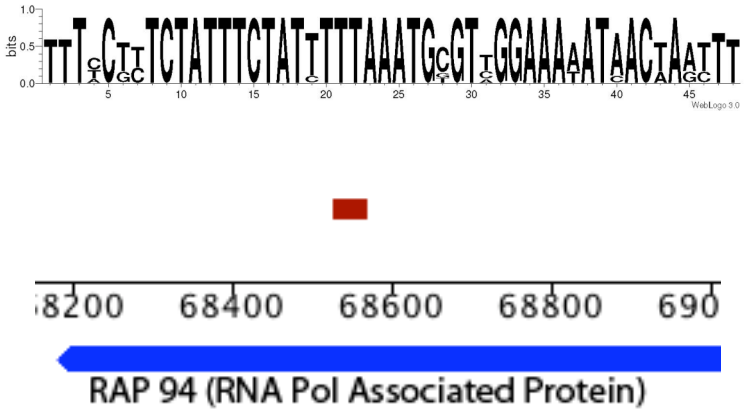


Figure 3-13. Logo and diagrammatic representation of hit 06.

Hit 06 is also entirely contained within the coding sequence of a gene, in this case the bottom strand RNA Polymerase Associated Protein (RAP94) (*m072L*). In the MYXV genome, hit 05 spans from 69528 to 68571, giving it a length of 43 nts.

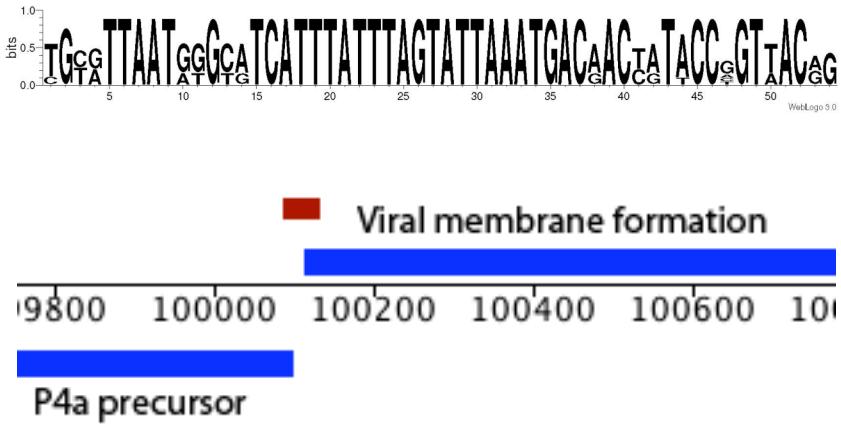


Figure 3-14. Logo and diagrammatic representation of hit 07.

Hit 07 is another bidirectional promoter, beginning just downstream of the translation start site of the bottom strand P4a Precursor (*m099L*) gene and extending through the non-coding sequence separating the P4a Precursor gene from the Viral Membrane Formation

(*m100R*) gene, and into the coding sequence of the latter. In MYXV it spans from 100085 to 100134, giving it a length of 49 nts.

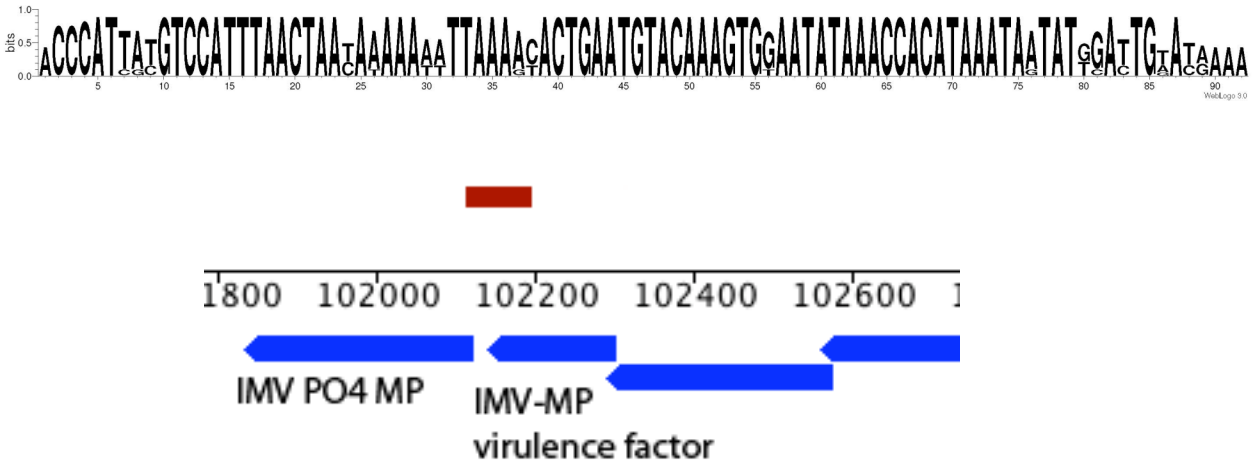


Figure 3-15. Logo and diagrammatic representation of hit 08.

Hit 08 contains the translation start site of the bottom strand IMV Phosphorylated Membrane Protein (IMV PO4 MP) (*m103L*) gene and extends into its coding sequence by about 10 nucleotides and extends through the non-coding sequence separating the IMV PO4 MP gene from the tail end of the IMV-MP/Virulence Factor (*m104L*) gene, another bottom strand gene. It therefore contains one promoter; that of the IMV PO4 MP gene. In MYXV it spans from 102112 to 102195, giving it a length of 83 nts.

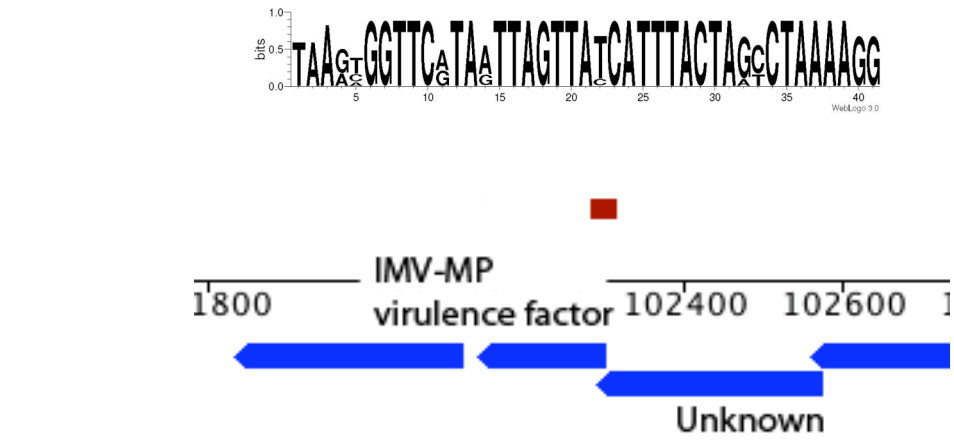


Figure 3-16. Logo and diagrammatic representation of hit 09.

Hit 09 contains the translation start site for the bottom strand IMV Membrane Protein /Virulence Factor (*m104L*) gene and overlaps with its coding sequence by about 20 nucleotides. It also overlaps with the Unknown Cop A15L (*m105L*) gene – another bottom strand gene that also overlaps part of the IMV-MP/Virulence Factor gene. In the MYXV genome, hit 09 spans from 102281 to 102317, giving it a length of 36 nts.

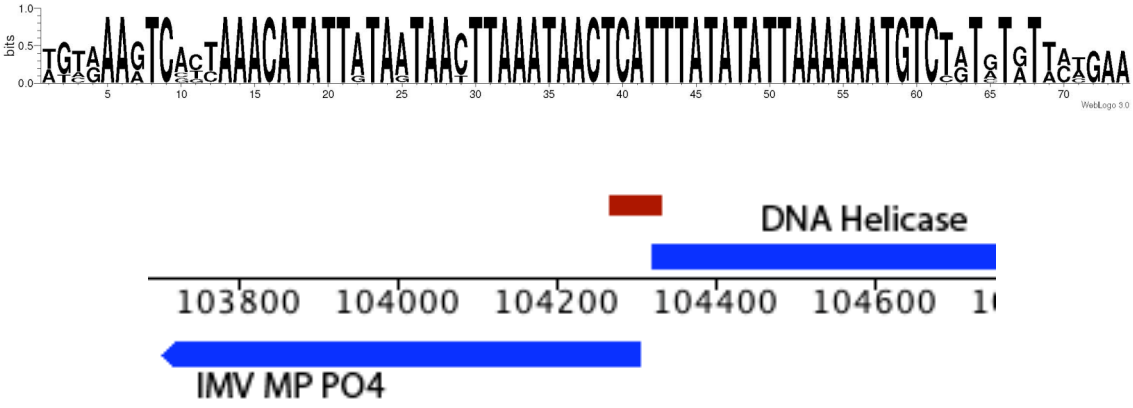


Figure 3-17. Logo and diagrammatic representation of hit 10.

Hit 10 is another bi-directional promoter, beginning downstream of the translation start site of the bottom strand IMV MP PO4 (*m107L*) gene (not to be confused with IMV PO4 MP, or *m103L*, which overlaps hit 08) and extending through the non-coding sequence separating this gene from the DNA Helicase Transcription (*m108R*) gene, and into the coding sequence of the latter. In MYXV, it spans from 104268 to 104334, giving it a length of 66 nts.

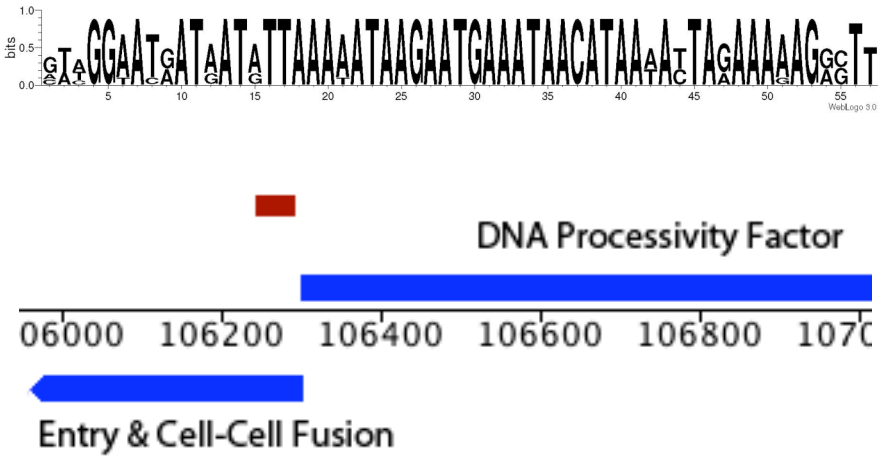
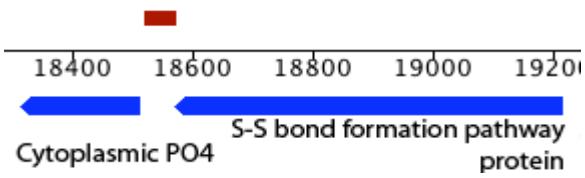
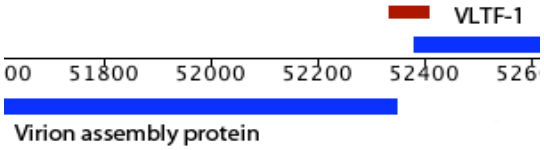
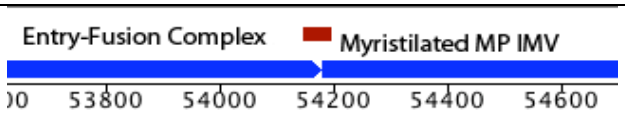
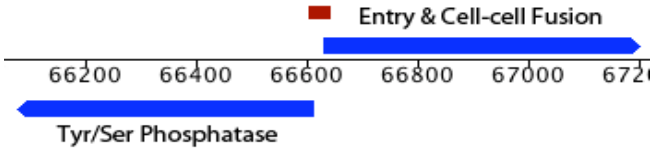
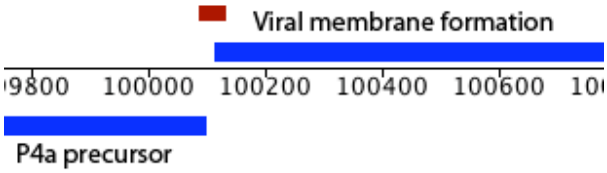
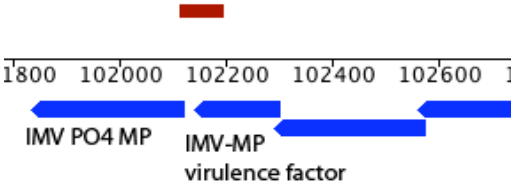


Figure 3-18. Logo and diagrammatic representation of hit 11.

Hit 11 does not contain the translation start sites of any genes and is entirely contained within the bottom stand Entry & Cell-Cell Fusion (*m110L*) gene, however, the top strand DNA Processivity Factor (*m111R*) gene's start site is within 10 nts downstream of the end of the hit so its promoter elements can be found in hit 11. In the MYXV genome, hit 11 spans from 106243 to 106294, giving it a length of 51 nts.

The set of 11 hits therefore consists of 2 hits from coding regions alone, and 9 hits that contain, between them, promoter elements corresponding to 13 different genes. These

promoter hits are summarized below (Table 3-4) along with the temporal classification of each of the genes in question, since the structure of the promoter elements are defined by the temporal class of the gene.

Hit	Gene	Strand	E/I/L	
1	Cytoplasmic Protein	-	E	
2	Virion Assembly Protein	-	L	
	Viral Late Transcription Factor 1	+	L	
3	Myristylated MP IMV	+	L	
4	Tyrosine/Serine Phosphatase	-	L	
	Entry & Cell-Cell Fusion	+	L	
7	P4a Precursor	-	L	
	Viral Membrane Formation	+	L	
8	IMV PO4 Membrane	-	L	

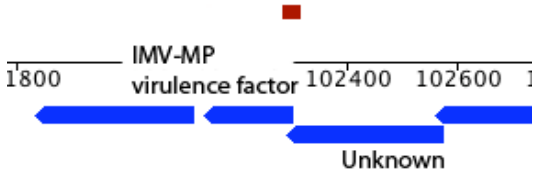
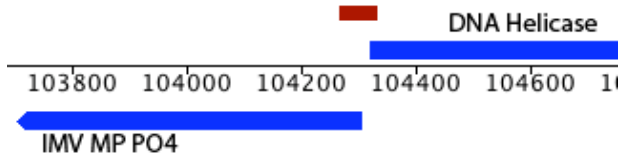
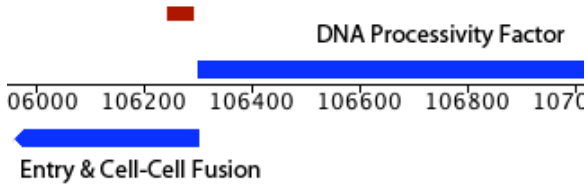
9	IMV Membrane Protein/Virulence Factor	-	L	
10	IMV Membrane Protein Phosphorylated DNA Helicase, Transcription	- +	L E/L	
11	DNA Processivity Factor	+	E	

Table 3-4 Summary of hits that contain promoters.

As is apparent from the diagrams shown inTable 3-4, many of these hits also overlap with an adjacent gene. In some cases they overlap with the 3' end of the gene, while in the cases of bidirectional promoters, they fall near the beginnings of the adjacent genes. This overlap is to be expected given the tight packing of genes within poxvirus genomes.

The identification of a sequence that contains the CSE among the hits determined to be most conserved lends strength to the claim that the CSE was, in fact, unusually conserved, while the observation that 9 out of 11 of the most highly conserved hits contain promoters supports the hypothesis that some poxvirus promoters maintain unusually high DNA identity. These promoters are examined in more detail in section 3.7.3.

3.6. Conservation scores

Having identified a set of hits to focus functional analysis on, the question still remained of whether or not the degree of conservation of these hits was unusual. In order to determine this, a scoring method was established to quantify the degree of conservation of the hits based on logos of the hits constructed using the Weblogo application, as outlined in section 2.3. In this scoring method, positions scores were obtained for each position in the hit based on the heights of the nucleotides appearing at that position in the logo.

To show that the hits were more conserved than would be expected, a set of sequences was selected to establish a baseline for the expected level of conservation. Known poxvirus promoters were selected to serve as controls since most of our hits contain promoters. Upstream sequences containing the promoters for 10 genes known to be highly conserved within the poxvirus family, shown below with their GenBank accession numbers in the MYXV genome (Table 3-5), were obtained for each of the 7 genomes using the VOCs database and aligned using ClustalW (Upton *et al*., 2003). Promoters are known to span the 30 nts upstream of the transcription initiation site which, in turn, is found within 10 nts upstream of the translation start site (Coupar, Boyle and Both, 1987). To ensure that the promoters were present in the sequences we used for these calculations, 100 nts upstream of the translation start sites for these genes were analyzed. These promoters were scored as described below, and scores were compared to those calculated for the hits.

Baseline1	**Ribonucleotide Reductase small subunit**	**m015L**
Baseline2	**DNA Polymerase**	**m034L**

Baseline3	**Unknown (Cop-G5R)**	**m049R**
Baseline4	**Complement Control/CD46/EEV**	**m144R**
Baseline5	**Holliday Junction Resolvase**	**m112R**
Baseline6	**IFN Resistance/eIF2 alpha-like PKR inhibitor**	**m156R**
Baseline7	**IFN Resistance/PKR inhibitor (Z-DNA binding)**	**m029L**
Baseline8	**IMV MP PO4 (Cop-A17L)**	**m107L**
Baseline9	**Thymidine Kinase**	**m061R**
Baseline10	**Uracil-DNA Glycosylase**	**m079R**

Table 3-5Promoters scored for comparison against conservation scores for hits. Upstream sequences were taken from the MYXV genome.

It is important to note that although many of the hits selected for functional analysis contain promoters, these promoters are not the same promoters as those selected as baseline sequences for calculating the expected conservation.

As mentioned in section 2.3, heights of the nucleotides were extracted directly from the Weblogo program using an in-house script, shown in Appendix B: In-house script for extracting character heights from Weblogo, which printed the heights of each letter at each position. Position scores were taken to be the height of the tallest (or most frequently appearing) nucleotide at a given position in the sequence divided by the number of different nucleotides appearing at that position. For instance, if the stack at a given position consisted of A (height 0.2), G (height 0.3) and T (height 0.8), the position score would be 0.8 / 3 or 0.267. Taking the height of only the most frequently appearing nucleotide into account in the score ensured that higher scores would be gained for more conserved positions since the greatest height possible is seen at positions that are perfectly conserved. Dividing by the number of nucleotides present at

a given position ensured that penalties would be given to positions where there is more variation. Thus, a position at which only one other nucleotide appears in addition to the most frequently occurring (tallest) nucleotide will score better than a position at which two different nucleotides appear (with lower frequencies) in addition to the tallest nucleotide, given that the height of the tallest nucleotide at both positions is the same

The *total information* was taken as the sum of all of the position scores along the full length of the hit (start and stop positions for the hits are provided in section 3.6), thereby awarding higher scores to longer hits, even more so if there is high conservation for much of the length. By this scheme, a shorter but very highly conserved sequence could score as highly as a longer, less conserved sequence.

Average information was taken to be the *total information* divided by the length of the sequences, thereby normalizing by size and awarding higher scores to more conserved hits.

Because the hits and the baseline sequences varied quite a bit in size, with the former in the range of 35-90 nts and the latter being 100 nts upstream of the start site, scores were also calculated for a fixed-length portion of each hit and upstream region. This ensured that the scores for the baseline sequences would not be diluted by the length they span, since the remainder of the upstream regions, less the promoters, could potentially reduce the overall conservation as these are not necessarily well conserved. 41 nts was selected as the fixed length since that was the length of the shortest hit, and because 41 nts is bigger than poxvirus

promoters, which range from 30-33 nts, so we would be sure to cover the promoter when calculating the fixed-length scores for the baseline sequences, despite some variation in where transcription initiates relative to where translation initiates, and cut out most of the excess sequence flanking the promoters, which might lower the score. Although a length of 41 nts does still factor in some excess flanking region in addition to the promoters in the baseline sequences, this occurs to a much lesser degree, with only 6-8 extra nucleotides being included in the score.

Total information^{41} and *average information*^{41} scores (with the 41 representing the fixed length) were calculated for the experimental set using the most well-conserved 41 nt stretch of each hit, and for the promoters using the 41 nt upstream of the translation start site of each of the genes. These two measures were regarded as the most accurate measures of conservation out of the four.

a)

Hit	Total Info (bits)	Average Info (bits)	Total Info^{41} (bits)	Average Info^{41} (bits)
hit 01	41.05	0.77	39.21	0.96
hit02	69.61	0.87	40.40	0.99
hit03	43.25	0.80	36.23	0.88
hit04	39.48	0.88	36.93	0.90
hit05	39.66	0.83	35.13	0.86
hit06	40.07	0.82	35.30	0.86
hit07	43.40	0.80	35.22	0.86
hit08	80.09	0.87	39.80	0.97
hit09	35.87	0.87	35.87	0.87
hit10	60.24	0.81	39.21	0.96
hit11	45.76	0.80	36.60	0.89

b)

Ortholog Group	Total Info (bits)	Average Info (bits)	Total Info^{41} (bits)	Average Info^{41} (bits)
Baseline1	19.06	0.45	18.99	0.46
Baseline2	32.60	0.71	28.33	0.69
Baseline3	40.88	0.83	34.29	0.84
Baseline4	21.63	0.46	19.90	0.49
Baseline5	37.58	0.82	33.21	0.81
Baseline6	22.59	0.48	19.65	0.48
Baseline7	19.32	0.39	17.08	0.42
Baseline8	32.95	0.69	27.75	0.68
Baseline9	22.44	0.46	18.28	0.45
Baseline10	27.20	0.59	25.37	0.62

c)

	Total Info (bits)	Average Info (bits)	Total Info^{41} (bits)	Average Info^{41} (bits)
t	4.14	4.77	6.43	6.38
Std. Deviation	11.800	0.116	4.620	0.113
Degrees of Freedom	19	19	19	19
p	0.001	0.000	0.000	0.000

Table 3-6Table showing conservation scores calculated for a) hits and b) control sequences. In Total Info^{41} and Average Info^{41} scores are being given only to the most highly conserved 41 nt portion in the

hits and the 41 nt upstream of the start site in the upstream regions. For each scoring method, Table 2 c) displays the values used in the t-test.

Comparing conservation score averages for the four different scoring schemes, the hits consistently scored higher than the baseline sequences. A student's t-test was conducted to determine whether or not the experimental and expected conservations scores differed significantly. The t-test showed that the scores for the hits were, in fact, significantly higher than the scores for the promoters selected as controls, supporting the hypothesis that these hits exhibit an unusually high degree of conservation.

Comparing the scores of the different hits within our selected set addresses the question of whether or not the CSE is unusually well conserved. Although hit 01, which contains the CSE, has an average score of 0.77, which is lower than most of the other hits, its Average Information[41] score, which is the score for the most conserved 41-nt stretch within hit 01, is 0.96 and is higher than the scores for seven of the remaining 10 hits, which suggests that even within this set of 11 highly conserved sequences, the hit containing the CSE is one of the most conserved.

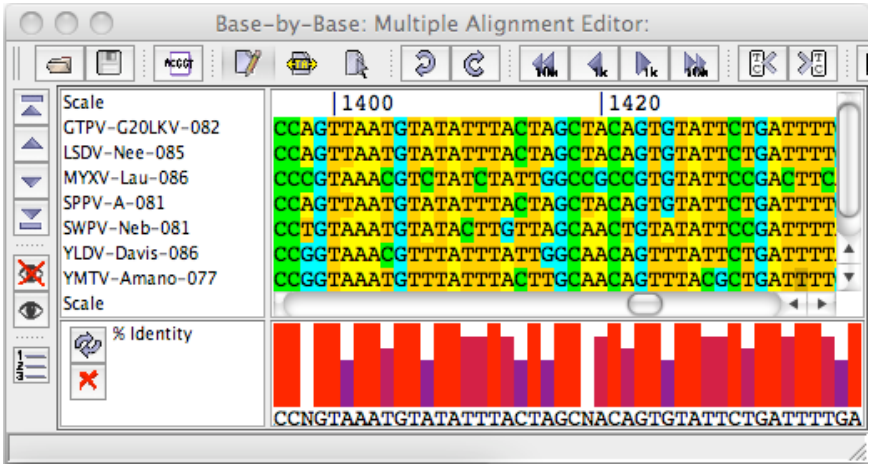
3.7. Functional analysis

3.7.1. Conserved protein motifs

Nine out of 11 of the hits contained at least one promoter, and often overlapped with the coding sequences of adjacent genes, as seen in Table 3-4. However, two of the hits, Hit05 and

Hit06, appeared only in coding regions and did not overlap with any known promoter regions or any other genes.

The interesting thing about high conservation in regions that are solely coding sequence is that conservation at the protein sequence level does not necessitate conservation at the DNA level, given codon degeneracy in the genetic code. A survey of protein and DNA alignments of the VETF gene, in which hit 05 is found, illustrates this point. As shown in Figure 3-19, a high degree of conservation is maintained in the protein sequence despite much lower conservation at the DNA sequence. Additional regions of the VETF gene demonstrating this effect are available in Appendix A, along with protein and DNA alignments of hits 05 and 06, both of which are well conserved at both the protein and DNA level.



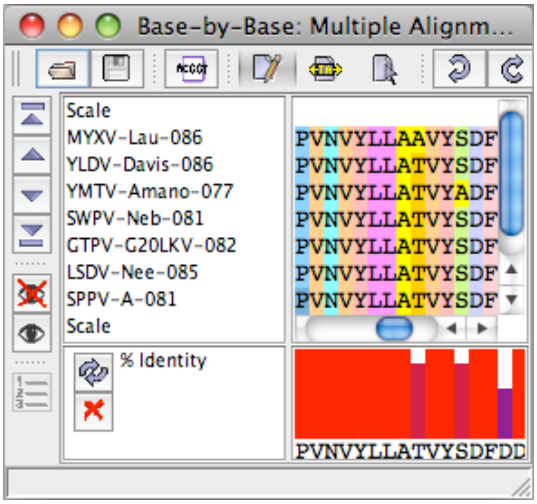


Figure 3-19 DNA (top) and protein (bottom) sequence alignments of the same gene region. Red/purple bars show percent identity.

Hit05 was found within the protein-coding region of the Viral Early Transcription Factor (VETF) gene. VETF is a promoter-binding protein with DNA-dependent ATPase activity that is involved in activating transcription of early genes in poxviruses (Li and Broyles, 1993). The ATPase activity is believed to be important to RNA synthesis since ATP hydrolysis is required for early gene transcription (Broyles and Fesler, 1990). Investigations into the interaction between the viral DNA-dependent RNA polymerase and promoter DNA have shown that RNA polymerase and VETF are both required in the formation of the protein-DNA complex (Broyles and Fesler, 1990). Experiments have subsequently shown that VETF is responsible for recruiting RNA polymerase to the viral early promoter, thereby activating transcription (Li and Broyles, 1993).

Out of general interest about the proteins containing these two hits, we investigated whether or not the VETF and RAP94 genes contained any known conserved domains. The amino acid sequence for hit05 was determined by searching the full protein sequence of the VETF gene for each of the hit's six-frame translation sequences. The amino acid sequence was found to be:

CNNEMFEKNMNNV

The EMBOSS PatMatMotif tool – a tool that searches the full PROSITE database of known protein motifs - was then used to query the PROSITE database for this amino acid sequence to see if it might be associated with a conserved protein domain or family. No matches were found. The PROSITE database was then directly searched for matches against the amino acid sequence for the full VETF gene using the ScanProsite tool in order to see if it contained any known conserved protein domains, and if so, whether or not the hit region was a part of these conserved protein domains. This search returned matches to two conserved domains. One match was to a helicase domain, superfamilies 1 and 2, which binds ATP. The other match was to the C-terminal helicase domain, superfamilies 1 and 2. These matches were distinct, non-overlapping matches in the protein sequence were linked by a 145 nt sequence, and it was in this unmatched linker region that hit06 was found, excluding it from the conserved domain matches (Figure 3-20).

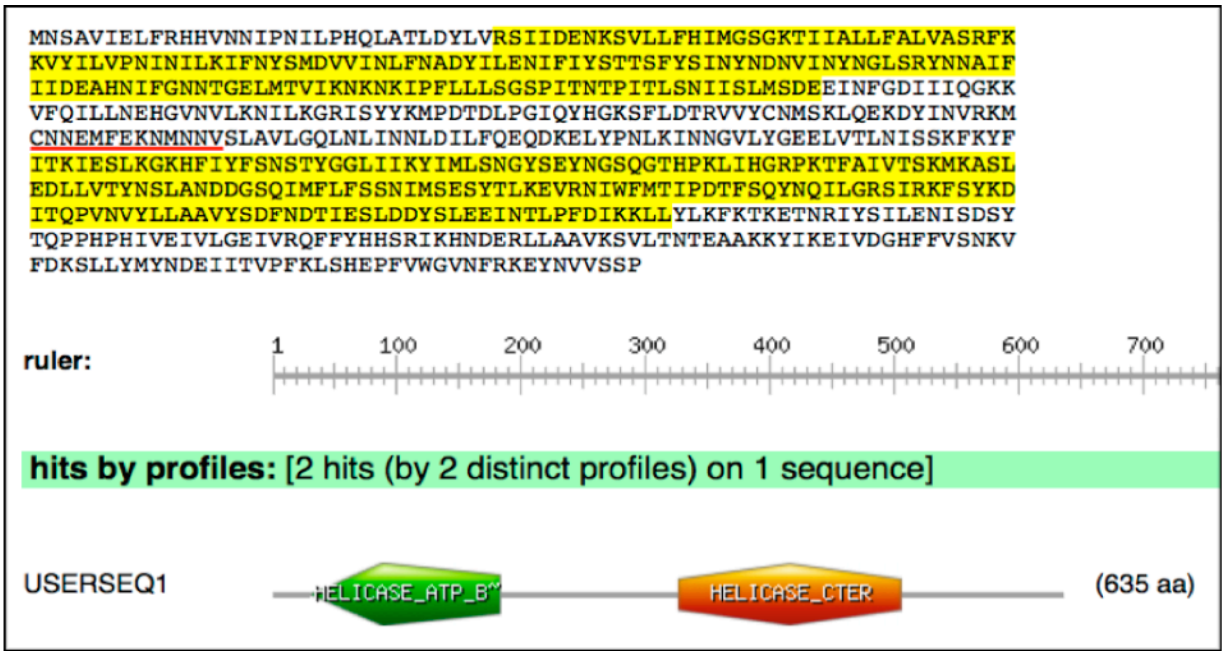


Figure 3-20. VETF amino acid sequence showing conserved domain matches and location of hit06.

Hit06 was found in the RNA Polymerase-Associated Protein (RAP94), a 94 kDa viral polypeptide that associates with DNA-dependent RNA polymerase molecules. It has been shown that only RNA polymerase molecules that contain RAP94 are able to functionally interact with the VETF gene, which is required for its promoter-specific DNA binding activities (Ahn, Gershon and Moss, 1994). This is interesting because both of the proteins that contain hit 05 and hit 06 respectively interact with promoters, which ties them to our promoter-containing hits. Association with VETF therefore enables polymerases to transcribe early genes from a double-stranded DNA template. RAP94-deficient polymerase molecules, on the other hand, are better able to transcribe non-specific single-stranded DNA templates (Ahn, Gershon and Moss, 1994). Based on these findings, RAP94 is believed to confer specificity to the RNA polymerase for promoters of early genes through its association with VETF(Ahn, Gershon and Moss, 1994).

The amino acid sequence for hit06 was determined in the same way as that of hit05; by searching the full protein sequence of the RAP94 gene for each of the hit's six-frame translation sequences. The amino acid sequence was found to be:

LVIFPTHLKIEIER

The EMBOSS PatMatMotif tool was again used to query the PROSITE database for this amino acid sequence. Again, no matches were found. Then the PROSITE database was again directly searched for matches against the amino acid sequence for the full RAP94 gene in order to see if it contained any known conserved protein domains. No matches were found.

It is not unexpected that hits 05 and 06 did not match any known conserved domains from the PROSITE database, since the PROSITE database stores amino acid motifs that characterize protein families. PROSITE database motifs are the most conserved parts of the proteins in these protein families and are therefore the bare minimum required to characterize a protein as belonging to that family. Given how short hit 05 and 06 are it is less likely that a match be found with a protein motif in the PROSITE database.

Moreover, the fact that no hits were found in the PROSITE database does not mean that these conserved sequences do not have conserved functions. Since the PROSITE database looks at all proteins belonging to the same protein groupings, the minimal motifs in the PROSITE database represent the minimal commonalities between proteins from a wide range of hosts, making it even less likely that a match be found between the hit sequences and the database should the hit be associated with a poxvirus-specific function. The *Vaccinia virusVaccinia virus* RNA Polymerase demonstrates this point; although the larger subunits show a high degree of amino acid similarity to the two largest of eukaryotic and prokaryotic cellular RNA polymerases, the smaller subunits, which interact with the face of the protein opposite the catalytic site and are proposed to interact with transcription factors, have no significant resemblance to smaller subunits of cellular RNA polymerases (Amegadzie, Ahn and Moss, 1992). An alignment of the VETF gene in 47 poxviruses (all sequenced poxviruses less the various strains of Vaccinia and *Variola virus*) showed that hit 05 is well conserved in all but 10 of the sequences aligned, supporting the idea that these sequences may be linked to poxvirus-specific functions (Figure 3-21).

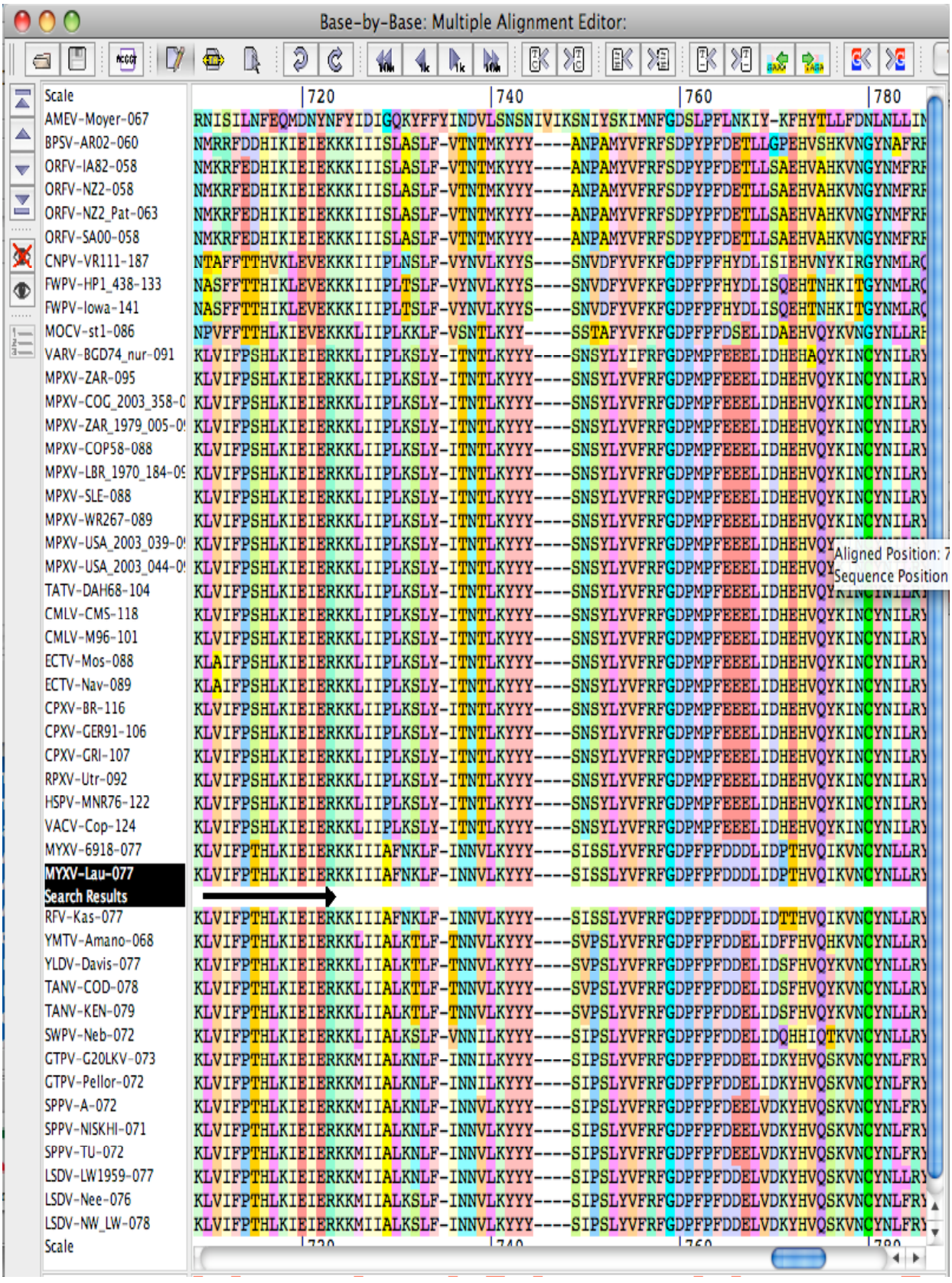

Base-by-Base: Multiple Alignment Editor:
Scale
720
740
760
780
AMEV-Moyer-067
BPSV-AR02-060
ORFV-IA82-058
ORFV-NZ2-058
ORFV-NZ2_Pat-063
ORFV-SA00-058
CNPV-VR111-187
FWPV-HP1_438-133
FWPV-Iowa-141
MOCV-st1-086
VARV-BGD74_nur-091
MPXV-ZAR-095
MPXV-COG_2003_358-0
MPXV-ZAR_1979_005-0
MPXV-COP58-088
MPXV-LBR_1970_184-0
MPXV-SLE-088
MPXV-WR267-089
MPXV-USA_2003_039-0
MPXV-USA_2003_044-0
TATV-DAH68-104
CMLV-CMS-118
CMLV-M96-101
ECTV-Mos-088
ECTV-Nav-089
CPXV-BR-116
CPXV-GER91-106
CPXV-GRI-107
RPXV-Utr-092
HSPV-MNR76-122
VACV-Cop-124
MYXV-6918-077
MYXV-Lau-077
Search Results
RFV-Kas-077
YMTV-Amano-068
YLDV-Davis-077
TANV-COD-078
TANV-KEN-079
SWPV-Neb-072
GTPV-G20LKV-073
GTPV-Pellor-072
SPPV-A-072
SPPV-NISKHI-071
SPPV-TU-072
LSDV-LW1959-077
LSDV-Nee-076
LSDV-NW_LW-078
Scale
Aligned Position: 7
Sequence Position

Figure 3-21 Protein sequence alignment of the RAP94 gene in all poxviruses (less the numerous strains of *Vaccinia* and *Variolavirus*) showing hit 06. Red/purple bars at the bottom show percent identity.

3.7.2. Codon Degeneracy

The simplest consequence of conservation within a coding region is that the amino acid sequence is subsequently conserved. However, given codon degeneracy and assuming unbiased codon usage (and it is worth noting that this may not be a safe assumption to make), it is not required that the DNA sequence be highly conserved in order for the amino acid sequence to be conserved, as discussed in section 3.7.1. Therefore, another hypothesis for the non-promoter sequences was that their conservation might be due to the low degeneracy of their constituent amino acids; since an amino acid with four- or six-fold degeneracy (four or six possible codons encoding it) leaves more room for variation at the DNA level than one with two-fold degeneracy. It follows that a string of amino acids with at most two-fold degeneracy leave little room for variation, since single nucleotide substitutions within this string have little chance of being silent. Were this the case, however, it should be noted that the observed conservation would therefore be linked with the low degeneracy of the amino acids and not necessarily with any conserved function.

To verify whether this was the case, histograms were made in Microsoft Excel showing the number of degenerate codons encoding the amino acid in each position of the protein sequences for hit05 (Figure 3-22a) and hit06 (Figure 3-22b), based on the genetic code.

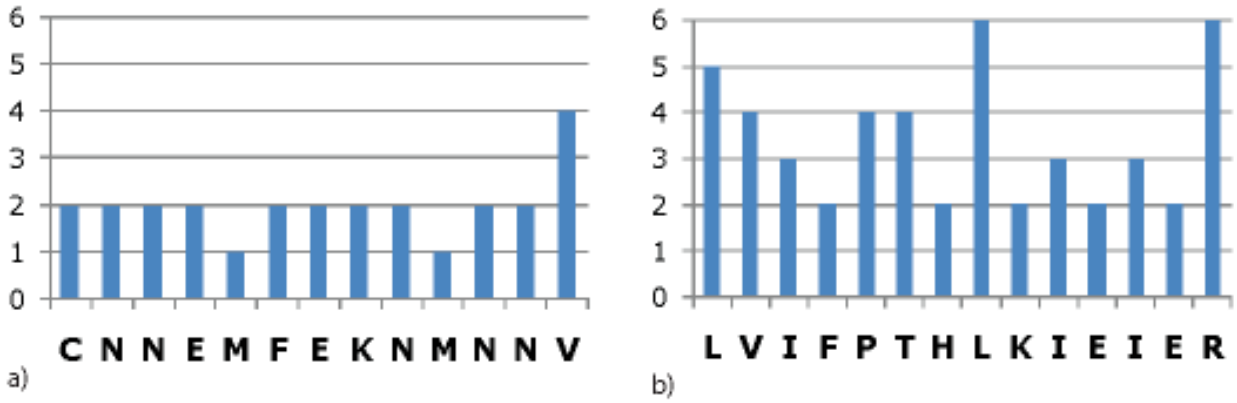


Figure 3-22. Histograms showing the degeneracy of each amino acid in the protein sequences corresponding to a) hit05 and b) hit06. Protein sequences were determined by querying the protein sequences of the genes containing the two hits for the putative amino acid sequences from each of the 6 possible frames.

Examining the codon degeneracy for the amino acids in hits 05 and 06 showed that hit 05 is made up of amino acids with low degeneracy, including two methionine residues for which there is only one codon. This may explain the high conservation of this region. Hit 06 however is made up of amino acids with 2- to 6-fold degeneracy, so the low degeneracy explanation does not hold for hit06.

This supports the hypothesis that an unknown conserved function within the DNA sequence of both of these hits may yet exist, since it has been shown that even well conserved poxvirus proteins have regions in which the protein sequence is conserved without high conservation in the DNA sequence (Figure 3-19 and Appendix A), as discussed in section 3.7.1.

3.7.3. Identifying promoter elements within hits

As discussed in section 3.5, 9 out of 11 hits contained at least one promoter. This meant that the high conservation observed in these 9 hits was partly accounted for by the presence of conserved promoter elements (where the term *promoter elements* refers to the different parts of the known consensus for promoters, refer to Figure 2-4). Whether or not there were any remaining portions of the hit that were unaccounted for by promoter element and an explanation for the high conservation in these remaining portions of the hits, remained to be determined.

Known elements of poxvirus promoters were identified in logos of the hits in order to characterize the hits to delineate promoter elements and visualize whether or not the hits are longer than expected once promoter elements are accounted for (Figure 3-23).

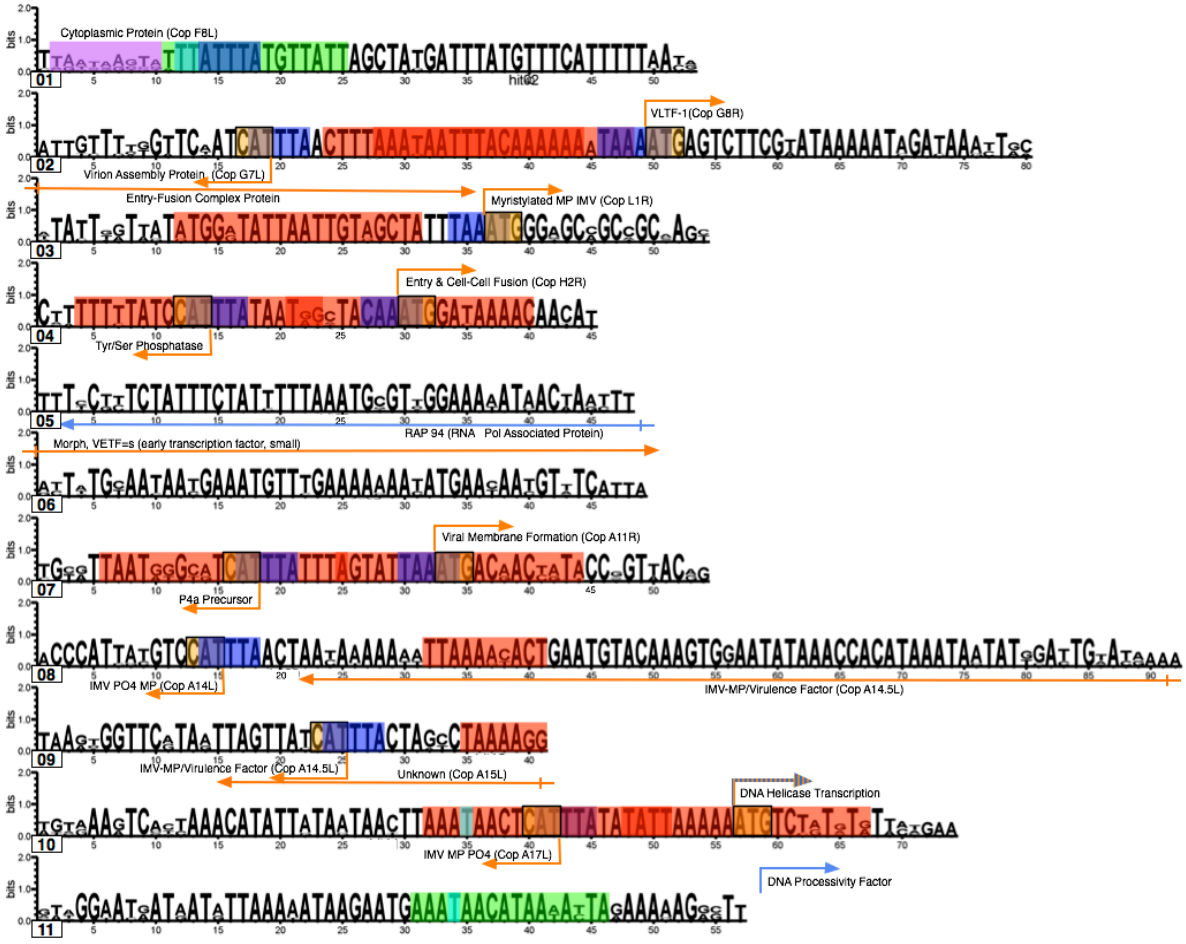


Figure 3-23. Annotated hit logos showing promoter elements. Blue arrows represent early genes, orange arrows represent late genes, and blue-and-orange striped arrows represent genes that are transcribed both early and late in the poxvirus life cycle. Highlighted promoter elements follow the colour key shown in the diagram of the known consensuses of promoters (Figure 2-4).

Although consensus sequences exist for poxvirus promoters, these consensus sequences allow for a great deal of variation since, by nature, the promoters themselves vary in sequence (Figure 2-4). For instance, intermediate and late promoters contain an Upstream Region of around 15 nts that is known to be A/T rich although it does not adhere to a particular sequence sequence consensus. The sequences of the promoter elements contained within the hits, however, are perfectly or near-perfectly conserved. Therefore it should be noted that although the presence of conserved promoter elements has been considered a suitable explanation for the

conservation of the regions of the hits in which these promoter elements fall, they are unusually well conserved even for poxvirus promoters.

In some cases, such as in hit 04, back-to-back promoters accounted for almost the whole length of the hit, while in others, lengthy stretches of highly conserved sequence flank the promoter elements. In hit 02, for instance, stretches of 16 nts and 27 nts lie to either side of the promoters (Figure 3-23). Several of the hits contained Met residues which likely contributed to their high degree of conservation, since Met has only one codon.

Interestingly hits 05 and 06, which fall in the middle of very large genes and are nowhere near the translation start sites of nearby genes, contain motifs that closely resemble poxvirus promoters elements, even with respect to their positions relative to ATG codons. These motifs are shown below in Figure 3-24following the colour scheme for poxvirus promoter elements in Figure 2-4.

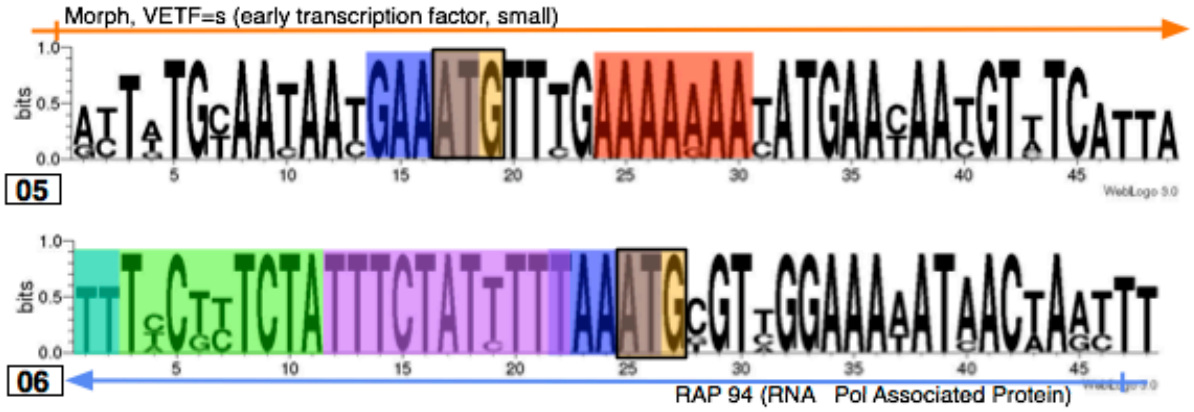


Figure 3-24. Hit 05 and 06 logos with promoter annotations.

As shown in Figure 3-24, hit 05 contains a GAAAT motif overlapping with the ATG codon that resembles the late promoter initiator site, which consists of a conserved TAAAT motif, and, towards the leftmost end of the hit, part of an AT-rich region, albeit a poorly conserved one. Downstream of the start site is an A-rich region that can be compared to the AAANAA motif of the intermediate promoter upstream sequence and another ATG codon. Verifying the amino acid sequence for this gene revealed that both ATG codons are in frame, which suggests these motifs could contain transcription and translation initiation sites (with transcription initiating in the TAAAT and AAANAA motifs and translation initiating at the ATGs), resulting in the formation of truncated proteins and not proteins with different amino acid sequences altogether.

Although one poxvirus gene that is expressed throughout the viral life cycle has been shown to have alternative transcription initiation sites – one early and one late – both of these start sites appeared upstream of the translation start site and therefore did not affect the protein product (Cochran, Puckett and Moss, 1985). Furthermore, it is not characteristic of poxvirus

genes to produce truncated proteins (Condit, 2007) and therefore it is not likely that these motifs act as active transcription and translation start sites for truncated proteins. However, literature searches did not provide any evidence refuting the possibility that sloppy translation initiation results in the production of truncated proteins. One way of testing for such a protein would be to clone an epitope at the 3' end of the gene and carry out a Western blot with an antibody against the epitope. If the Western blot shows proteins of two different sizes, a truncated protein is likely being produced.

The context surrounding the ATG motif in hit 06 closely resembles a poxvirus early promoter (Figure 3-24). To investigate this resemblance more closely, the sequence upstream of the beginning of hit 06 was included in a comparison of hit 06 with the known structure of an early promoter (Figure 3-25).

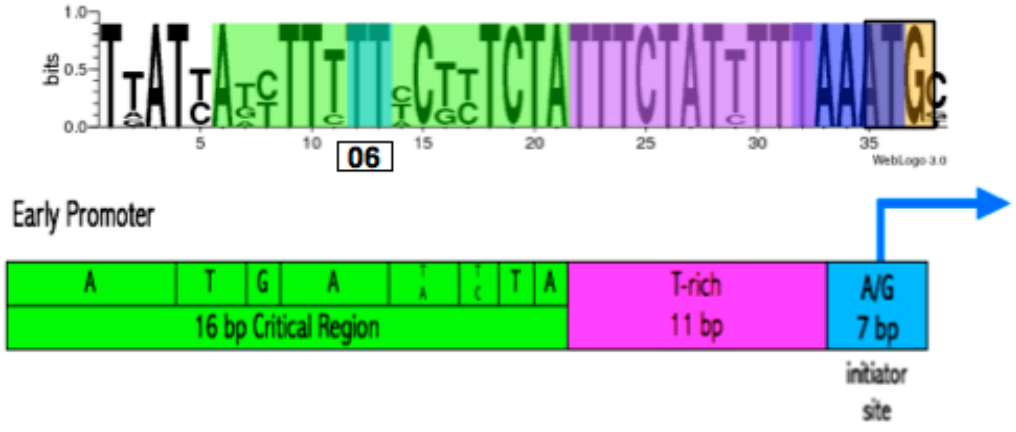


Figure 3-25. Comparison of hit 06 and its upstream region with the known structure and sequence of poxvirus early promoters.

The perfectly conserved TAAAT motif that overlaps with the ATG closely resembles the initiator site of poxvirus late promoters while also satisfying the more loosely defined early promoter initiator site. Upstream of the TAAAT motif is a T-rich region of comparable length to the 11 nt T-rich region known to appear in the same position in early promoters (Figure 3-25). Upstream of that, the sequence from positions 1 to 10 of hit 06 is comparable to the latter half of the 16 nt critical region of early promoters. In particular, the perfectly conserved TCTA motif in positions 8 to 11 and the TT in positions 1 to 2 of hit 06 compare to the same motifs appearing at positions -21 to -18 and -28 and -29 relative to the transcription start site in the early promoter. Upstream of the TT at positions 1 to 2 of the hit, however, the resemblance diminishes significantly. Verifying the amino acid sequence for hit 06 revealed that the ATG codon is not in frame and would therefore result in the formation of an entirely different protein as oppose to a truncated version of the RAP94 protein. Considering the likeness of the sequence surrounding this ATG to an early promoter, it seems likely that it be an active translation start site. Although literature searches did not provide any evidence of ORFs beginning at this site, no evidence was found refuting an alternative out-of-frame start site either.

The annotation of promoter elements within the hits thus shows that only hits 04 and 07 are mostly accounted for by the presence of more than one promoter, while the rest contain regions that are not accounted for by the presence of conserved promoter elements but still exhibit a very high degree of conservation that may be accredited to a novel conserved function.

3.7.4. Identifying sequence motifs within hits

As is apparent from the identification of conserved promoter elements within the hits, short motifs in the range of 5 to 10 nts can give hints as to the function of a sequence. For instance, the presence of the sequence TAAAT, which is the known sequence for the initiator site of poxvirus late promoters, not only identifies the sequence immediately upstream as a putative promoter but also suggests that a translation start site may be nearby. To search for other such hints, a *de novo* motif-finding program called MEME was used to search for short motifs in the hits (Bailey *et al.*, 2006).

MEME, which is run as a web-based tool available at http://meme.nbcr.net/meme4_1_1/cgi-bin/meme.cgi, takes protein or DNA sequences in fasta format as input. The user can specify a range corresponding to the minimum and maximum motif length (for example it can search for motifs of 5-8 nts) and also set parameters corresponding to the expected number of occurrences of motifs per sequence (ie. 0, 1 or multiple per sequence). Thus MEME is similar to Java Pattern Finder in that it can be used to identify sequence matches of a specific length but unlike JaPaFi, the length of the sequence matches can be given as a range and MEME is not limited to identifying sequence matches that appear in every sequence of the given set. Rather, MEME identifies motifs that appear in any number of sequences in the set, be it one that appears numerous times in all sequences, once in each of two sequences, or numerous times in a single sequence. For this reason, the user can specify the maximum number of highest scoring motifs for viewing, since the entire list of motifs identified in any or all sequences would be very lengthy and most motifs would be of little significance. Also unlike JaPaFi, MEME is limited to processing data that consists of no more than 60 000

characters, and therefore would not have been capable of processing poxvirus genomes of 150 – 350 kB.

The output from MEME consists of a list showing a user-specified number of high-scoring motifs, with the motifs being ranked based on their e-values, which estimate the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and number of occurrences, that one would find in a similarly sized set of random sequences (Bailey and Elkan, 1994). Also shown are p-values, which represent the probability of a random string having the same match score or higher (Figure 3-26). P-values are calculated from the match score of the site with the position specific scoring matrix for the motif and take into account factors such as how many occurrences of the motif there are and what their distribution is across the given sequences, and how conserved the motif is (Bailey and Elkan, 1994). For each motif identified, a logo of the motif based on all occurrences of the motif in the given sequences is presented, as well as a table summarizing which sequences it appeared in, which strand, its start position, its p-value, and the sequence context in which it appears (10 nts to either side of the motif). In addition to the logo and the table, a BLOCKS format diagram is also output, comparing the locations of the different occurrences of the motifs within the sequences in which they appear.

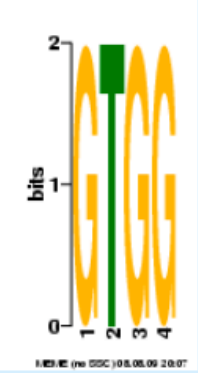
P N **MOTIF 4** width = 4 sites = 2 llr = 15 E-value = 1.6e+003

SEQUENCE LOGO

Information Content
8.0 (bits)

Relative Entropy
10.8 (bits)

Download LOGO
Without SSC:[EPS][PNG]
With SSC:[EPS][PNG]



NAME	STRAND	START	P-VALUE	SITES
09_GTPV.fasta	+	3	5.72e-04	TAA GTGG TTCATAATTA
08_GTPV.fasta	-	63	5.72e-04	TATTATTTAT GTGG TTTATATTCC

Motif 4 block diagrams

Name	Lowest p-value	Motifs
09_GTPV.fasta	5.72e-04	+4
08_GTPV.fasta	5.72e-04	-4
SCALE		1 25 50 75

Figure 3-26. MEME sample output for one motif, MOTIF 4.

3.7.5. Motifs within the hits

To investigate the motifs shared among any or all of the hits, MEME was used to search the 11 hit sequences for the top 15 motifs of 2-10 nts shared between the hits. Since MEME accepts DNA sequences and not alignments as input, the hit sequences as they appear in the MYXV genome acted as our model sequences for motif search and analysis. As this was an exploratory analysis to see if any motifs are shared between the hits, there were no expectations in terms of what sequences the program might identify, where in the hits they may appear, or

what functions they may serve. The significance of the motifs that the program identified were judged based on their e-values (where lower e-values were optimal), how conserved the motifs were (as shown in their logos), and how many occurrences there were of the motifs. Motifs that only appeared in two sequences were regarded as occurring by random chance; even more so if they were not perfectly conserved at every position or if they were on the shorter end of the 2 – 10 nt range specified when running MEME.

The highest-scoring motif was a 4 nt sequence occurring once in each of hits 07 and 03, with an e-value of $3.6*10^{2}$ and p-values of $3.38+10^{-4}$ (Figure 3-27).

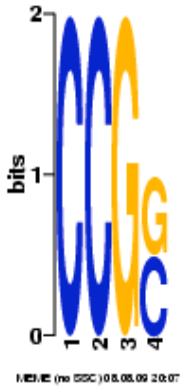


Figure 3-27. Logo of highest-scoring motif identified within the hits by MEME motif finder.

No literature was found describing a conserved function associated with CCG[C/g] motifs in poxviruses, although CCGG motifs in the promoter region of an adenovirus gene have been shown to act as DNA methylation sites, resulting in the inactivation of the gene (Langner, Vardimon and Renz, 1984). To comment on the motif's e-value of $3.6*10^{2}$, the e-value of the motif was compared to NCBI's default threshold e-value of 10, where BLAST matches with e-values greater than 10 are not reported. Given that the e-value of the motif was 10-fold higher

than NCBI's default threshold e-value, and given the low number of occurrences and the short length of the motif – making it more likely that it randomly appeared in the two hits – this motif was not considered a significant hint as to the conserved functions of the hits.

Similarly, 13 of the remaining high-scoring motifs had e-values greater than 10 (by 1 to 3 orders of magnitude) and generally had only two occurrences. Only one motif strayed from this trend, with 11 occurrences in 7 different hits (Figure 3-28). This logo contained a perfectly conserved ATG and was expected since ATGs have already been identified as translation initiation sites falling within the promoters contained within the hits, in addition to Met residues that have been identified within many of the hits.

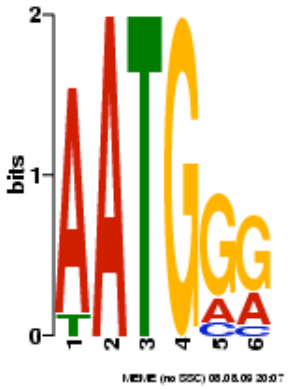


Figure 3-28. Logo of motif containing ATG codon.

In addition to a perfectly conserved ATG, this motif also contains a well conserved A residue in the position upstream of the ATG which is not unusual since all but two of the promoters in the hits are late promoters, which contain a conserved TAAAT motif that overlaps the ATG. What is unusual about this motif is the two positions downstream of the conserved ATG, since these fall in the coding region of genes, defining what the second residue is. It seems

unlikely that there be a consensus in the protein sequence for such a large set of genes (all late genes). However, as indicated by the motif logo, neither of these positions is very well conserved, and this motif has an e-value of $2.4*10^2$, which is also above the NCBI default threshold.

Thus no significant motifs were identified as being shared between the hits.

3.7.6. Motifs within early, intermediate and late promoters

MEME was run on the upstream sequences of 5 genes of each temporal class in turn, where the upstream sequences consisted of 100 nts upstream of the translation start sites for the selected genes. Although 5 is a small number of genes to conduct motif searches on and ideally a larger number of genes would have been investigated to lend statistical significance to our findings, we were limited by the number of intermediate genes we could use since there are only 5 poxvirus genes that have been identified as belonging to this temporal class. The gene VBRC gene names and gene families of the genes selected for motif analysis are shown below (Table 3-7).

Early Genes	
MYXV-Lau-019	Ribonucleotide Reductase - small subunit
MYXV-Lau-039	DNA Pol
MYXV-Lau-066	Thymidine Kinase

MYXV-Lau-084	Uracil-DNA glycosylase
MYXV-Lau-102	VITF-3 34 kDa subunit
Intermediate Genes	
MYXV-Lau-094	VLTF-2 late transcription factor 2
MYXV-Lau-095	VLTF-2 late transcription factor 3
MYXV-Lau-058	VLTF-1
MYXV-Lau-045	DNA-binding phosphoprotein
MYXV-Lau-049	RNA Helicase/NPH-II
Late Genes	
MYXV-Lau-086	VETF-s (early transcription factor small)
MYXV-Lau-091	NPH-I/Helicase, virion
MYXV-Lau-034	IFN resistance/PKR inhibitor (Z-DNA binding)
MYXV-Lau-074	Tyr/Ser phosphatase
MYXV-Lau-077	RAP94 (RNA pol assoc protein)

Table 3-7 **Early, Intermediate and Late genes selected for motif search and analysis.**

This trial was treated as a control and it was hypothesized that MEME would identify known conserved promoter elements, recognizable by their sequence, which should reflect the consensuses of the promoters for each class, and by their positions relative to the start site as outlined in Figure 2-4.

In examining the output, the positions of the identified patterns within the sequences were in fact more important than the motifs themselves and how conserved they were, since the critical region and linker region of the early promoter and the upstream sequence and linker region of the intermediate and late promoters are loosely defined (Figure 2-4) whereas the distance between these elements and the transcription start must be quite precise for the promoters to carry out their functions.

High-scoring motifs for early, intermediate and later upstream region trials had e-values of $2.7*10^3$, $8.0*10^3$ and $1.2*10^2$, respectively, and were all above the NCBI-defined default threshold of 10. The highest-scoring motif was found to have 3 occurrences in 3 different sequences, and the remaining 14 high-scoring motifs appeared in only two different sequences, which shows that these may be random matches and not motifs. Moreover, many of the occurrences of the motifs identified were not within the 35 nt stretches that contain the promoters, and those patterns that were identified within this stretch did not align with other occurrences (Figure 3-29).

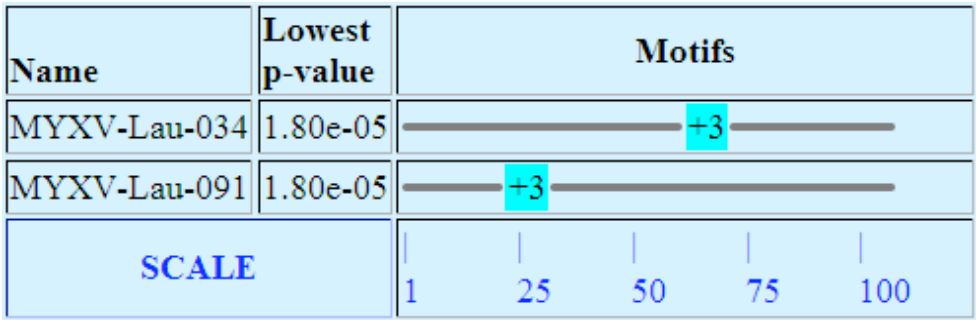


Figure 3-29. Diagram showing the locations of a motif identified between two late promoters. Translation start sites are located at the 100 nucleotide mark, with promoters appearing between 70 and 100. + and – signs refer to the strand.

Promoter elements were not detected in running MEME on early, intermediate and late promoters. This is not surprising since conserved promoter elements do not necessarily manifest themselves as highly conserved sequences so much as trends in the nucleotide composition of the regions defined as promoter elements.

3.7.7. Motifs shared between the hits and early, intermediate and late promoters

Lastly, MEME was used to search for conserved elements shared between the hits and each of the three types of upstream sequences in turn, hypothesizing that the program would pick up promoter elements appearing in the 30 nts upstream of the translation start site in the upstream sequences (therefore between positions 70-100) and appearing within 30-35 nts upstream of the transcription start sites, which are near the translation start sites contained within the hits. More specifically, motifs were expected between the early gene upstream regions and the early promoters contained within hits 1 and 11, between the intermediate gene upstream regions and the early/late gene promoter in hit 10 and possibly other late promoters in the remaining hits, and between the late gene upstream regions and the late promoters appearing in the hits.

When searching for motifs between the hits and early upstream sequences, no significant matches were identified, with the highest-scoring motif appearing in only two early upstream regions and two hits. The 6 nt motif itself did not match any of the known elements of poxvirus

promoters and was not found within 40 nts of the translation start site in the two early upstream regions that were searched. With an e-value of 6.8 * 10^2, it was above the NCBI default threshold e-value. The remaining top-scoring motifs only appeared in two sequences each, often two upstream regions or an upstream region and a hit, and their sequences did not resemble known promoter elements. Of the motifs identified within the 5 early upstream regions, only 2 appeared within the actual promoter (Figure 3-30, motifs 3 and 9). Therefore, no significant motifs were identified as being shared between the hits and early promoters, despite the fact that early promoter elements have been identified in two of the hits, nor were any significant motifs identified as being shared among the early promoters themselves, since the exact sequences of promoters vary.

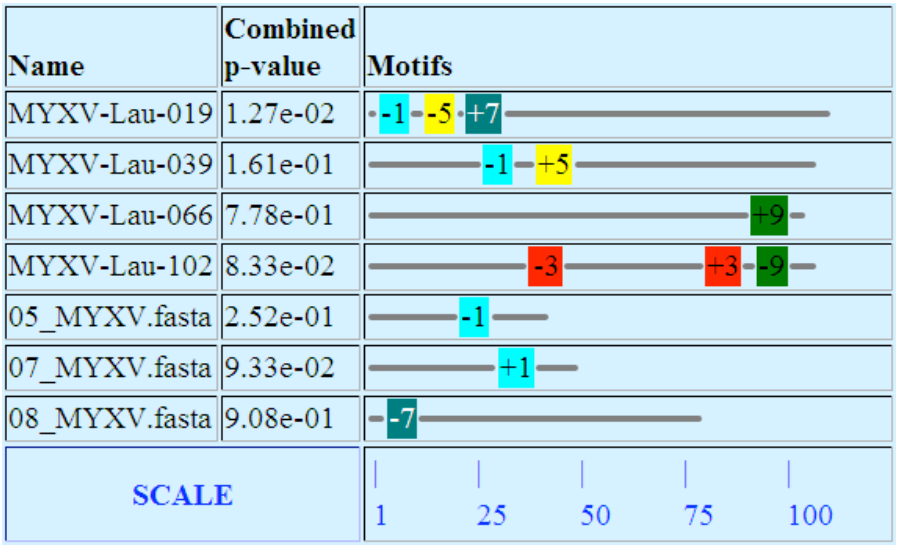


Figure 3-30. Summary of motifs identified between hits and early gene upstream sequences. In early upstream sequences (MYXV-Lau-019, -039, -066 and -102) translation start site is at 100, with promoter between 70-100. + and – signs refer to the strand.

When searching for motifs between the hits and intermediate upstream sequences, there were many more motifs occurring in the actual promoter than there had been in the early

upstream sequences, however most of these, save for two, had high e-values in the order of 10^2 and low numbers of occurrences (Table 3-7).

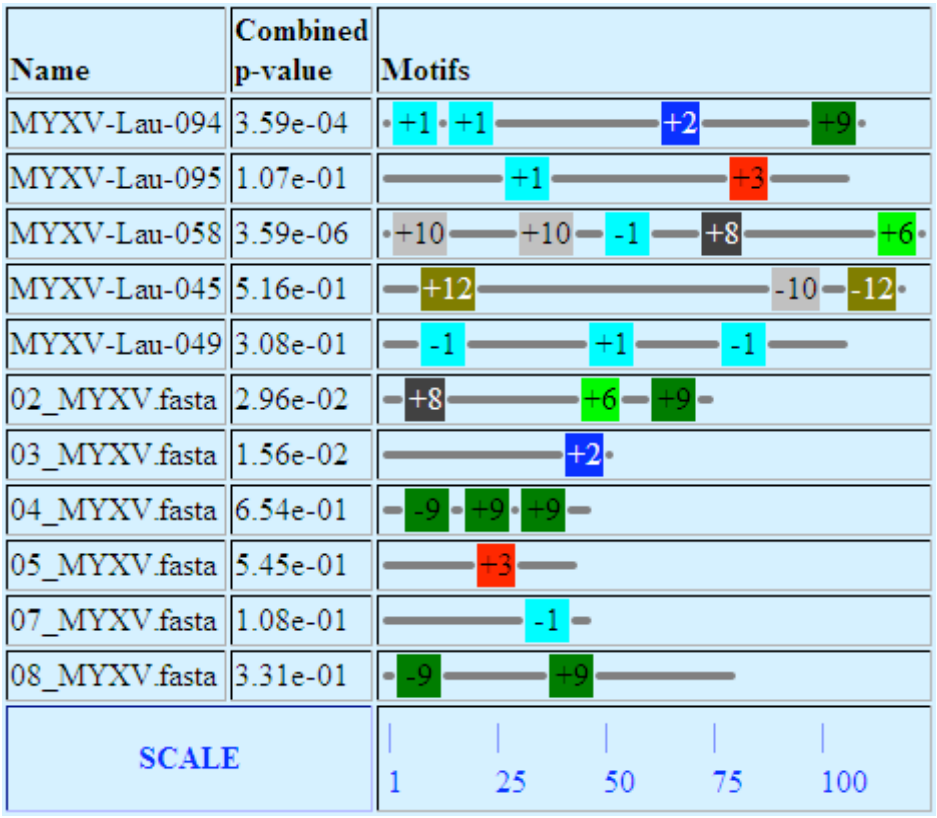


Figure 3-31. Summary of motifs identified between hits and intermediate upstream sequences.

Of the two motifs that had more multiple occurrences, one had 7 occurrences in upstream regions and one occurrence in hit 07, and the occurrences in the upstream sequences did not align with one another or the positions of known promoter elements. Therefore only one of these motifs was of interest as a result of its sequence, conservation and the frequency and distribution of its occurrences (Figure 3-32).

Motif 9, which appears in hits 02, 04, and 08 (shown in dark green in Figure 3-31) and included a highly conserved ATG, was a recognizable putative promoter that appeared in one upstream sequence and multiple times in each of 3 hits.

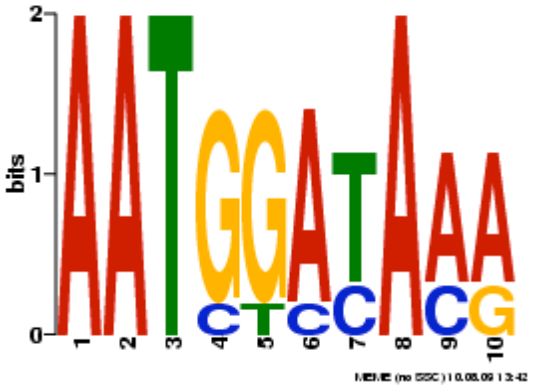


Figure 3-32. Logo of motif 9 found in hits and intermediate upstream sequences. E-value of $2.3*10^4$ and 7 occurrences in 1 upstream region and 3 different hits.

This motif occurred once in hit 2, where it was part of the coding region of the VLTF-1 gene and not associated with either of the promoters contained in hit 2. It also occurred twice in hit 08 and in one of these instances it contained the translation initiation site for the IMV PO4 MP gene, while the other instance was in the coding region of the IMV-MP/Virulence Factor gene. Lastly, it occurred three times in hit 04, with two of these instances corresponding to the start sites of the Tyr/Ser Phosphatase gene and the Entry & Cell-Cell Fusion gene, and the third instance occurring in the overlap of the two promoters.

The sequence of the motif satisfies the known consensus for the transcription initiator sites in both intermediate and late promoters, with the conserved ATG marking the translation initiation site and the highly conserved A upstream of the ATG forming part of the TAAA[AAA]T motif. The e-value of $2.3*10^4$ for this motif was 3 orders of magnitude higher than the NCBI default threshold e-value however, limiting its statistical significance.

Lastly, searching for motifs in the hits and late upstream sequences gave similar results, with only two motifs occurring more than twice. Of these two high-scoring motifs, one had three occurrences in three different hits and 2 occurrences in late upstream sequences. The occurrences in the hits did not coincide with any promoter elements, nor did the upstream sequence occurrences fall in the promoters. The second of these high-scoring motifs, however, appeared in 15 locations across 6 hits and 4 upstream sequences (Figure 3-33). Its occurrences in upstream regions align at the translation start site and most of its occurrences in the hits coincide with active translation start sites.

Name	Lowest p-value	Motifs
05_MYXV.fasta	6.20e-06	+1
03_MYXV.fasta	6.20e-06	+1 +1
04_MYXV.fasta	2.33e-05	-1 +1
MYXV-Lau-074	2.33e-05	-1 +1
08_MYXV.fasta	2.95e-05	-1 +1
MYXV-Lau-077	2.95e-05	+1
10_MYXV.fasta	8.62e-05	-1
MYXV-Lau-091	1.32e-04	+1
MYXV-Lau-086	1.64e-04	+1
02_MYXV.fasta	2.56e-04	+1 +1
SCALE		1 25 50 75 100

Figure 3-33. Distribution of motif occurrences for highest-scoring motif identified in hits and late upstream sequences.

Like the top-scoring motif from the intermediate gene set, this motif contains the translation initiating ATG, with the remainder of the motif extending three nucleotides to either side of it (Figure 3-34). With an e-value of $6.8*10^{-1}$, it is the only motif identified that can be considered statistically significant according to NCBI's default threshold e-value.

Figure 3-34. Logo of highest-scoring motif in hits and late gene upstream regions. E-value of $6.8*10^{-1}$ and 15 occurrences in 4 upstream regions and 6 different hits.

Superimposing this motif with the high-scoring motif from the intermediate upstream sequence set reveals that the two motifs overlap very well.

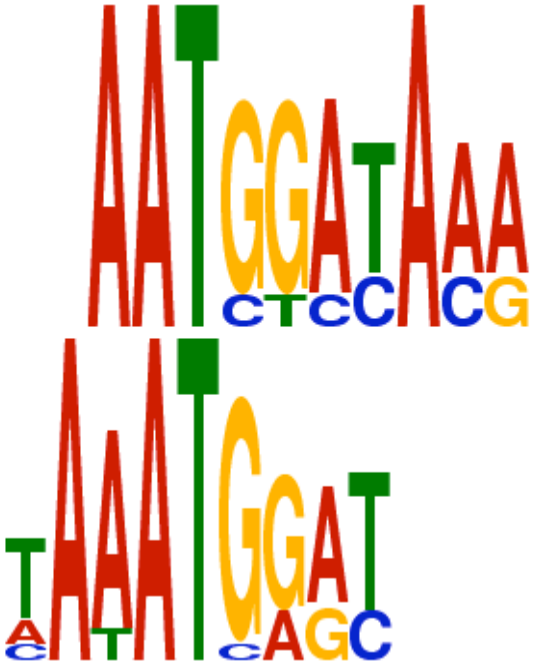


Figure 3-35. Superimposition of intermediate gene high-scoring motif (top) and late gene high-scoring motif (bottom)

What was unusual about both of these motifs is that if the ATG is a translation initiation site, the remainder of the motif extends into the coding sequence of the genes by 2 residues beyond the start codon in the case of the intermediate motif, 1 residue in the case of the late motif. In placing restrictions on the nucleotides appearing at these positions the possible residues in position 2 of the peptide were limited to residues with a G or A in the first nucleotide position, an A or G in the second and a T orC in the third, asshown below (Figure 3-36).

gat	**D**
gac	**D**
ggt	**G**
ggc	**G**
aat	**N**
aac	**N**
agt	**S**
agc	**S**

Figure 3-36 Possible position 2 residues, as dictated by motifs identified between the hits and intermediate and late promoters.

To determine how many genes followed this trend and verify that it was specific to late genes, a survey was conducted of all MYXV genes to see what proportion of late genes reflect this disposition to D, G, N or S in position 2. Among late genes, the most frequently occurring position 2 residues in descending order were D (23%), A (17%), and S (14%), while the sum of the frequencies of the four amino acids dictated by the motif (D + G + N + S) made up 48% of all late genes. Early genes were found to demonstrate comparable preferences for D (17%), K (17%) and S (13%), with the sum of the frequencies of the four amino acids dictated by the motif (D + G + N + S) making up 30% of all early genes. Frequencies were also measured for intermediate genes however since only 5 intermediate genes are known, these results were not considered to be statistically strong and are therefore not shown. In the set consisting of all MYXV genes, without isolating the temporal classes, position two preferences were found to be D (17%), A (12%) and E (11%) with the sum of the frequencies of the four amino acids dictated by the motif (D + G + N + S) making up 39% of all genes. These results are summarized below (Table 3-8).

All genes	D	17%
	A	12%
	E	11%
	D+G+N+S	67%
Late genes	D	23%
	A	17%
	S	14%
	D+G+N+S	48%
Early Genes	D	17%
	K	17%
	S	13%
	D+G+N+S	30%

Table 3-8 Summary of most frequently occurring position 2 residues among all, late and early genes.

Looking at these frequencies from a slightly different angle, the temporal class breakdown of all genes in which D, G, N or S appear in position 2 was determined, hypothesizing that the majority of the genes with each of these residues appearing in position 2 would be late genes. These proportions could, however, be skewed by the sheer numbers of genes belonging to each class, with the late class having 81 classified members and the early class having 23. These counts revealed that 66% of all genes containing D residues in position 2 were late genes, versus 14% early genes and the remainder being unclassified genes. Similar breakdowns were calculated for G, N and S and are shown below (Table 3-9). These numbers show that between early and late genes, the predisposition for D, G, N and S position 2 residues is much more prominent in late genes.

D	
Early	0.14
Late	0.66
?	0.21
G	
Early	0.00
Late	0.43
?	0.57
N	
Early	0.00
Late	0.43
?	0.57
S	
Early	0.18
Late	0.65
?	0.18

Table 3-9 Summary of temporal class breakdowns of all genes with D, G, N or S occurring at position 2.

As another statistical study of the trends observed in late gene position 2 residues, the relative occurrence was calculated for late genes with D, G, N or S in position 2 versus all genes. The relative occurrence is simply the ratio of the proportion of late genes that contain the aforementioned position 2 residues relative to the ratio of all genes with these position 2 residues.

$$= \frac{(\#\text{ late genes with D, G, N, S})}{(\#\text{total late genes})}$$

(# total genes with D, G, N, S / #total genes)

= 0.48/0.32

= 1.23

The relative occurrence (>1) suggests that late genes have a predisposition towards these residues in position 2, i.e. above baseline frequency.

Thus, these statistics all support the observation that late genes have a temporal-class specific predisposition to the residues D, G, N and S in position 2. The next step was to explore why.

3.7.8. Kozak Sequence

Closer inspection of the late promoter motif reveals that it bears a likeness to the Kozak sequence, a sequence that occurs in eukaryotic RNA and plays a major role in translation initiation. The Kozak sequence functions by slowing down the speed of scanning by the ribosome, and the ribosome requires this sequence or some variation on the Kozak consensus to recognize the initial AUG. The full Kozak consensus sequence is known to be GCCA/GCCAUGG, and while adherence to this consensus optimizes recognition of the initial AUG, two positions in particular – an A or G at -3 and a G at -4 relative to the ATG - are also known to affect translation efficiency (Figure 3-37) (Olafsdottir *et al*., 2008).

Figure 3-37. Consensus of the Kozak sequence, the eukaryotic mRNA signaling sequence.

While many of the viral promoters in poxvirus genomes do not strictly adhere to the Kozak sequence and these genes still get translated sufficiently, it seems intuitive that late genes would have a stronger Kozak sequence because many of the late genes are translated into structural proteins of which many units are needed for building the progeny virus particles. A strong Kozak sequence would then enhance the amount of protein synthesized.

A literature search revealed that the presence of the Kozak sequence in the context surrounding the ATG codon in late genes had been discovered by others, and that Sanchez-Puig and Blasco had conducted mutation studies to investigate the effect of the -3 and +4 nucleotides on translation efficiency in *Vaccinia virusVaccinia virus* (Sanchez-Puig and Blasco, 2008). In surveying a large set of late genes, they found that while the genes showed a preference for A in the -3 position (with 101 out of 160 genes studied reflecting this preference) and G or A in the +4 position (with 65 out of 160 genes reflecting this preference), site-directed mutagenesis studies did not reveal any change in translation efficiency as a result of mutations at these positions (Sanchez-Puig and Blasco, 2008). As an explanation for their inability to show a change in translation efficiency, the authors suggested that the promoter used in their β-gal construct may have been strong enough to overcome the effects of changes in the -3 and +4 positions.

These experiments therefore refute the claim that the motif identified acts as a Kozak sequence, enhancing translation efficiency in late genes, despite the motif containing a conserved ATG at +1 and a conserved G at +4. That said, the authors themselves seem to suggest that their inability to show a change in translation efficiency as a result of mutations in the Kozak sequence may be due to external factors.

4. Conclusions & Future Work

4.1. Conclusions

The discovery of the conserved sequence element (CSE) by Brunetti et. al. raised the question of whether or not a 42 nt sequence that is perfectly or near-perfectly conserved in 7 different poxviruses from 4 different genera is unusual in its length and degree of conservation. In this study, we have discovered, by using the Java Pattern Finder program, that there are in fact 10 comparable sequences present in this set of genomes. Thus, the CSE is actually part of a larger conserved sequence that is only one of 11 hit sequences that are unusually well conserved among these genomes.

Nine out of 11 of these hits contain conserved promoter elements, as summarized below, and we hypothesized that the presence of particularly well-conserved promoters partially accounted for most of the conservation observed in these hits. Using a scoring method we developed based on position scores that represent the degree of conservation at each position of a sequence alignment, as observed in alignment Logos, we found that the conservation scores for the hits were significantly higher (student's t-test) than those obtained for control set 10 different promoters. Moreover, many of these promoter-containing hits were longer than expected once promoter elements were accounted for.

The 11 hits were analyzed for conserved functions using a number of sequence-based bioinformatics methods, including promoter element searches based on the consensus sequences of the three classes of poxvirus promoters and motif searches within the hits themselves, with the following results.

Hit 01 contained the CSE, which was later shown to act as a promoter in poxviruses. Hit 01 also included 10 nts upstream of the beginning of the CSE although it contained more differences than the CSE, which explains why a longer sequence was identified with JaPaFi. This flanking region is believed to be part of the promoter for the Cytoplasmic Protein gene, a bottom strand gene that starts downstream of the beginning of the hits.

Hits 02, 04, 07 and 10 all contain bidirectional promoters, where promoter elements for two divergently transcribed, opposite strand genes overlap in the non-coding sequence between the genes. In the cases of hits 04 and 07, these bidirectional promoters make up almost the entire hit, which may explain the conservation of these hits since the promoter elements place constraints on the nucleotide makeup of these regions. In the cases of hits 02 and 10, however, the hits contain lengthy stretches flanking the bidirectional promoters that are also highly conserved, suggesting that these regions may have conserved functions in addition to the well conserved bidirectional promoters.

Hits 03, 08 and 09 also contain promoters, which may accounts for parts of the conservation observed in these hits since promoters place sequence constraints on the DNA.

Furthermore, the majority of each of these hits falls within the coding sequence of nearby genes. Where promoters and coding sequences overlap, this places two constraints on these sequences; protein sequence conservation as a result of the conservation of these genes, and DNA sequence conservation due to conserved promoter elements. Further investigations of these sequences should be conducted, including a survey of the codon degeneracy of the amino acids in their protein sequences, since low codon-degeneracy can limit the variability of the underlying DNA sequence.

Hits 05 and 06 fell entirely within the coding sequences although motifs resembling poxvirus promoters were identified in both; in hit 05 conserved elements of a promoter that would have resulted in a truncated protein were identified while in hit 06 a motif resembling an early promoter that would have resulted in an out-of-frame alternative translation start site was identified. These motifs raised the question of whether or not protein products were produced from transcripts initiating at these alternate start sites. The hit sequences also raised the question of why conservation was observed at the DNA level when it is not required in order for the protein sequence to be conserved. The codon degeneracy of the nucleotides in the protein sequences of these hits showed that hit 05 is mostly made up of codons with two-fold degeneracy. This limits the possible variation at the DNA level to a degree that might partially explain the conservation of the hit. Hit 06, however, consists mostly of codons with three- to six-fold degeneracy. Its protein sequence can, therefore, accommodate much more variation at the DNA level, suggesting a novel conserved function in this DNA sequence. Querying the protein sequences of both of these hits against the Prosite database, neither one was found to be a part of known conserved protein domains. Protein alignments of the VETF gene and the RAP94 gene,

in which hits 05 and 06 were found, respectively, showed that both of these sequences are well conserved among a larger set of genomes, suggesting that they may have pox-wide conserved functions which would not necessarily have any hits in the PROSITE database. This still does not explain the high degree of conservation observed at the DNA level, suggesting that other novel conserved functions may yet exist in the DNA. Further investigations into conserved functions that would result in DNA sequence conservation should be conducted, as well as investigations into potential family-wide conserved functions of the proteins.

A single promoter element was identified in hit 11, acting on the DNA Processivity Factor gene, however, this only accounted for part of the sequence, suggesting that the remaining portion may yet have an unknown conserved sequences.

Motif searches identified a motif surrounding the translation starts sites of late genes. This motif brings together the highly conserved TAAAT motif that makes up the initiator site of late promoters and the Kozak sequence, which serves as a signal to the ribosome in eukaryotic translation. It was shown by subsequent analysis of the context surrounding the start sites of all late genes that late genes consistently adhere to the Kozak constraints as they pertain to the -3 and +4 positions in particular. The idea that late genes, which usually encode structural proteins that are required in abundance, would mimic the eukaryotic translation signals is conceptually sound, especially given that the -3 and +4 positions in particular affect not only the recruitment of ribosomes to the site but also the amount of protein produced (Olafsdottir *et al*., 2008). Despite this, however, site-directed mutagenesis studies have shown that changing these positions does not always affect translation efficiency (Sanchez-Puig and Blasco, 2008).

Thus, sequence analysis of these 11 highly conserved sequences has provided a number of inferences regarding conserved functions that may be contributing in part to the conservation observed in these sequences; namely that late promoters contain a Kozak sequence that enhances translation efficiency of late genes, and hits 05 and 06 may be part of pox-wide conserved protein domains. Hits 05 and 06 may also contain alternative transcription initiation sites resulting in truncated or altogether different proteins. These should be tested for in future wet-lab experiments. Our speculations as to what the roles of these sequences might be include origins of replication, transcription factor binding sites, sites of protein interaction and packaging signals.

4.2. Future Work

4.2.1. Expanding the set of genomes

Since this analysis has been conducted entirely on the 7 poxvirus genomes in Brunetti's set, a logical next step would be to investigate more genomes. A set consisting of one model species from each genus in the poxvirus family may be better suited to identifying regions with functions that are conserved within the whole family, although it may be best if the expansion of the set be limited to genera that share relatively high degrees of sequence similarity. For instance, including GC-rich genomes in the set, such as those named in section 1.1.2, might skew the results since these share low sequence similarity with most members of the poxvirus family.

Similar analysis can also be conducted on different virus families in order to identify sequences that may have very different family-specific conserved functions. Coronaviruses, for instance, contain a ribosomal signal in the DNA sequence between two open reading frames. This signal forms a pseudoknot in the DNA, causing the ribosome to shift frames and translate a different ORF. Applying JaPaFi to a set of coronaviruses might identify other DNA sequences with conserved functions.

Moreover, applying JaPaFi to a set of genomes including viruses from different families may lead to the identification of sequences with highly conserved virus-wide functions.

4.2.2. Signal vs. Noise

An obvious question for future work is that of signal vs. noise, and the development of a statistical measure along the same lines as an e-value in BLAST that indicates the likelihood that a pattern identified be a conserved sequence indicative of a conserved function (signal) as oppose to a pattern occurring in different locations of various genomes as a result of chance that does not reflect a conserved function (noise). A first step towards a statistical assessment of the significance of hits goes back to the idea of an expected level of conservation, as in the comparison of the conservation scores of hits to the scores of a set of promoters selected as controls. The question becomes “what are the chances that a particular sequence in one virus is conserved in a particular set of related sequences”. We began creating a mathematical model

that would allow us to predict how many hits of a certain length and number of differences can be expected given that we know how closely related the genomes are. Essentially, we wanted to address the question of:

> ***How many sequences of length n would you expect to be conserved with k differences between a set of genomes that vary to x degree?***

This problem can be treated as a Bayesian probability problem, which takes the form of *what is the probability of A given B.* Creating this model requires a measurement of how related a set of more than 2 genomes is overall. This is different from pairwise percent identity measures. A program was written by Daniel Godlovitch, a PhD candidate in the Mathematics and Statistics Department at the University of Victoria, which achieves this (Appendix B). The program creates a consensus based on an alignment of all genomes in the set, then counts the number of positions where each genome varies from the consensus, giving a measure of the similarity of the genomes.

However, as we would require a way of converting what we know of how many hits are obtained from different parameter values, and our sense of when the number of hits shows that the hits may be significant and when there is an inundation of hits that are likely noise, it quickly became apparent that even this approach is reliant on a statistical measure like an e-value for discerning signal from noise. The development of a statistic of significance for these hits therefore must be postponed until a successor with more mathematical training can do so as a follow-up to the discussion of signal vs. noise in section 3.3.

Alternately, a conservation score could be calculated as outlined in section 3.6 for the full multiple sequence alignment of the genomes in the set, less the gapped regions at the genome extremities. This would give an average position score for the whole genome alignment representing the expected degree of conservation for these genomes, and sequences within these genomes that score higher than this score could be deemed unusually well conserved. Since the calculation of position scores can be tedious, especially for sequences of length in the scale of full poxvirus genomes, an algorithm would be needed to calculate these scores based on a multiple sequence alignment of the genomes.

5. Bibliography

1. Ahn, B.Y., Gershon, P.D., Moss, B. (1994). RNA Polymerase-associated ProteinR ap94 Confers Promoter Specificity for Initiating Transcription of *Vaccinia virus* Early Stage Genes.*Journal of Biological Chemistry***269**(10): 7552-7557.

2. Amegadzie, B. Y., Ahn, B. Y., Moss, B. (1992). Characterization of a 7-kilodalton subunit of *Vaccinia virusVaccinia virus* DNA-dependent RNA polymerase with structural similarities to the smallest subunit of eukaryotic RNA polymerase-II. *Journal of Virology***66**(5): 3003-3010.

3. Assarsson, E., Greenbaum, J.A., Sundstrom, M., Schaffer, L., Hammond, J.A., Pasquetto, V., Oseroff, C., Hendrickson, R.C., Lefkowitz, E.J., Tscharke, D.C., Sideny, J., Grey, H.M., Head, S.R., Peters, B., Sette, A. (2008). Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes. *Proceedings of the National Academy of Science***105**(6): 2140-2145.

4. Bailey, T.L., Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in miopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*: 28-36.

5. Bailey, TL, Williams, N., Misleh, C., Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research***34**: W369-W373.

6. Baldick, C.J. Jr., Keck, J.G., Moss, B. (1992). Mutational Analysis of the Core, Spacer, and Initiator Regions of *Vaccinia virusVaccinia virus* Intermediate-Class Promoters. *Journal of Virology***66**(8): 4710-4719.

7. Barrett, J.W., Sun, Y., Nazarian, S. H., Belsito, T.A., Brunetti, C.R., McFadden, G. (2006). Optimization of codon usage of poxvirus genes allows for improved transient expression in mammalian cells. *Virus Genes***33**(1): 15-26.

8. Barsky, M., Stege, U., Thomo, A., Upton, C. (2006). A New Algorithm for Fast All-Against-All Substring Matching. *Lecture Notes in Computer Science, Proc. of the 13th Symposium on String Processing and Information Retrieval (SPIRE'06),* **4209/2006**: 360-366.

9. Barsky, M. (2006). All-against-all Approximate Substring Matching, M.Sc. Thesis, University of Victoria, Department of Computer Science.

10. Bhattacharya, S., Dasgupta, R. (2009). A TALE OF TWO GLOBAL HEALTH PROGRAMS Smallpox Eradication's Lessons for the Antipolio Campaign in India. *American Journal of Public Health* **99**(7): 1176-1184.

11. Blanton, J. D., Self, J., Niezgoda, M., Faber, M., Dietzchold, B., Rupprecht, C. (2007). Oral vaccinatin of raccoons (*Procyon lotor*) with genetically modified rabies virus vaccines.

*Vaccine***25**(42): 7296-7300.

12. Brodie, R., Smith, A.J., Roper, R.L., Tcherepanov, V., Upton, C. (2004). Base-by-base: Single nucleotide-level analysis of whole viral genome alignments. *BMC Bioinformatics***5**(96).

13. Broyles, S., Fesler, B.S. (1990). *Vaccinia virusVaccinia virus* gene encoding a component of the viral early transcription factor. *Journal of Virology***64**(4): 1523-1529.

14. Broyles, S. (2003). *Vaccinia virusVaccinia virus* transcription. *Journal of General Virology***84**: 2293-2303.

15. Brunetti, C.R., Amano, H., Ueda, Y., Win, J., Miyamura, T., Suzuki, T., Li, X., Barrett, J.W., McFadden, G. (2003). Complete Genomic Sequence and Comparative Analysis of the Tumorigenic Poxvirus Yaba Monkey Tumor Virus. *Journal of Virology***77**(24): 13335-13347.

16. Chung, C.S., Chen, C.H., Ho, M. Y., Huang, C. Y., Liao, C. L., Chang, W. (2006). *Vaccinia virusVaccinia virus* proteome: identification of proteins in *Vaccinia virusVaccinia virus* intracellular mature virion particles. *Journal of Virology***80**(5): 2127-40.

17. Cochran, M. A., Puckett, C., Moss, B. (1985). In vitro mutagenesis of the promoter region for a *Vaccinia virusVaccinia virus* gene: evidence for tandem early and late regulatory signals. *Journal of Virology***54**(1): 30-37.

18. Collier, A., Hu, S.L., Coombs, R., Arditti, D., Corey, L. (1989). Clinical and virologic responses to a recombinant vaccinia HIV-1 GP160 vaccine (HIVAC-1e). *International Conference on AIDS*: 543.

19. Condit, R.C. (2007). Vaccinia, Inc. - Probing the Functional Substructure of Poxviral Replication Factoires. *Cell Host & Microbe.***2**(4): 205-207.

20. Coupar, B.E., Boyle, D., Both, G. (1987). Effect of in vitro Mutations in a *Vaccinia virusVaccinia virus* Early Promoter Region Monitored by Herpes Simplex Virus Thymidine Kinase Expression in Recombinant *Vaccinia virusVaccinia virus*. *Journal of General Virology***68**: 2299-2309.

21. Crooks, G.E., Hon, G., Brenner, J.M., Chandonia, S.E. (2004). Weblogo: A Sequence Logo Generator. *Genome Research***14**(6): 1188-1190.

22. Da Silva, M., Upton, C.(2005). Using purine skews to predict genes in AT-rich poxviruses. *BMC Genomics.***6**(22): 22.

23. Da Silva, M., Upton, C.(2009).*Vaccinia virus* G8R Protein: A Structural Ortholog of Proliferating Cell Nuclear Antigen (PCNA). *PLoS ONE***4**(5).

24. Davison, A.J., Moss, B.(1989). Structure of *Vaccinia virusVaccinia virus* Late Promoters. *Journal of Molecular Biology***210**(4): 771-784.

25. Davison, A.J.(1989). Structure of *Vaccinia virusVaccinia virus* Early Promoters. *Journal of Molecular Biology*. **210**(4): 749-769.

26. deCastro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Petra, S., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acid Research*.**34**: W362-W365.

27. Earl, P.L.Cotter, C., Moss, B., VanCott, T., Currier, J., Eller, L., McCutchan, F., Birx, D.L., Michael, N. L., Marovich, M.A., Robb, M., Cox, J.H. (2009). Design and evaluation of multi-gene, multi-clade HIV-1 MVA vaccines. *Vaccine***27**(42)5885-5895.

28. Eaton, H.E., Metcalf, J., Brunetti, C.R. (2008). Characterization of the promoter activity of a poxvirus conserved element. *Canadian Journal of Microbiology***54**(6): 483 - 488.

29. Esteban, D.J., Da Silva, M., Upton, C. (2005). New bioinformatics tools for viral genome analyses at Viral Bioinformatics - Canada. *Pharmacogenomics***6**(3)271-280.

30. Ferrier-Rembert, A., Drillien, R., Tournier, J., Garin, D., Crance, J. (2008). Short- and long-term immunogenicity and protection induced by non-replication smallpox vaccine candidates in mice and comparison with the traditional 1st generation vaccine. *Vaccine***26**(14): 1794-1804.

31. Fick, W.C., Viljoen, G.J.(1999). Identification and characterisation of an early/late bi-directional promoter of the capripoxvirus, lumpy skin disease virus. *Archives of Virology***144**(6): 1229-1239.

32. Garza, N.L. Hatkin, J. M., Livingston, V., Nichols, D.K., Chaplin, P.J., Volkmann, A., Fisher, D., Nalca, A. (2009). Evaluation of the efficacy of modified vaccinia Ankara (MVA)/IMVAMUNE against aerosolized rabbitpox virus in a rabbit model. *Vaccine***27**(40): 5496-5504.

33. Goebel, S.J., Johnson, G.P., Perkus, M.E., Davis, S.W., Winslow, J.P., Paoletti, E. (1990). The complete DNA sequence of *Vaccinia virusVaccinia virus*. *Virology***179**(1): 517-63.

34. Guarner, J., Johnson, B.J., Paddock, C.D., Shieh, W., Goldsmith, C.S., Reynolds, M.G., Damon, I.K. Regnery, R.L., Zaki, S.R. (2004). Monkeypox Transmission and Pathogenesis in Prairie Dogs. *Emerging Infectious Diseases***10**(3).

35. Hardison, R.C. (2003). Comparative Genomics. *PLoS Biology***1**(2): 156-160.

36. Heuser, J. (2005). Deep-etch EM reveals that the early poxvirus envelope is a single membrane bilayer stabilized by a geodetic honeycomb surface coat. *The Journal of Cell Biology***169**(2): 269-283.

37. Jacobs, B.L., Jeffrey, O.L., Karen, V.K., Karen, L.D., Stacy, D.W., Susan, A.H., Shukmei, W., Trung, H., Carole, R.B. (2008). *Vaccinia virusVaccinia virus* Vaccines: Past, Present and Future. *Antiviral Research***84**(1), doi:10.1016/j.antiviral.2009.06.006.

38. Katsafanas, G.C., Moss, B. (2007). Linkage of transcription and translation within cytoplasmic poxvirus DNA factories provides a mechanism to coordinate viral and usurp host functions. *Cell Host Microbe***2**(4): 221-228.

39. Knutson, B.A., Liu, X., Oh, J., Broyles, S.S. (2006). *Vaccinia virusVaccinia virus* intermediate and late promoter elements are targeted by the TATA-binding protein. *Journal of Virology***80**(14): 6784-6793.

40. Langner, K.D., Vardimon, L., Renz, D., Doerfler, W. (1984). DNA methylation of three 5' C-C-G-G 3' sites in the promoter and 5' region inactivate the E2a gene of adenovirus type 2. *Proceedings of the National Academy of Science***81**(10): 2950-2954.

41. Li, J., Broyles, S.S. (1993). Recruitment of *Vaccinia virusVaccinia virus* RNA Polymerase to an Early Gene Promoter by the Viral Early Transcription Factor. *Journal of Biological Chemistry***268**(4): 2773-2780.

42. Lun, X., Yang, W., Alain, T., Shi, Z., Muzik, H., Barrett, J.W., McFadden, G., Bell, J., Hamilton, M. G., Senger, D.L., Forsyth, P. A. (2005). Myxoma virus is a novel oncolytic virus with significant antitumor activity against experimental human gliomas. *Cancer Research***65**(21): 9982-9990.

43. Meseda, C.A., Mayer, A.E., Kumar, A., Garcia, A.D., Campbell, J., Listrani, P., Manischewitz, J., King, L.R., Golding, H., Merchlinsky, M., Weir, J.P. (2009). Comparative Evaluation of the Immune Responses and Protection Engendered by LC16m8 and Dryvax Smallpox Vaccines in a Mouse Model. *Clinical and Vaccine Immunology***16**(9): 1261-1271.

44. Mohamed, M.R.Niles, E.G. (2003). UUUUUNU Oligonucleotide Stimulation of *Vaccinia virus* Early Gene Transcription Terminatin, in trans. *Journal of Biological Chemistry***278**(41).

45. Moss, B., Ahn, B., Amegadzie, B., Gershon, P.D., Keck, J.G. (1991). Cytoplasmic Transcription System Encoded by *Vaccinia virus*. *The Journal of Biological Chemistry.***266**(3): 1355-1358.

46. Olafsdottir, G., Svansson, V., Ingvarsson, S., Marti, E., Torsteinsdottir, S. (2008). In vitro analysis of expression vectors for DNA vaccination of horses: the effect of a Kozak sequence. *Acta Veterinaria Scandinavica***50**(1): 44.

47. Osorio, J. E., Iams, K.P., Meteyer, C.U., Rocke, T.E. (2009). Comparison of Monkeypox Viruses Pathogenesis in Mice by In Vivo Imaging. *PLoS One***4**(8). e6592.

48. Parrino, J., Graham, B.S.(2006). Smallpox vaccines: Past, present, and future. *Journal of Allergy and Clinical Immunology***118**(6): 1320-1326.

49. Roscoe, D.E., Holste, W.C., Sorhage, F.E., Campbell, C., Niezgoda, M., Buchannan, R., Diehl, D., Niu, H.S., Rupprecht, C.E. (1998). Efficacy of an oral vaccinia-rabies glycoprotein recombinant vaccine in controlling epidemic raccoon rabies in New Jersey. *Journal of Wildlife Diseases***34**(4): 752-763.

50. Sanchez-Puid, J.M., Blasco, R. (2008). AUG context and mRNA translation in *Vaccinia virus*. *Spanish Journal of Agricultural Research***6**: 73-80.

51. Shen, Y., Nemunaitis, J. (2005). Fighting Cancer with *Vaccinia virus*: Teaching New Tricks to an Old Dog. *Molecular Therapy***11**(2): 180-195.

52. Shuman, S. (1998). *Vaccinia virus* DNA topoisomerase: a model eukaryotic type IB enzyme. *Biochimica et biophysica acta***1400**(1-3): 321-37.

53. Tartaglia, J., Perkus, M. E., Taylor, J., Norton, E. K., Audonnet, J. C., Cox, W. I., Davis, S. W., van der Hoeven, J., Meignier, B., Riviere, M. (1992). NYVAC: a highly attenuated strain of *Vaccinia virus*. *Virology***188**(1): 217-32.

54. Upton, C., Slack, S., Hunter, A.L., Ehlers, A., Roper, R.L. (2003). Poxvirus Orthologous Clusters: toward Defining the Minimum Essential Poxvirus Genome. *Journal of Virology***77**(13): 7590-7600.

55. Upton, C., Hogg, D., Perrin, D., Boone, M., Harris, N. (2001). Viral Genome Organizer: A system for analyzing complete viral genomes. *Virus Research***70** (1-2): 55-64.

56. Vos, J. C., Stunnenberg, H.G. (1988). Derepression of a novel class of *Vaccinia virus* genes upon DNA replication. *The EMBO Journal***7**(11): 3487–3492.

57. Wallace, J.C., Henikoff, S. (1992). PATMAT: a searching and extraction program for sequence, pattern and block queries and databases. *Computer applications in the biosciences***8**(3): 249-254.

58. Wittek, R., Menna, A., Muller, H.K., Schumperli, D., Boseley, P.G., Wyler, R. (1978). Inverted Terminal Repeats in Rabbit Poxvirus and *Vaccinia virus* DNA. *Journal of Virology***28**(1): 171-181.

59. World Health Organization. Smallpox Fact Sheet - Historical Significance. *World Health Organization.* 2009 йил 05-07 <http://www.who.int/mediacentre/factsheets/smallpox/en/index.html>.

60. Yu, Z., Li, S., Brader, P., Chen, N., Yu, Y.A., Zhang, Q., Szalay, A.A., Fong, Y., Wong, R.J. (2009). Oncolytic vaccinia therapy of squamous cell carcinoma. *Molecular Cancer***8**(45).

6. Appendices

6.1. Appendix A

DNA and protein alignments of corresponding regions in the Morphogenesis/Viral Early Transcription Factor gene, *m081R.* These demonstrate that conservation at the protein level does not necessitate high conservation at the DNA level. Also shown are protein and DNA alignments of hits 05 and 06 and that the coding region hits, where conservation is observed at both the protein and DNA sequence level, are, therefore, unusual.

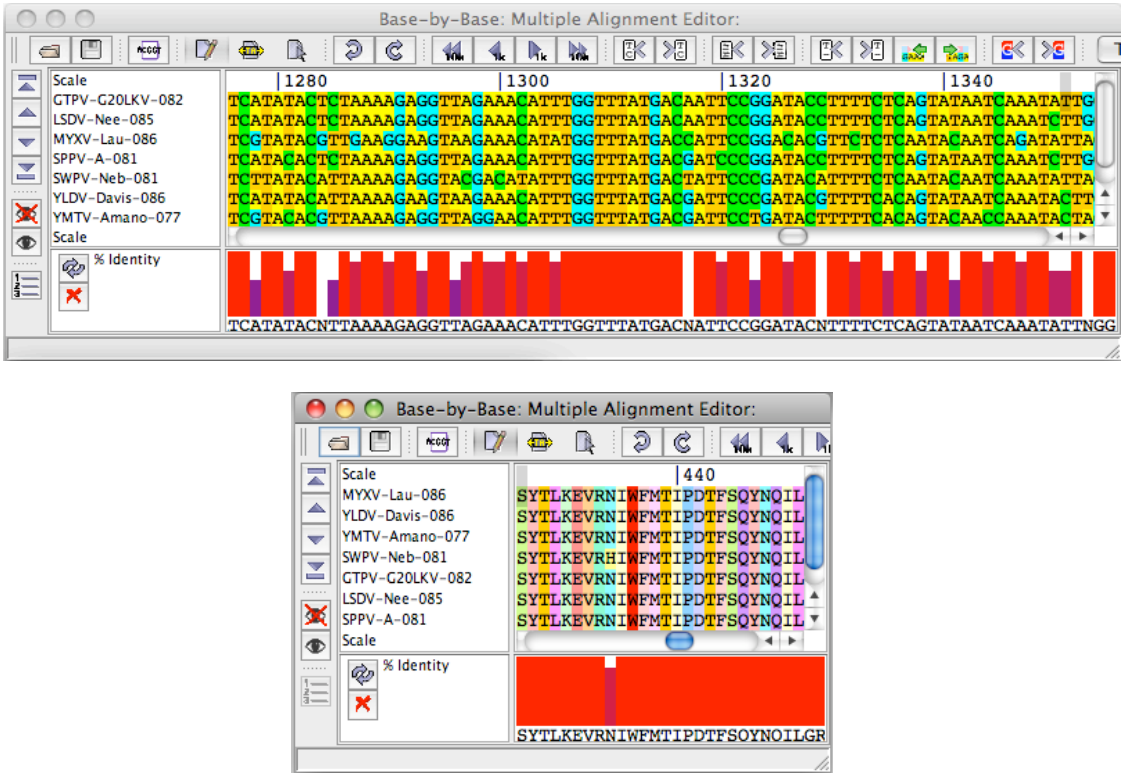


Figure 6-1 DNA and protein alignments of a superconserved region in the VETF gene.

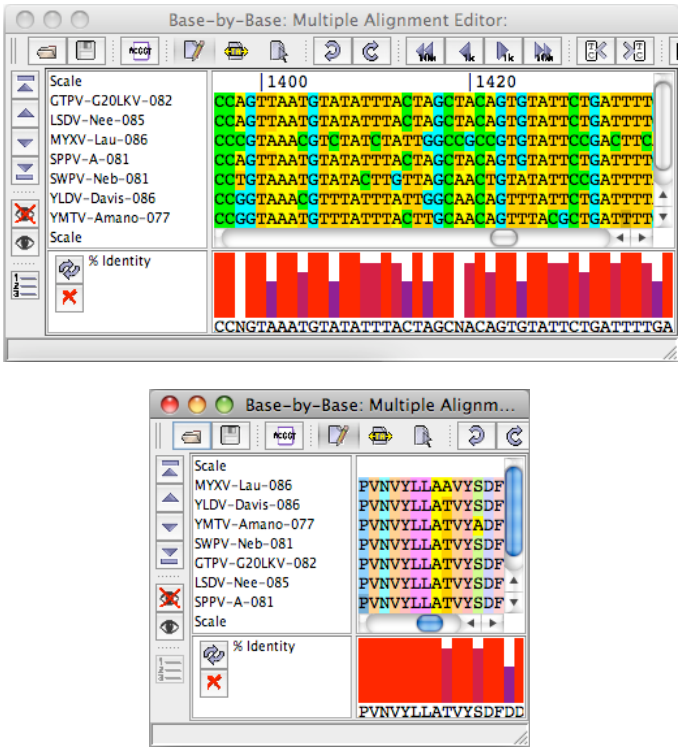


Figure 6-2 DNA and protein alignments of a superconserved region in the VETF gene.

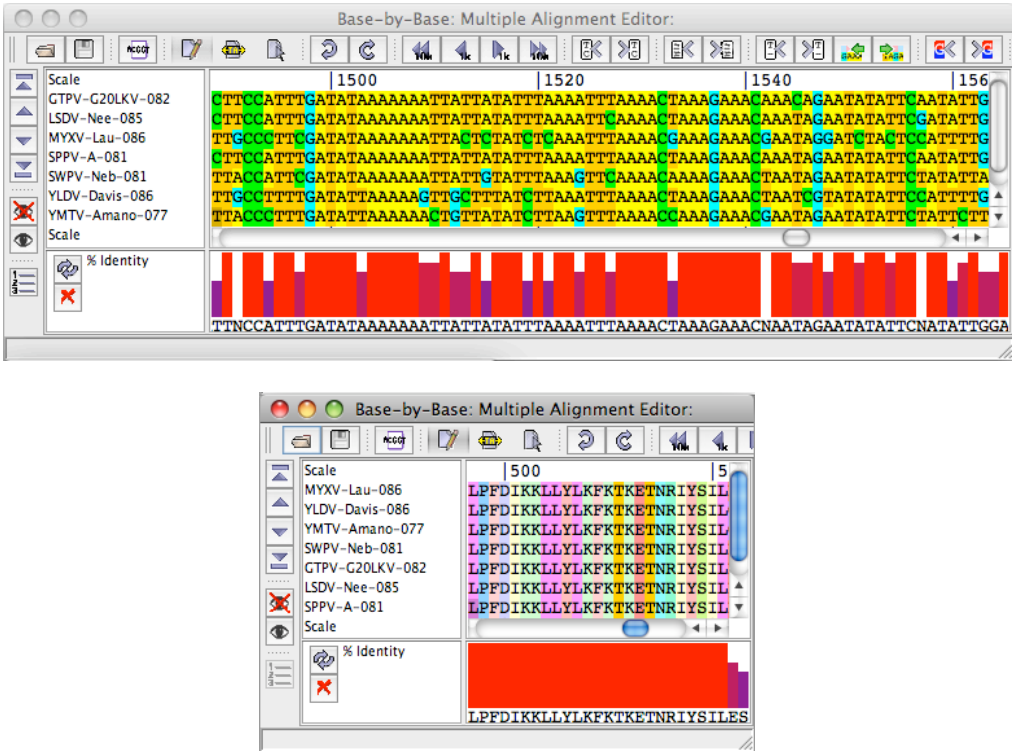


Figure 6-3 DNA and protein alignments of a superconserved region in the VETF gene.

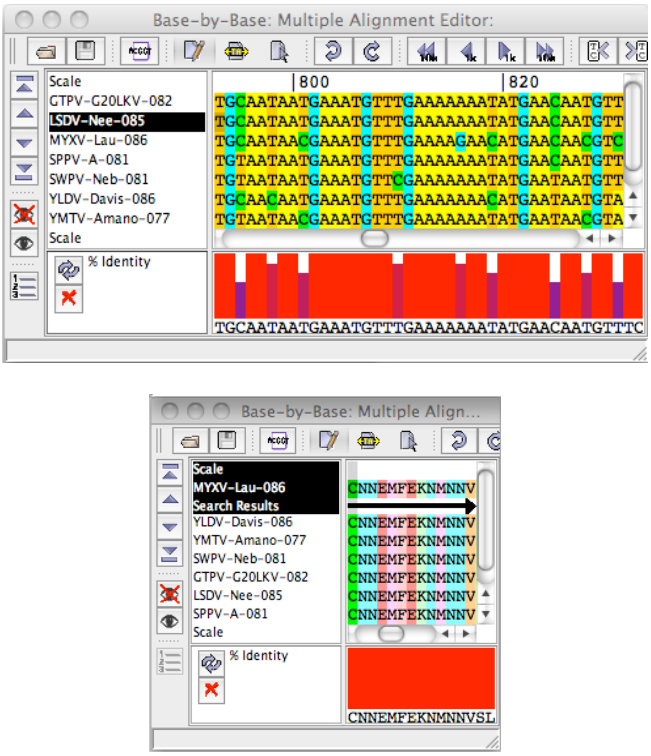


Figure 6-4 DNA and protein alignments of hit 05.

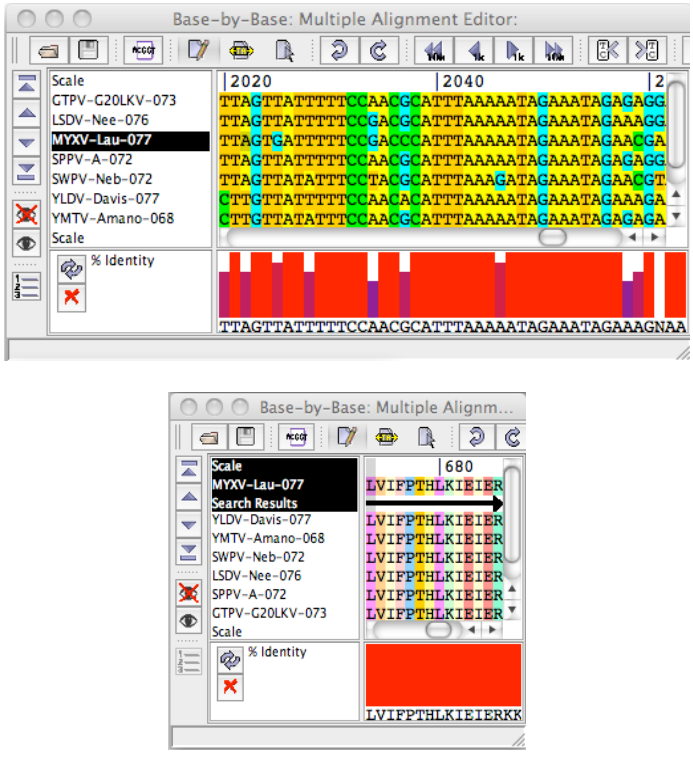


Figure 6-5 DNA and protein alignments of hit 06.

6.2. Appendix B: In-house script for extracting character heights from Weblogo

The Weblogo application was downloaded as an executable application and run locally. The output was aliased as “zork.fasta” and the following script was given zork.fasta as input and used to extract the character heights and display them in text format.

```
#!/usr/bin/perl -Wall

use logo;

open (FASTA, "zork.fasta");
my @inputdata = <FASTA>;
close (FASTA);

my %stuff = (input_kind => 1,
             smallsamplecorrection => 0,
             stretch => 0);

my ($hd,$de,$kd) = logo::getHeightData(\@inputdata,\%stuff);

my $i;
print "zork ", scalar @$hd;
for ($i=0;$i<scalar @$hd;$i++) {
  my $j;
  print "$i: ", scalar @{$hd->[$i]};
  for ($j=0;$j<scalar @{$hd->[$i]};$j++) {
    print $hd->[$i]->[$j];
  }
}
```

6.3. Appendix C: AGS program for measuring genome similarity

The AGS program ("Aliya's Gene Sequence" program), written by Daniel Godlovitch, builds a consensus based on an alignment of all genomes in the set. It iterates through the alignment and at each position, adds the most frequently occurring nucleotide to the consensus. If there are two nucleotides that appear with the same frequency, both are added to the consensus for that position. The number of positions where each genome varies from the consensus is then counted and the sum or average of these numbers reflects the similarity of the genomes.

```
Ancestor="-"; %Declare an empty string to make the ancestor sequence
from
SS=size(Seq);
Count=zeros(1,4);
for i=1:SS(2)
 Count=zeros(1,4); %This is the vector of one base pair from all the
different genes
 Countstring="ACGT";
 Temp=Seq(:,i); %Pick the column of Seq for comparison
  for j=1:SS(1) %Counts the number of each nucleotide
   switch(Temp(j)) %Adds one to the appropriate element of Count, using
the switch command
    case "A"
     Count(1)=Count(1)+1;
    case "C"
     Count(2)=Count(2)+1;
    case "G"
     Count(3)=Count(3)+1;
    case "T"
     Count(4)=Count(4)+1;
   endswitch
  end
 Base="X";
 for k=1:4
  if (Count(k)==max(Count))
   Base=[Base Countstring(k)];
  end
 end
 Base=Base(2:length(Base)); %Vector of most common n-tides
```

```
 Nn=length(Base)*rand;
 for i=1:length(Base)
  if i>Nn %pick one of them at random
   Ancestor=[Ancestor Base(i)]; %append it to Ancestor
   break
  end
 end
end

Ancestor=Ancestor(2:length(Ancestor))
error=zeros(1,SS(1));
for i=1:SS(1)
 for j=1:SS(2)
  if(Seq(i,j)~=Ancestor(j))
   error(i)=error(i)+1;
  end
 end
end
error
```

It should be noted that this program requires that all genomes be the same length, so member genomes must be truncated. It is yet to be determined where the appropriate place to truncate would be. One option is to truncate all genomes to the length of the shortest genome (Figure 6-6 A), while another option is to exclude all ITRs (Figure 6-6 B).

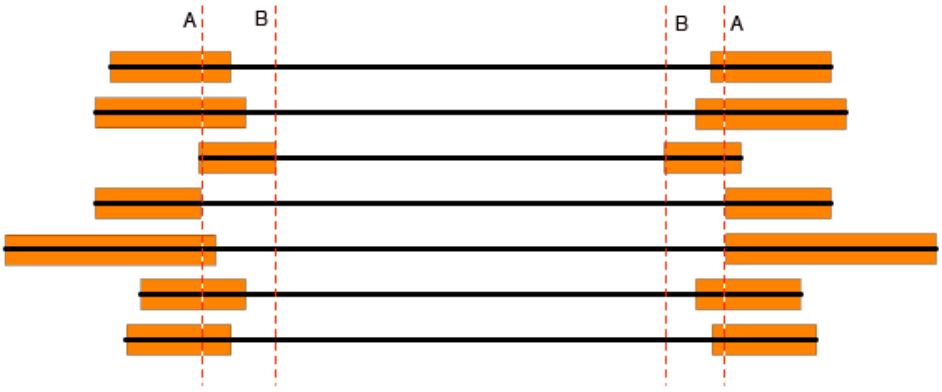


Figure 6-6 Places to truncate genomes for AGS program.