2010

Characterization and Evolution of the SerH Immobilization Antigen Genes in Tetrahymena Thermophila

Christopher Thomas Mcginness
Cleveland State University

CHARACTERIZATON AND EVOLUTION OF THE *SER*H IMMOBILIZATION ANTIGEN GENES IN *TETRAHYMENA THERMOPHILA*

CHRISTOPHER THOMAS MCGINNESS

Bachelor of Arts in Classical Languages and Philosophy

Xavier University

May, 2006

submitted in partial fulfillment of requirements for the degree

MASTER OF SCIENCE IN BIOLOGY

at the

CLEVELAND STATE UNIVERSITY

May, 2010

This thesis has been approved

for the Department of BIOLOGY GEOLOGY AND ENVIRONMENTAL SCIENCES

and the College of Graduate Studies by

_____

Thesis Chairperson, Dr. F. Paul Doerder

_____

Department & Date

_____

Dr. Robert Krebs

_____

Department & Date

_____

Dr. Girish Shukla

_____

Department & Date

ACKNOWLEGMENTS

I wish to extend my appreciation and thanks to my advisor, Dr. F. Paul Doerder, without whose guidance and assistance this would not have been possible.

I also wish to thank Dr. Robert Krebs and Dr. Girish Shukla for their contributions and Dr. Crystal Weyman for her advice and the experience I gained from her lab.

CHARACTERZATION AND EVOLUTION OF THE *SER*H IMMOBILIZATION ANTIGEN GENES IN *TETRAHYMENA THERMOPHILA*

CHRISTOPHER THOMAS MCGINNESS

ABSTRACT

Tetrahymena thermophila express a major cell surface protein known as the immobilization (i-) antigen which coats the entire cell surface including the cilia. *T. thermophila* has several different i-antigens which are expressed in a mutually exclusive manner under varying environmental conditions. When the genes for these i-antigens were sequenced it was found that they encode proteins consisting of imperfect repeats with cysteine periodicity. The best characterized of these i-antigens are those specified by alleles at the *Ser*H i-antigen locus. The H proteins all contain a section of 8-cysteine containing imperfect repeats. The presence of repeats allows for the possibility that *Ser*H genes evolve, at least in part, by concerted evolution, a process in which the repeats of the gene evolve together so that the repeats within one sequence would be more similar to each other than they would be to the repeats of any other sequence. A previous study (Mol. Biol. Evol. **23**: 608-614) found evidence that *Ser*H genes evolve via a mix of vertical transmission and concerted evolution. This study characterized 20 *Ser*H alleles from wild samples and further explored the mode of evolution of *Ser*H i-antigen. Using bioinformatic tools, *Ser*H alleles were characterized with respect to nucleotide diversity, repeat structure, codon usage, and sequence evolution. The encoded proteins were examined for amino acid composition, cysteine periodicity, and potential secondary structure. A model of the i-antigen structure was presented. Standard bioinformatic tests for evolution provided no evidence that *Ser*H genes are positively selected. Neighbor-

joining trees of the 8 cysteine-containing confirm that *Ser*H genes evolve through a mix of primarily vertical transmission and concerted evolution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGRURES

Figure Page

CHAPTER I

INTRODUCTION

Free living ciliates such as *Paramecium* and *Tetrahymena* express a major cell surface protein known as the immobilization (i-) antigen which coats the entire cell surface including the cilia (Beale and Kacser 1957). These i-antigens were first discovered in *Paramecium*, then in *Tetrahymena* and other ciliates (Beale and Kacser 1957; Nanney 1960). These proteins are known as immobilization antigens because they were historically studied using a simple assay in which antibody mixed with living cells caused them to cease swimming. By assaying various wild isolates and inbred strains grown under different environmental conditions, it was found that cells possess an array of i-antigens, each detectable by a specific antibody (Nanney 1960). The i-antigen present was said to be the serotype. Using antibodies to biochemically purified antigen, Smith et al (1992) showed that the various i-antigens are usually present as a single molecular species, consistent with original observations that i-antigens are expressed in a mutually exclusive manner. Because 1) they are in direct contact with the environment, 2) they are so variable and 3) cells are never without i-antigen, they are likely important molecules. The function of i-antigens is unknown.

With the advent of recombinant DNA technology, i-antigen genes of *Paramecium* and then *Tetrahymena* were cloned and sequenced. The most striking similarity between the different i-antigens of different ciliate species is that they are composed of imperfect repeats containing periodic cysteines. This includes i-antigens of *Paramecium* (Prat et al. 1986; Nielsen et al. 1991), *Tetrahymena* (Deak and Doerder 1995; Tondravi et al. 1990), the fish parasite *Ichthyophtherius* (Clark et al. 1995) and the predatory ciliate *Lembadion bullinum* (Peters-Regehr, Kusch, and Heckmann 1997), as well as the diplomonad parasite *Giardia lamblia* (Gillin et al. 1990) and the parasitic fungus *Pneumocystis carinii* (Kovacz et al. 1993). This structure of imperfect repeats containing periodic cysteines seems to be conserved among surface antigens (Simon and Schmidt 2006) and will be discussed in greater detail below as it suggests a common tertiary structure.

The i-antigens of *Paramecium* and *Tetrahymena* are similarly rich in alanine, cysteine, serine and threonine (Gerber et al. 2002) and are attached to the outer membrane by GPI linkage (Azzouz et al. 1990, Ron et al 1992). Eleven serotypes (i-antigens) have been described for *Paramecium tetraurelia* stock 51. The genes encoding these i-antigens have been sequenced (Preer et al. 1985; Nielson et al., 1991; Scott et al., 1993; Breuer et al., 1996) and encode large molecules of 2100-2700 amino acids with molecular weights on reducing SDS gels of 250–300 kDa (Forney et al. 1983; Hansma and Kung 1975). When the first i-antigen gene of *Paramecium* (the gene encoding serotype 51A) was sequenced (Preer et al. 1985), it was found to encode a protein having 37 imperfect repeats each containing eight cysteines. The sequence also showed that *Paramecium* do not use the standard genetic code. In both *Paramecium* and *Tetrahymena*

(Horowitz and Gorovsky, 1985) traditional stop codons UUG and UUA specify glutamine.

T. thermophila i-antigen genes encode much smaller antigens ranging from 25 to 59 kDa (Gerber et al. 2002; Smith et al. 1992). Whereas the larger i-antigens from *Paramecium* contain 30-37 imperfect repeats marked by eight periodic cysteines (Nielson et al., 1991; Scott et al., 1993; Breuer et al., 1996) the smaller *T. thermophila* H i-antigens contain 3.5 imperfect repeats (three full repeats and one half repeat)with eight periodic cysteines. As with the surface proteins of many parasitic protists mentioned above (Kusch and Schmidt, 2001), the *T. thermophila* genes encode modular proteins, composed of tandem imperfect repeats containing even numbers of periodic cysteines, specifically 6, 8, 10 or 12 (Table I).

As in *Paramecium*, *T. thermophila* is capable of expressing multiple i-antigens. These include the L, H, T, J, K, I, S, M, and P i-antigens which are expressed under varying environmental conditions (Table I) (Smith et al., 1992; Saad and Doerder 1995; Doerder personal communication). Each of these variant surface proteins is specified by genes at unlinked loci (Doerder and Berkowitz, 1986).

Table I: Properties of the Various Antigens of *T. thermophila*

Antigen	Condition of Expression	No. of Isoforms	Molecular Weight	Cysteines Per Repeat
H1	20 – 36°C	1	52,000	8
H2	20 – 36°C	1	44,000	8
H3	20 – 36°C	1	52,000	8
H4	20 – 36°C	1	49,000	8
H5	20 – 36°C	1		8
H6	20 – 36°C	1		8
J	20 – 36°C	1		10
K	20 – 36°C	1		unknown
T1	> 36°C	1	25,000	unknown
T2	> 36°C	1	25,000	unknown
T3	> 36°C	5-8	36,000 – 52,000	unknown
S	0.2M NaCl in medium	1	50,000	unknown
I	antibody in medium	2 (4)	30,000 – 32,000 56,000 – 59,000	12
L	< 20°C	3-5	41,000 – 52,000	6
M	20 - 36°C mutant	1	51,000	unknown
P	20 - 36°C mutant	1	53,000	unknown

Sources: (Doerder and Berkowitz 1986; Smith et al. 1992; Saad and Doerder 1995; Doerder 2000 Gerber et al. 2002; Doerder and Gerber 2000; Doerder 2000; Doerder unpublished)

Various temperature conditions induce the expression L, T, and H (J and K). The *Ser*L genes are expressed at temperatures 20°C (Juergensmeyer, 1969). The *Ser*T genes are expressed at temperatures above 36°C (Smith et al. 1992) and have 3 allelic variants (Phillips, 1967). The *Ser*H, *Ser*J and *Ser*K genes are expressed when the *Tetrahymena* are grown in the intervening normal temperature range (Loefer and Owen 1961; Saad and Doerder 1995). The expression of genes at these three loci is mutually exclusive and involves dominant epistasis (Doerder 2000). The stability of mRNA plays a major role in regulating expression of the *Ser*H genes (Love et al. 1988; McMillan et al. 1995; Tondavi et al. 1990). For example, the half life of the H3 encoding mRNA is 1 hour at 30°C, but only 3 minutes at 40°C (Love et al. 1988). The *Ser*S gene is expressed in media

containing 0.2 M NaCl (Grass 1972), and the *Ser*I genes are expressed when H-expressing cells are grown in allele-specific anti-H (Juergensmeyer 1969). The P and M antigens are expressed on the surface of mutant *Tetrahymena* strains (Doerder and Berkowitz 1987). *Ser*H, *Ser*L and *Ser*J genes have been sequenced (Doerder 2000; Doerder and Gerber 2000; Gerber et al. 2002). Gerber used five alleles and Katz used 11 *Ser*H alleles, of which six were atypical. Though the genome of *T. thermophila* has been sequenced, there is only tentative identification of the genes encoding other i-antigens except for *Ser*I. That is to say that there are sequenced areas in the genome whose putative amino acid sequences seem to indicate that they code a surface antigen (Doerder, personal communication). Of the surface antigens, the *Ser*H locus is the most studied and best characterized.

1.1 The *SerH* Gene

The *Ser*H locus is highly polymorphic. Inbred strains constructed by 1960 (Nanney 1960) had four alleles, more than any other genetic locus. Subsequently, the Doerder lab found additional alleles in wild isolates which were studied both by Gerber et al. (2002) and Katz et al. (2006). The study by Katz (2006) formed the basis of this project. That study utilized 11 *Ser*H alleles, five of which were considered typical. The *Ser*H alleles encode single polypeptides of about 400-450 amino acids with electrophoretic mobilities of 44,000 - 52,000 and isolectric points of 4.1 to 4.5 (Doerder and Berkowitz 1986). There is relatively little cross-reactivity among the antisera for the *Ser*H alleles (Loefer and Owen 1961), except anti-H1 and H3 and anti-H3 and H1 occasionally cross react (Doerder and Berkowitz 1986). Also, cells expressing *Ser*H6 are

recognized by both anti-H1 and anti-H3 sera (Gerber et al. 2002). The H proteins typically consist of 3.5 imperfect repeats each containing 8 periodic cysteines (Fig. 1, Fig. 2). They also contain a leader sequence, which contains 10 cysteines and an ER translocation signal, as well as a carboxyl terminus of about 27 amino acids containing the GPI addition site.

*Ser*H alleles from *T. thermophila* populations showed considerable sequence variation (Gerber et al. 2002); however, the overall protein structure, the hydrophobicity, and the cysteine periodicity were highly conserved. The *Ser*H alleles also showed ratios of nonsynonymous to synonymous amino acid substitution rates that were < 1 in pairwise maximum likelihood tests suggesting that these alleles are not under any positive selection (Gerber et al. 2002). The same study also performed both Tajima (1989) and Fu and Li (1993) tests for selection. Both tests showed no evidence of positive selection. These results were interpreted to mean that *Ser*H alleles are selectively neutral, with deleterious alleles removed by purifying selection (Gerber et al. 2002).

Fig. 1 Overall Structure of the *Ser*H Gene

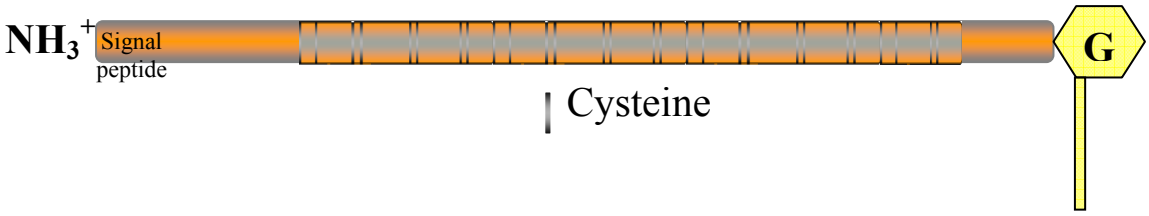


Fig 2a Amino Acid Sequence for a *Ser*H3 Allele Showing 3.5 Repeats and Periodic Cysteines

*Ser*H3 (Genbank AF190647.1 Deduced Amino Acid Sequence)

MQNKTIII CLIISQLLVSVFSSAGGQAN CTGVAAGTD CASV CGVPTVAGTGTTA CSWVSSSTLTT CTVTD CT CLTTGTVTGITNLNDQF CTS CKGSTSNTYANGAGTA

CVAASAS CNSTIRGTTA---WTVGD CTV CTPTTPALVGST CKA CNTISSAWTDAN CAA CASTSTPKGNTNFANSAGTA

CVNASAT CASGSRGTTAANAWTVAD CLA CTPATPVFVPAASPAVTTS CVA CSAATSGLNDAL CNA CASSASPAAKTTFANTAGSA

CVASSAT CTAGSRGTTAANAWTAAD CLA CTPATPAVQFGASPATTSS CVA CNTINSGWTDAN CNS CAMAASPQTKNIVAKADGSA

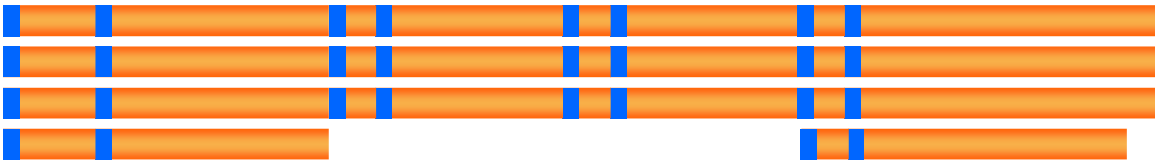
CVAAVFS CTQSARGSNK---WTNAD CAA CNGTAANANQ--YASADGST

CQATQASSTFSGQIFVSILLVLSALLI*

Aligned by the cysteines (highlighted) in the repeats. This particular *Ser*H3 allele has 3.5 imperfect repeats.

Pink – Putative ER Translocation Signal Black – Leader Sequence Red – First Repeat Blue – Second Repeat Green – Third Repeat Orange – Fourth Repeat (Half Repeat) Indigo – Ending Sequence (including GPI linkage site)

Fig 2b Cartoon Representation of Repeats in *Ser*H



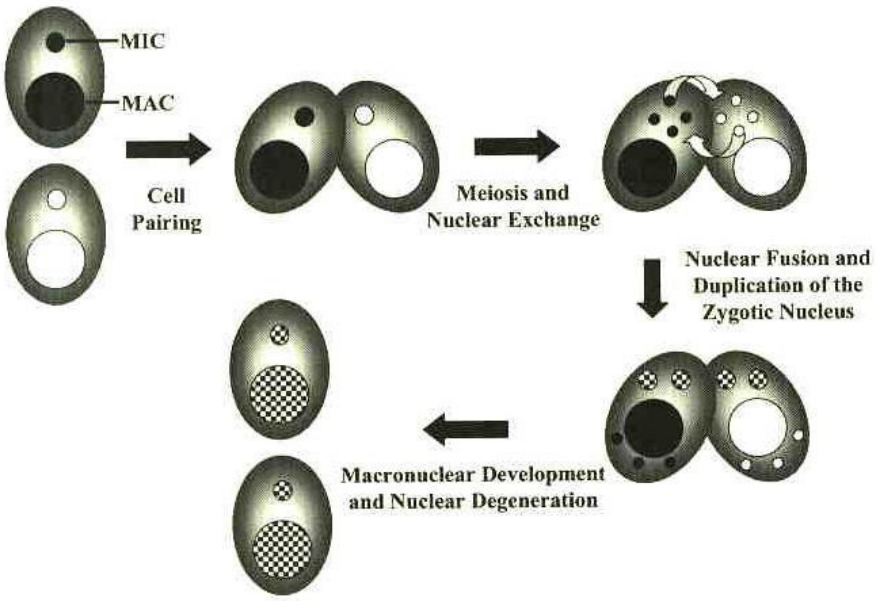
Blue represents the cysteine location. Note that the first part of the final half repeat is most similar to the front quarter of the full repeats and the second part most similar to the last quarter of the full repeats.

1.2 Ciliate Nuclear Dimorphism

In ciliates, including *T. thermophila,* the cell contains two separate nuclei. The germline micronucleus (MIC) is diploid, transcriptionally inactive and gives rise to meiotic products which are exchanged during conjugation. The much larger macronucleus (MAC) is the site of gene expression (reviewed in Jahn and Klobutcher 2002; Yao, Duharcourt and Chalker, 2002; McGrath and Katz, 2004).

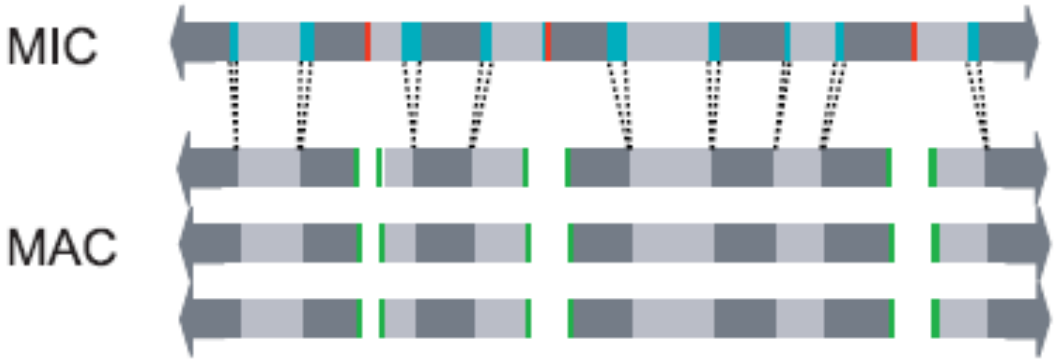
The macronucleus is a rearranged copy of the micronucleus (reviewed in Jahn and Klobutcher 2002). This rearrangement occurs during macronuclear development at conjugation (Fig. 3) and involves many editing processes such as DNA sequence elimination, chromosome fragmentation, addition of telomeres and differential gene amplification (Fig 4). The fragmentation of the five micronuclear chromosomes in *T. thermophila* is controlled by cis-acting elements known as the chromosome breaking sequences, which are flanked by telomere addition zones (reviewed in Jahn and Klobutcher 2002).

Fig. 3 The Process of Conjugation in *T. thermophila*



The process of conjugation in *T. thermophila*. Germ line micronuclei are exchanged at conjugation, the macronucleus is degraded and the fertilization nuclei give rise to a new macronucleus. (Jahn and Klobutcher 2002).

Fig. 4 Relationship between the Macronucleus and the Micronucleus



The five micronuclear chromosomes are split into the approximately 225 maconuclear chromosomes in *T. thermophila* which are maintained at a copy number of ~45. (Eisen et al. 2006). Blue indicates internally deleted sequences, red indicates chromosome breakage sites and green indicates telomeres.

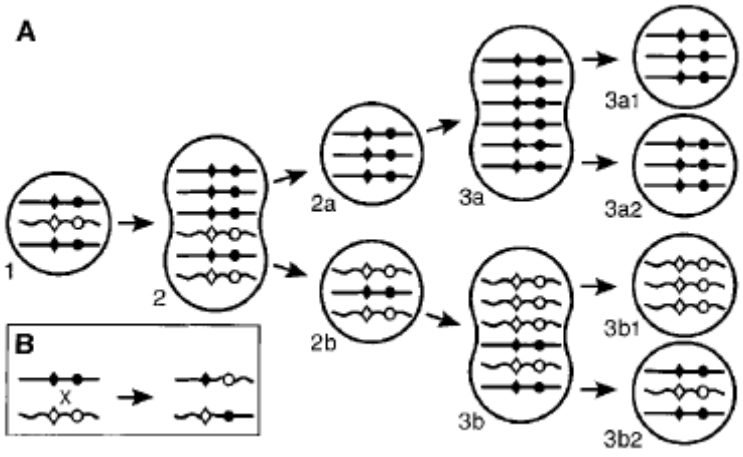
The mature *T. thermophila* macronucleus contains about 45 copies each of ~225 macronuclear chromosomes (Eisen et al. 2006) and ~ 10,000 copies of the ribosomal DNA (rDNA) chromosome Yao et al. 2002). The sequenced MAC genome is estimated to contain 27,000 protein encoding genes (Eisen et al. 2006). Because the total genome is about 22%GC, genes typically are recognized, in part, by their higher %GC reflecting the need to use G and C containing codons.

1.3 Macronuclear Sorting and Recombination

Though the micronucleus is diploid, the much larger macronucleus with ~45 copies of most genes is functionally haploid. This is because during vegetative propagation the macronucleus undergoes what is called phenotypic assortment (Fig 5). It was discovered early that if a *T. thermophila* cell at the end of conjugation begins with copies of two alleles (e.g., *Ser*H1/*Ser*H4), its descendants (via binary fission) will assort into lines in which the macronucleus is wholly one allele or the other at that locus. This

discovery was first made in the mating-type system (Allen and Nanney 1958) and then found to apply to all macronuclear genes (Doerder et al. 1991). Molecular analysis shows that assortment is due to actual loss of the allele. This means that the macronucleus is functionally homozygous and haploid, even though it has >10X the gene copy number of the micronucleus. This also means that PCR typically amplifies only one allele. Because the MAC is made up of around 225 chromosomes (Eisen et al. 2006) it can be assumed that, barring recombination, the loci on each chromosome assort together. These loci form a co-assortment group (Wickert et al. 2000). It is also of note that genetic markers destined for a particular MAC co-assortment group are found together in the MIC, uninterrupted by any markers destined for a different co-assortment group (Wickert et al. 2000).

Fig. 5 Phenotypic Assortment



Wickert et al 2000

A: Phenotypic (macronuclear) assortment: The circles represent the MAC, the oblong shapes an amitotically dividing MAC. The lines represent a single MAC chromosome (3 of 45 copies shown).
B: Crossing over of the MAC chromosomes can create recombinants, but are rare (Wickert et al. 2000).

Recombination during macronuclear division in *T. thermophila* is a fairly rare event, estimated at less than 1% among two sets of co-assorting markers (Longcor et al. 1996). The distance between the markers in one instance of recombination that was found was estimated to be greater than 100kb (Longcor et al. 1996). There was, however, significant recombination between two independently obtained nonsense mutations in the *SerH1* allele (Deak and Doerder 1998). The homozygotes for mutations *SerH1-1* and *SerH1-2* did not express the H1 i-antigen, but the heterozygote (*SerH1-1*/ *SerH1-2*) always expressed the H1 allele under the appropriate conditions. Molecular study of these mutations, separated by 726 nucleotides, showed a 30% recombination rate generating a stable, functional *Ser*H1 allele in heterozygotes (Deak and Doerder 1998). Because this recombinant allele is found in the macronucleus, it was not transmitted to progeny at the next conjugation.

1.4 The Evolution of *SerH*

The *Ser*H genes of *T. thermophila* and their presumed homologs in other ciliates are little studied with respect to their mode of evolution. There are two principle studies, Gerber et al. (2002), which was discussed above, and Katz et al. (2006). The Katz study of *Ser*H alleles, looking only at the imperfect cysteine repeats, suggested that these alleles evolve by a mix of concerted evolution as well as vertical transmission. This particular study was undertaken after Lynch (Lynch and Connery 2003) noted that the *Ser*H polymorphism observed by Gerber et al. (2002) gave an unusually high estimate of effective population size (N_e). In fact, it placed the effective population size of *T.*

thermophila on the same order as some prokaryotes. This was much higher than any previous estimate and because of this Katz attempted to determine if there was anything unusual at the *Ser*H locus as compared to other *T. thermophila* genes, both nuclear and mitochondrial. Katz found that the *Ser*H locus of *T. thermophila* has a significantly higher amount of variation than do five other nuclear loci including actin, *Ef1*α, SSU – rDNA, α*tub* and one mitochondrial mt-SSU (Katz et al. 2006).

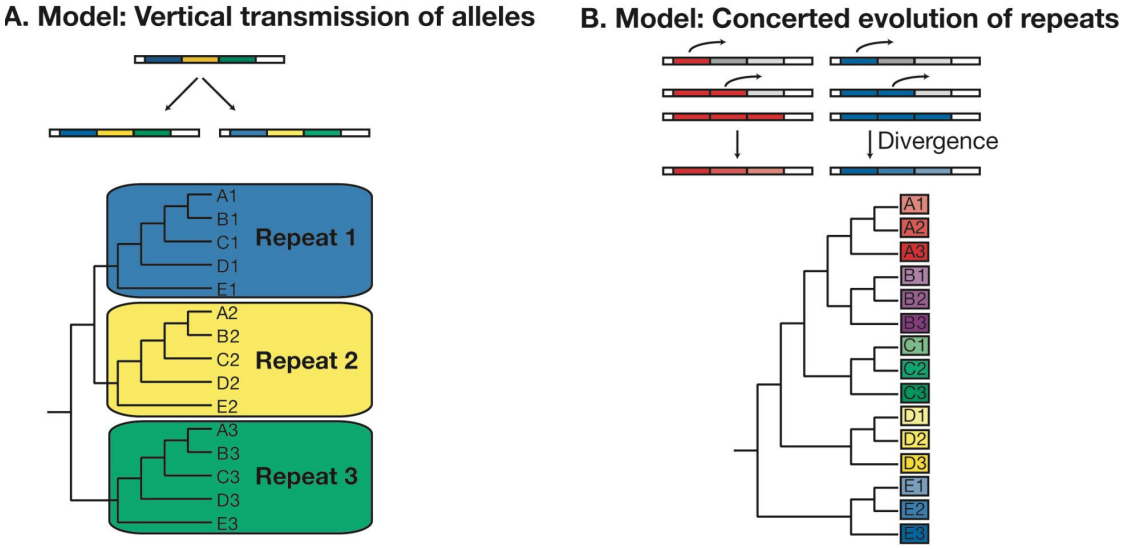
In an attempt to explain the high variation at the *SerH* locus Katz et al. (2006) investigated the evolution of *Ser*H by using sequences of the first three full individual imperfect repeats of 11 *Ser*H alleles to construct phylogenetic trees. The phylogenetic relationship among the repeats showed a mode of evolution that was a mix between concerted and vertical transmission.

Vertical transmission (Fig. 6) refers to the passing of genetic information from one generation to the next via standard hereditary mechanisms. If a gene composed of repeats, such as the *SerH* gene, is undergoing vertical transmission it would be expected that each repeat is more similar to that same repeat in another allele than to the other repeats in the same allele. That is, repeat 1 from allele A would be more similar to repeat 1 from alleles C, D, and E, than it would be to the repeats 2 and 3 from allele A.

Concerted evolution describes the situation where members of a repetitive gene family evolve in concert as a unit. A mutation in one repeat spreads through the family. There are several mechanisms that can lead to this homogenization of repeats including whole gene duplication, transposition and gene conversion (reviewed Nei and Rooney 2006). If concerted evolution is at work, repeats within one allele will show more with one another than with the repeats of other alleles. Examples of concerted evolution are

well documented in the operons encoding rRNAs 16S, 23S, and 5S in *Escherichia coli* (Blattner et al. 1997), For example, the sequence divergence of all seven 16S rRNA genes is only 0.195% in *E. coli* strain K-12, whereas, in a closely related bacterial species, *Haemophilus influenzae,* the coding regions of six ribosomal operons are entirely identical (Fleischmann et al. 1995). In most eukaryotic organisms, the genes encoding ribosomal RNA are tandemly arrayed and undergo concerted evolution (Arnheim et al. 1980, Coen et al. 1982, Schlotterer and Tautz 1994). The *RNU2* locus of primates which encodes U2 snRNA is another well documented case of concerted evolution. The *RNU2* locus has been evolving concertedly >35 Myr since the divergence of baboons and humans. Thus the repeat units of the tandem array are essentially identical within each species, but differ between species (Pavelitz et al. 1995).

Fig. 6 Vertical Transmission and Concerted Evolution



A: The repeats from each allele show more similarity to the same repeat of other alleles than to the repeats from within that same allele. **B**. The repeats from within each allele are more similar to one another than to the repeats from other alleles (From Katz et al. 2006).

In this study *SerH* nucleotide sequences were examined from 20 wild lines of *Tetrahymena thermophila* collected in the Allegheny National Forest. These sequences along with the published sequences of the *Ser*H1, *Ser*H3, *Ser*H4, *Ser*H5, and *Ser*H6 alleles, obtained from Genbank, and their putative amino acid sequences were analyzed *in toto* to determine the phylogenetic relationships among samples. The sequences of the repeats making up the *SerH* genes were analyzed to determine whether the mixed evolutionary mechanism proposed in Katz holds true. In addition, the ER-transmembrane signal containing leader sequences (amino end) as well as the carboxyl terminus containing the GPI-linkage site were analyzed to provide as much information as possible about the *Ser*H alleles.

CHAPTER II

MATERIALS AND METHODS

Twenty-one wild *Ser*H sequences were provided by the Doerder lab, one was discarded as it was not unique. Also, the five sequenced *Ser*H alleles were retrieved from Genbank. This study focuses on sequence variation and not population parameters thus justifying elimination of the duplicate sequences. Alleles and their properties are shown in Table II.

2.1 Collection, Culture, DNA extraction and Sequencing

Tetrahymena thermophila were collected in water samples from two ponds in the Allegheny National Forest in PA. The samples were isolated and grown at 28° in PPY medium consisting of 1% w/v proteose peptone, 0.15% w/v yeast extract and 0.005M $FeCl_3$. Whole cell DNA was isolated using a modified microwave extraction method (Goodwin, 1993).Typically, cells of 8-14 ml of a 2-3 day old culture (stationary phase) were pelleted by low speed centrifugation, washed once with 0.01M pH 7.4 Tris-HCl buffer (PPY only), and resuspended in ~0.25 ml of residual medium or buffer to which was added 100 µl of DNA extraction buffer (50 mM Tris-HCl pH 7.4, 50 mM EDTA, 3% w/v SDS). Following three 10 second pulses of a 750 watt microwave on high setting,

another 200 µl extraction buffer and 10 µl of RNaseA (10 mg/ml) were added, and the mixture was incubated at room temperature for 10 min. The DNA mixture was then transferred to a 1.5 ml microcentrifuge tube, extracted with 400 µl phenol/chloroform, and after centrifugation for 5 min at 12g, the aqueous layer (~400 µl) was transferred to a fresh tube. After addition of 15 µl 3M sodium acetate, DNA was precipitated with 300 µl isopropanol (20 min. on ice or overnight at -20°) and the pellet was washed with 70% ethanol. The DNA was dried under a vacuum, resuspended in 20-30 µl dH_2O and used directly or diluted in standard PCR amplification using Z-Taq (TaKaRa) and AT/CT primers (AT: GTAAAACAAAACTATAATAATTTG; CT: TCAAAAAGTGCAATTTTAAATTC). These products were cloned into SURE cells using the pGEM Teasy kit. The plasmids were purified with a Qaigen kit and were sequenced at the CSU sequencing facility.

2.2 Sequence Analysis

Sequence files were first assembled into complete alleles (~1.3 kb) with the Geneious software package (Drummond et al. 2009). These files were then translated by Geneious using the ciliate nuclear genetic code to obtain the putative amino acid sequences.

Both the nucleotide and putative amino acid sequences were exported as FASTA files. Seqverter (http://www.genestudio.com/seqverter.htm) was used to combine individual FASTA files into multiple FASTA files which were imported into the MEGA 4 software package (Tamura et al. 2007). Alignments of both the nucleotide and putative amino acid sequences were performed by ClustalW in MEGA 4 with a gap opening

penalty of 15 and a gap extension penalty of 6.66 for nucleotide alignments and a gap opening penalty of 10 and extension penalty of 0.1 for amino acid alignments. Alignments were then adjusted by eye if necessary. Sequence analysis for G + C content, codon usage and relative synonymous codon usage were performed with MEGA 4. Comparisons of codon usage were made to the *P. tetraurelia* 51A gene, obtained from Genbank (Nielson et al. 2000) and from a collection of 283 *T. thermophila* protein coding genes at the Codon Usage Database < http://www.kazusa.or.jp/codon/> (Nakamura et al. 2000). The graph of G + C content was made using the G/C content/skew program at the Nano+Bio-Center website < http://nbc11.biologie.uni-kl.de/framed/left/menu/auto/right/GC/>. Each sample was run with a window of 100 and stepping of 10. The average G + C content was then graphed against nucleotide position from 0 – 1241. The DNAsp software package (Rozas et al. 2003) was used to calculate codon usage bias (NC). DNAsp was also used to calculate and graph nucleotide diversity along the sequence.

The deduced amino acid sequences were then all run through the big-PI predictor (Eisenhaber et al. 1999) at the ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server from the Swiss institute of bioinformatics (Gasteiger et al. 2003) to determine the GPI anchor site. SOSUI (Hirokawa et al. 1998) and the NetSurfP server (Peterson et al. 2009) were used to predict any alpha-helices in the *Ser*H proteins.

2.3 Analysis for Selection

To determine the ratio nonsynonymous to synonymous nucleotide substitutions (dN/dS or ω) the full *Ser*H nucleotide sequences were aligned by translation in Geneious

and exported as a fasta file. Then dN/dS analysis was done with the HyPhy software package (Pond et al. 2005) on the Datamonkey.org web server (Pond and Frost 2005) using the SLAC algorithm with the REV nucleotide substitution model (Pond and Frost 2005, 2).

Two other tests for neutrality were then performed on the aligned sequences. Tajima's D test (Tajima 1989) and Fu and Li's D test (Fu and Li 1993) were performed with the DNAsp software package (Rozas et al. 2003).

2.4 Phylogenetic Analysis

From the alignments, MEGA 4 was used to calculate pairwise distance between the sequences and to build phylogenetic trees via neighbor joining with bootstrap tests of 500 replicates.

The full sequences, both nucleotide and amino acid, were then broken into the leader (amino terminus), each full (8-cysteine containing) imperfect repeat, the 4-cysteine half imperfect repeat and the carboxyl terminus (Fig. 2). The putative amino acids sequences were separated and converted into an individual FASTA file. The nucleotide sequences were then separated in the same manner using the putative amino acid sequences as guides.

The 77 sequences (77 nucleotide and 77 putative amino acid) of complete repeats were then aligned with ClustalW MEGA 4 as above, pairwise distance was computed and phylogenetic trees were built from the alignments by neighbor-joining as they were from the complete sequences. The nucleotide repeat sequences were then split into 4 cysteine containing half repeats (each half labeled "a" or "b") using the putative amino acid

sequences as guides. These 179 half-repeat sequences were then aligned in MEGA 4 and from them, neighbor joining trees were built as in Katz (2006).

The neighbor joining trees were all analyzed to determine whether the *Ser*H showed a pattern of evolution consistent with vertical transmission, concerted evolution, or a mixed evolution.

Table II: Information about the *Ser*H Sequences Used in This Study

Sequence	Obtained	Genbank Accession #	Number of Repeats	Length (BP)	Length (AA)	Water Name	Date	State
18217-2	Cloned		3.5	1265	422	CRWP	8/16/2001	PA
18217-4	Cloned		3.5	1262	422	CRWP	8/16/2001	PA
18218-4	Cloned		3.5	1274	424	CRWP	8/16/2001	PA
18219-4	Cloned		3.5	1251	416	CRWP	8/16/2001	PA
18221-4	Cloned		3.5	1275	424	CRWP	8/16/2001	PA
18222-4	Cloned		3.5	1250	416	CRWP	8/16/2001	PA
18226-2	Cloned		3.5	1249	416	CRWP	8/16/2001	PA
18230-1	Cloned		3.5	1241	415	CRWP	8/16/2001	PA
18234-1-1	Cloned		3.5	1265	421	CRWP	8/16/2001	PA
18235-2	Cloned		3.5	1274	423	CRWP	8/16/2001	PA
18247-2	Cloned		3.5	1274	424	SG29	8/17/2001	PA
18249-1	Cloned		3.5	1250	416	SG29	8/17/2001	PA
18252-4	Cloned		3.5	1275	424	SG29	8/17/2001	PA
18253-3	Cloned		3.5	1273	424	SG29	8/17/2001	PA
18253-4	Cloned		4.5	1429	473	SG29	8/17/2001	PA
18262-1	Cloned		3.5	1275	424	SG29	8/17/2001	PA
18266-2	Cloned		4.5	1429	475	SG29	8/17/2001	PA
18267-1	Cloned		3.5	1249	415	SG29	8/17/2001	PA
18278-3-1	Cloned		3.5	1267	421	CRWP	8/21/2002	PA
18280-1	Cloned		2.5	959	319	CRWP	8/21/2002	PA
Genbank H1	Genbank	U15793.2	3.5	1311	436			MA
Genbank H3	Genbank	AF190647.1	3.5	1272	423			MA
Genbank H4	Genbank	AF425241.1	3.5	1191	397			PA
Genbank H5	Genbank	AF425242.1	3.5	1185	429			PA
Genbank H6	Genbank	AF425243.1	3.5	1263	421			PA

CHAPTER III

RESULTS

This project analyzed the sequences of 20 unique *Ser*H sequences from cells collected in two ponds (12 and eight sequences respectively) of the Allegheny National Forest in western Pennsylvania. These 20 were supplemented with 5 of the alleles used by Katz et al. 2006, the same alleles labeled H1-H6 in Gerber et al., (2002) obtained from Genbank (Table II). For the 20 unpublished alleles, chromatogram files were edited and assembled as described in Materials and Methods.

3.1 Characterization of *Ser*H Genes

The nucleotide sequences of the 20 non-identical wild lines plus the 5 *Ser*H allele exemplars from Genbank had an average G + C content of 40.6%. This is consistent with a study of 75 protein coding nuclear gene sequences in which the average G + C content was 38% with a range of 25-49% (Wuitschick and Karrer 1999). It is, however, considerably higher than the 27.6% G + C content of predicted exons from the complete macronuclear sequence (Eisen et al. 2006). The deduced amino acid sequences showed *Ser*H to be rich in alanine (18.7%), threonine (15.8%), serine (12.6%) and cysteine (9.37%) (Fig. 7). These sequences did not contain any histidine and only three sequences

had glutamic acid (1 residue each out of ~ 400). Aside from the lack of glutamic acid and histidine the deduced amino acid composition of *Ser*H was very similar to that of the *Paramecium tetraurelia* 51A surface antigen (Preer et al. 1985) (Fig. 8). Notably, these proteins are rich in small amino acids and relatively deficient in large amino acids (e.g., phenylalanine and tryptophan).

A codon usage analysis of these 25 nucleotide sequences showed that 11 codons were not used and five more were used five or fewer times out of 10,520 codons (Fig. 9). The codon usage pattern indicates a slight bias for AT rich codons as compared to their GC rich counterparts. Four codons had a relative synonymous codon usage (RSCU) index of above 2.5: AGA (5.89), CCU (3.03), GCU (2.97) and GGU (2.78) (Fig 10). RSCU values are the number of times a particular codon is observed, relative to the number of times that the codon would be observed for uniform synonymous codon usage (i.e. all the codons for a given amino-acid have the same probability). In the absence of any codon usage bias, the RSCU values would be 1.00 (Sharp et al. 1986). A RCSU value above 1.00 indicates a bias towards the use of a codon. Each of these codons has been found to be the most highly used for their respective amino acids (arginine (for which only one of six codons is used), proline, alanine and glycine respectively) in *T. thermophila* (Wuitschick and Karrer 1999). A comparison of the codon usage of these 25 *Ser*H sequences with a database of 283 *T. thermophila* protein coding genes showed a similar pattern of codon usage with the exception of the codons for the amino acids not used in *Ser*H (glutamate and histidine) and one codon encoding the five most abundant amino acids in *Ser*H (GCU-Ala, ACU-Thr, UCU-Ser, UGU-Cys and GGU-Gly) which were used at a much higher rate in *Ser*H than over the 283 *T. thermophila* genes (Fig. 11).

Fig. 7 Average Amino Acid Use of 25 *Ser*H Deduced Amino Acid Sequences

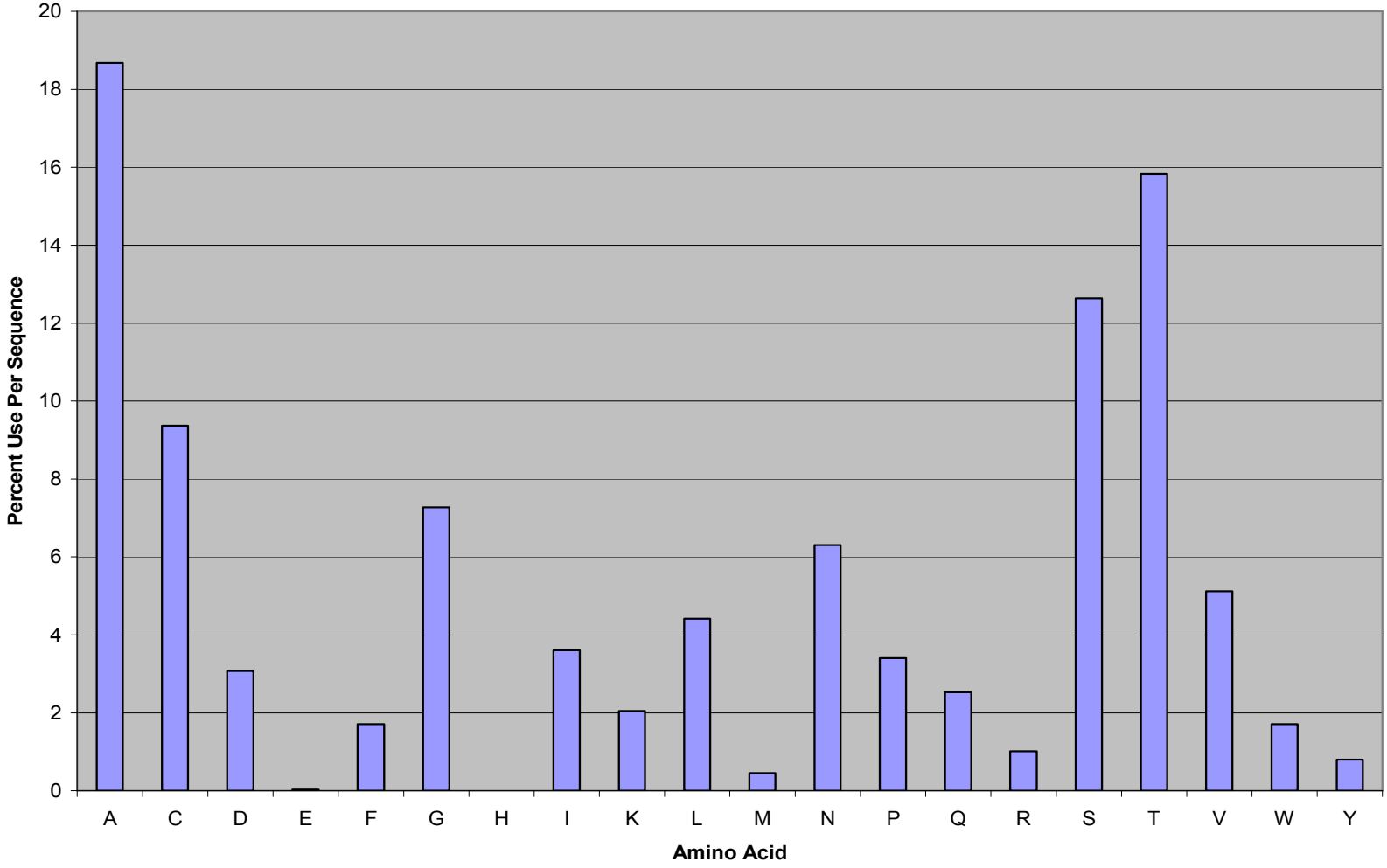


Fig. 8 Comparison of the Amino Acid Composition of the 25 *T. thermophila Ser*H Sequences to the *P. tetraurelia* 51A Surface Antigen

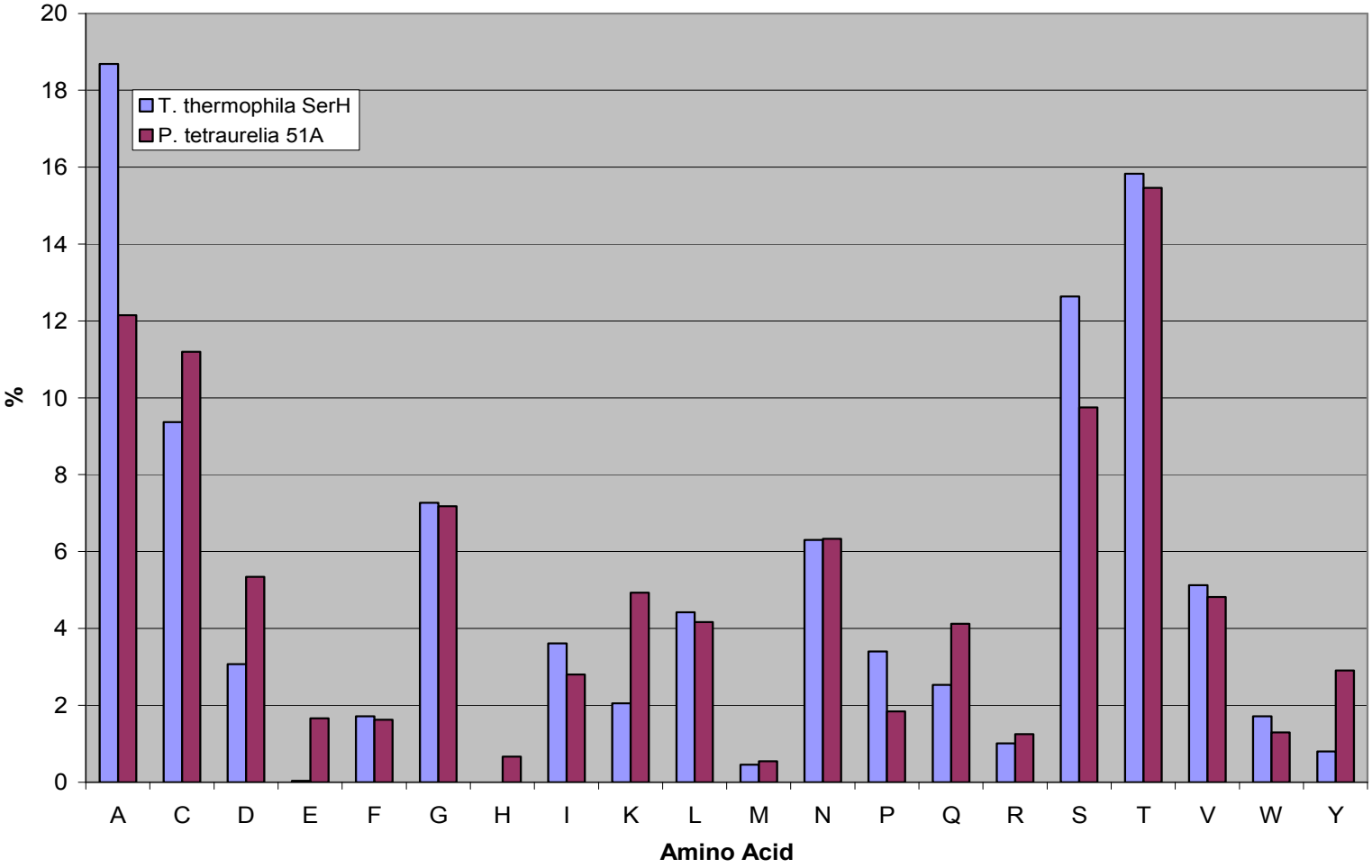


Fig. 9 Codon Usage of the 25 *Ser*H Sequences

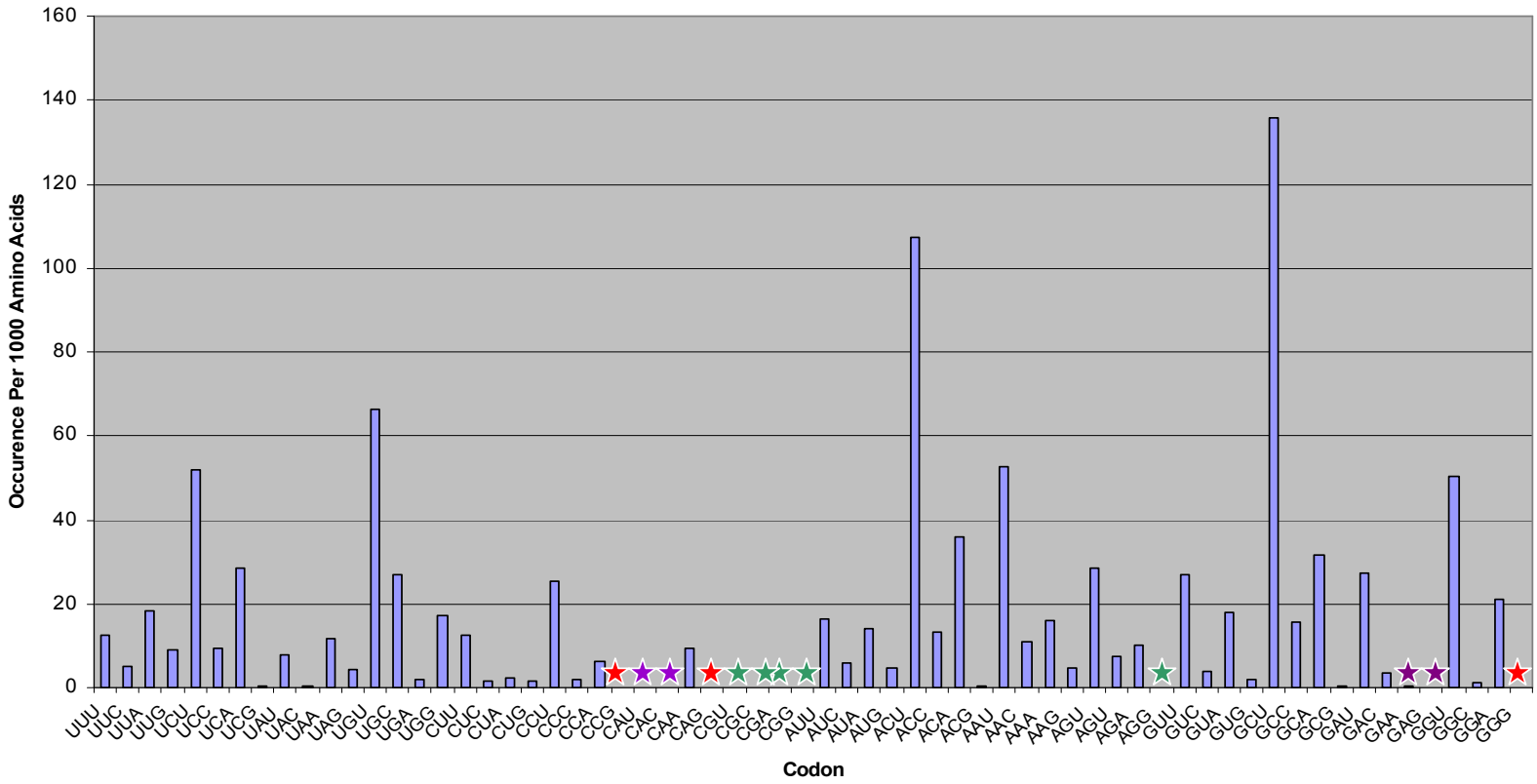


Fig. 10 Relative Synonymous Codon Usage of 25 *Ser*H Sequences

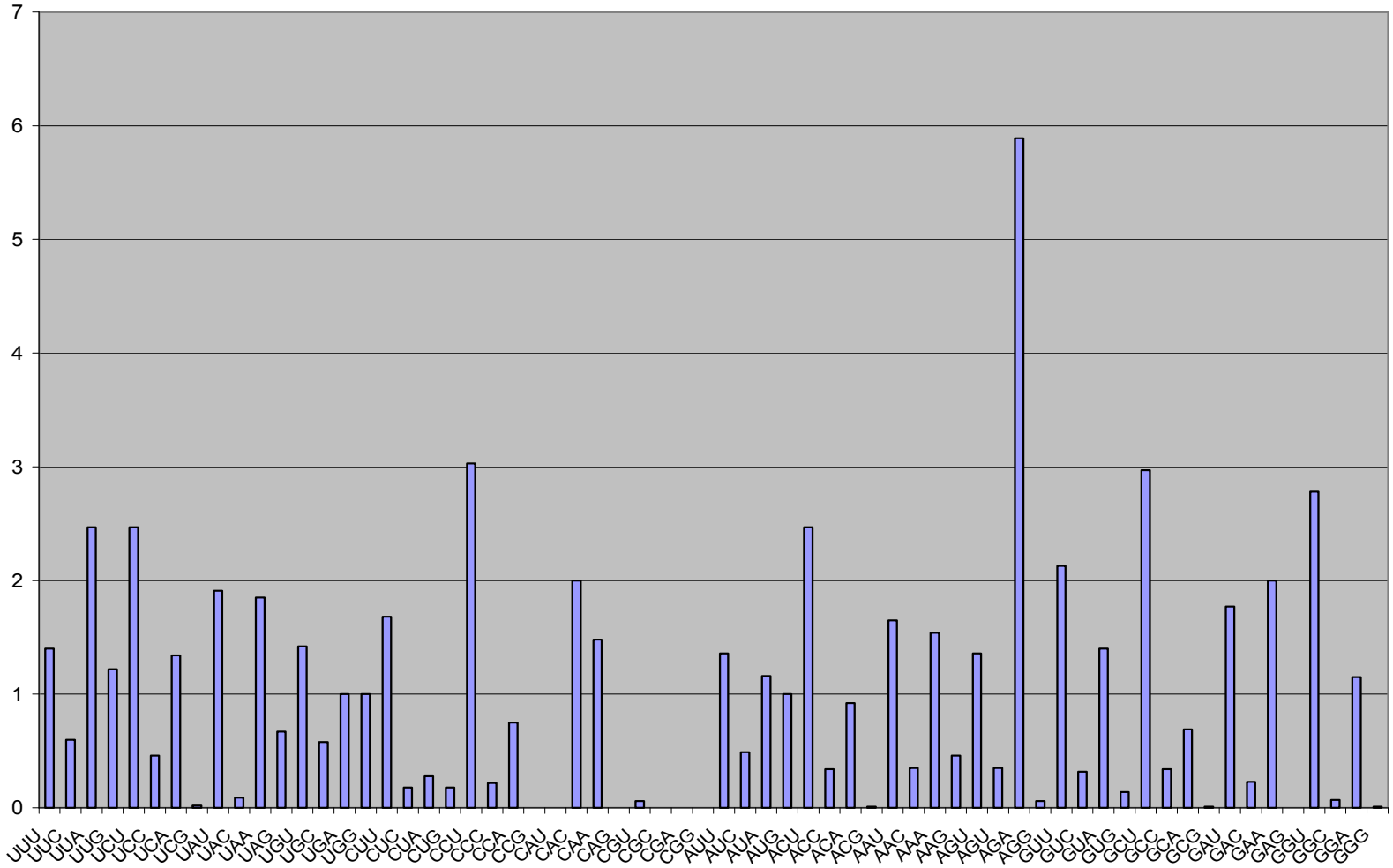
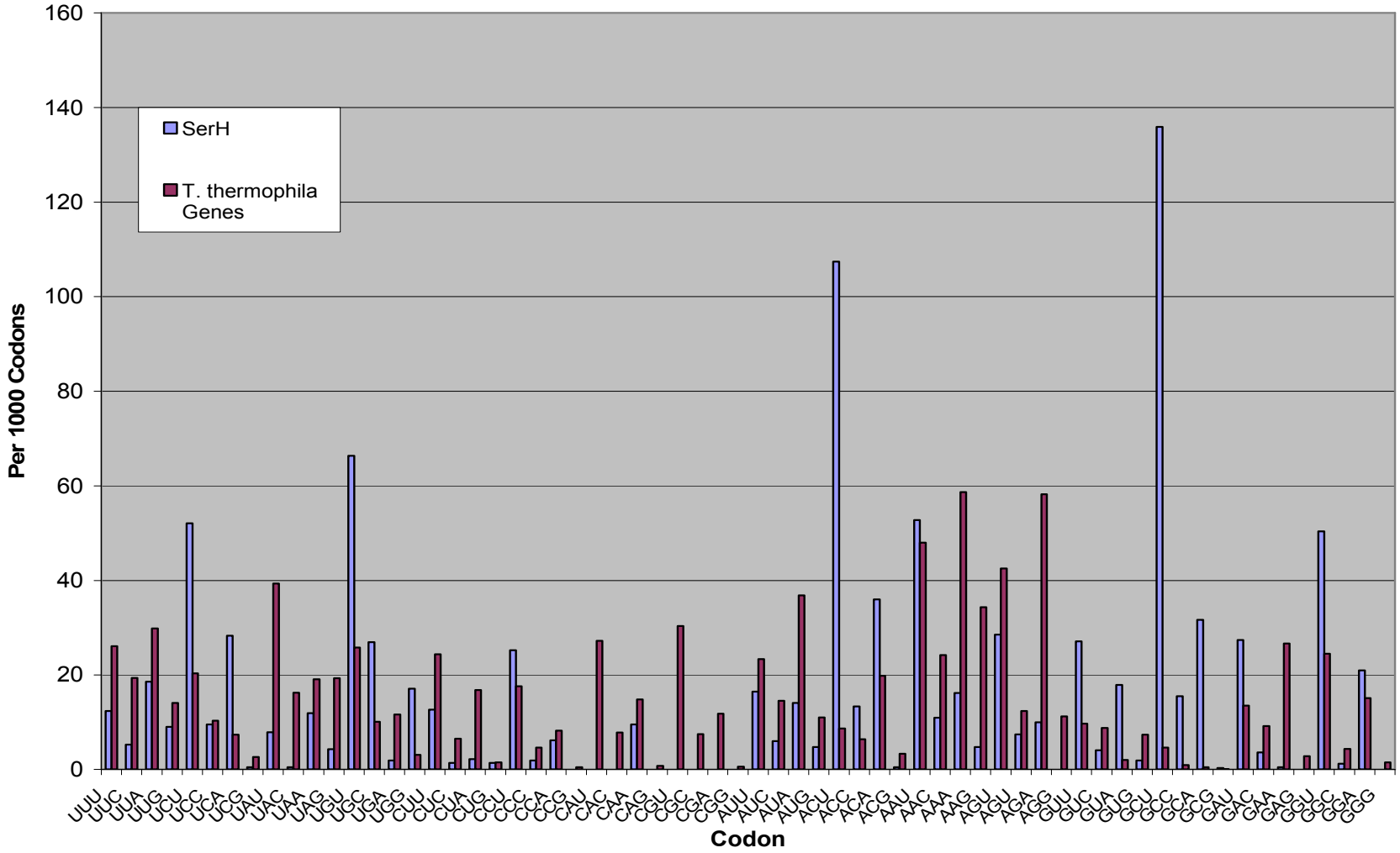


Fig. 11 Comparison of the Average Codon Usage of 25 *Ser*H Sequences to the Average of 283 *T. thermophila* Genes



A codon bias analysis found the average codon bias (Nc) of 25 *Ser*H sequences to be 34.37. Nc, or number of effective codons is a codon bias analysis in which the homozygosity for each amino acid is estimated from the squared codon frequencies. This returns a value between 20 (when only one codon is effectively used for each amino acid) and 61 (when codons are used randomly). If the calculated N_c is greater than 61 (because codon usage is more evenly distributed than expected), it is adjusted to 61 (Wright 1990). This Nc value is indicates that *Ser*H is a highly expressed gene in *T. thermophila* (Salim et al. 2007), which is consistent with other studied that show i-antigens are highly expressed (Leak and Forney, 1996; Gerber et al. 2002).

Using the big-PI predictor (Eisenhaber et al. 1999) on each of the 25 deduced amino acid sequences, a consensus GPI anchor area was found in the carboxyl terminus. Each *Ser*H sequences has a consensus CQAT at the beginning of the carboxyl terminus and in nearly every case the predicted GPI anchor site was in the 6 amino acid residues following the threonine, most commonly immediately after the threonine. However, the sequences with 4.5 repeats were predicted to have a weak GPI linkage site on a threonine nine amino acids after CQAT and the sequence with 2.5 repeats was also predicted to have a weak GPI linkage site at a serine three after CQAT. This is consistent with the idea that *Ser*H alleles with 4.5 or 2.5 repeats are not expressed (see below).

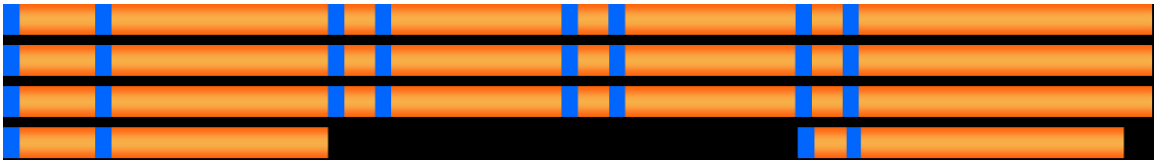
Secondary structure predictions using SOSUI and NetSurfP (see Materials and Methods) suggested no alpha helical areas in the cysteine repeat section of the *Ser*H deduced amino acid sequences except within the first few residues of the first repeat.

3.2 Cysteine Periodicity

As described in Materials and Methods, nucleotide sequences of repeats were aligned using MEGA 4 as based on alignment of translated amino acid sequences. As shown in Fig. 2, repeats contain indels (dashes) when compared to each other. Nucleotide alignment based on amino acid alignment accounts for indels and also provides for detection of synonymous vs. non-synonymous substitutions.

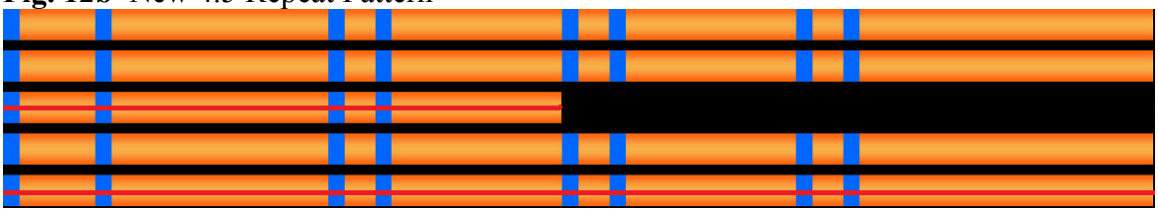
All translated *SerH* sequences showed the expected imperfect repeats each containing 8 cysteines. They also showed the expected half repeat. Though most (22) had the typical 3.5 repeats, one had 2.5 and two had 4.5. The half repeat, containing four cysteines, was usually last. However, two sequences (18253-4 and 18266-2) had two full repeats followed by a half repeat followed by two more full repeats (Fig. 12b). The final repeat of these two sequences was equal to the third (half) repeat repeated twice.

Fig. 12a. Different Cysteine Repeat Patterns: Standard 3.5 Repeats



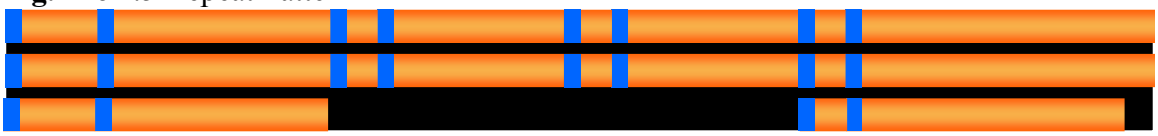
The normal pattern of *Ser*H cysteine repeats three full repeats followed by a half repeat that resembles the front quarter and back quarter of the full repeats. Twenty-two of the sequences showed this pattern.

Fig. 12b New 4.5 Repeat Pattern



Novel pattern of repeats seen in two sequences. 4.5 consisting of two full repeats followed by the half repeat followed by two more full repeats. Also, as indicated by the red line, the final repeat is the same sequence as the half repeat 2x.

Fig. 12c 2.5 Repeat Pattern



This pattern of 2.5 repeats was seen in one sequence. It consists of two full repeats followed by a half repeat.

The complete sequences themselves were highly variable. The average pairwise percent divergence between the nucleotide sequences was 21.2% (Table III) and 26.9% for the deduced amino acid sequences These values were very similar to those previously reported by Katz, 21.4% nucleotide divergence and 31.7% amino acid divergence (Katz et al. 2006). The percent sequence divergence varied across the molecule (Table III).

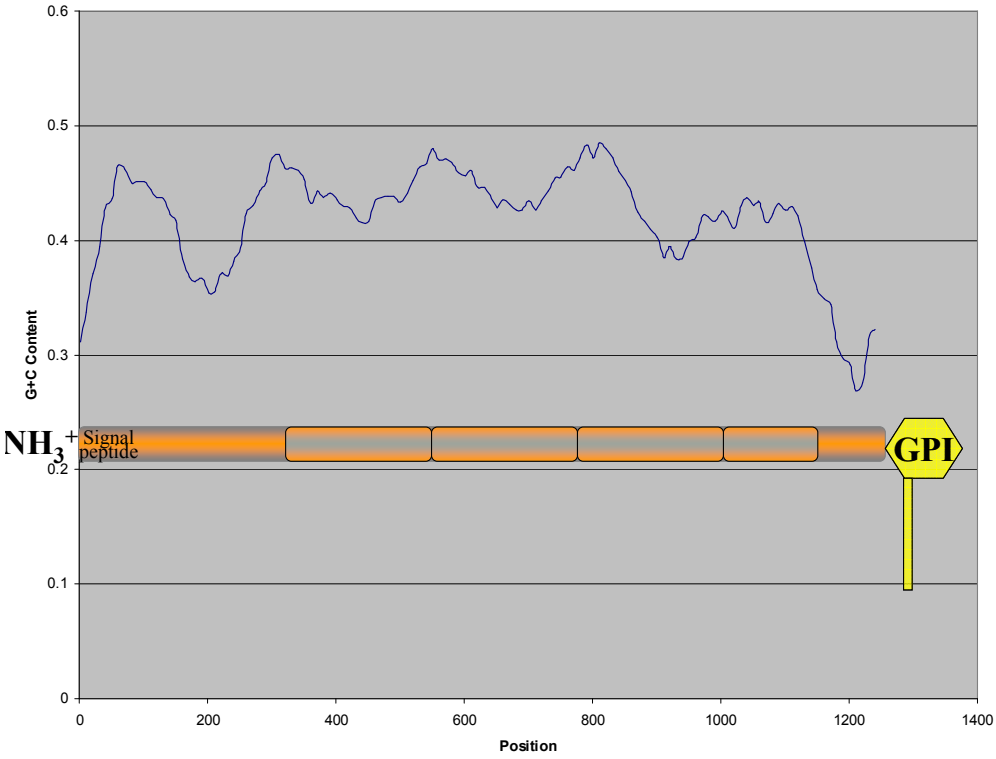
When the sequences of all the full repeats (i.e. each imperfect 8-cysteine repeat taken as an individual sequence, excluding 4-cysteine half repeats) were analyzed, the nucleotide divergence was found to be 39.9% and that of the putative amino acid sequences was 46% (Table III).

The G + C content was higher within the repeats (44.1%) than it was in either the leader sequence (37.8%) or the carboxyl terminus (31.2%) (Fig.13).

Table III: Comparative Sequence Analysis of *Ser*H Sections.

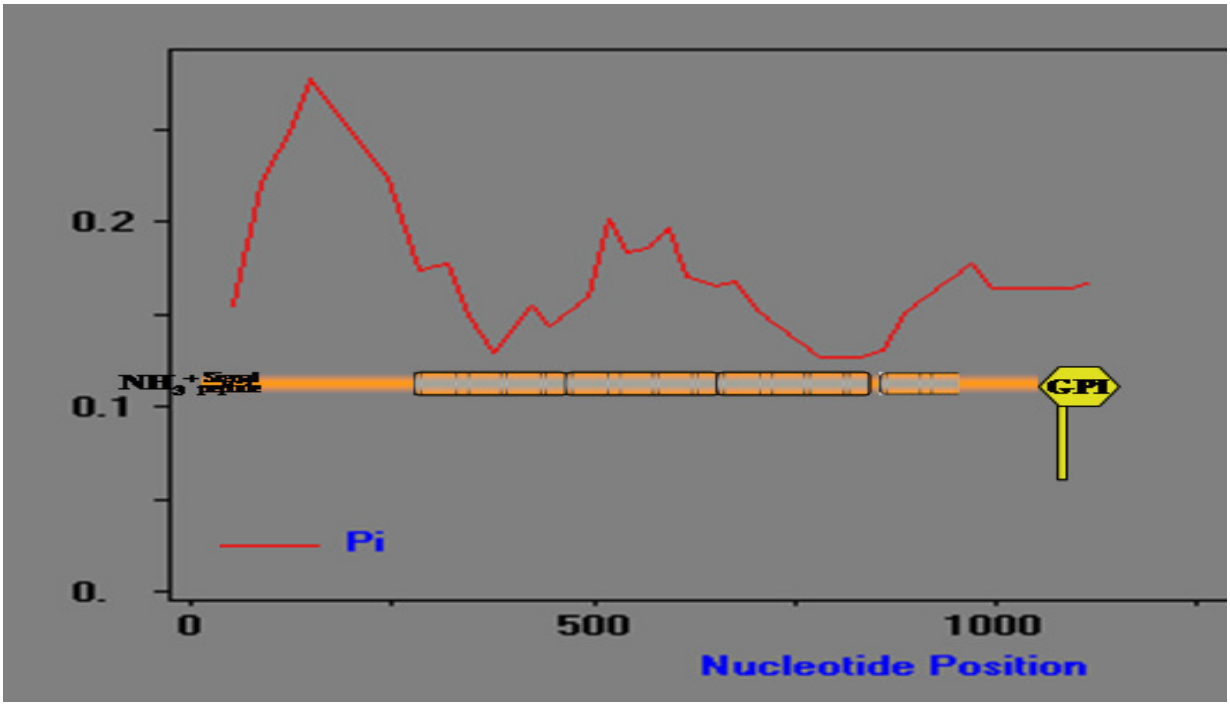
	G + C Content	Mean Pairwise Distance Nucleotide	Mean Pairwise Distance Amino Acid
Leader Sequence	37.8%	0.216	0.294
First Full Repeat	44%	0.212	0.349
Second Full Repeat	45%	0.185	0.231
Third Full Repeat	43.8%	0.170	0.219
Carboxyl Terminus	31.2%	0.018	0.047
All Full Repeats	44.1%	0.399	0.460
Whole Sequence	41.6%	0.212	0.269

Fig. 13 Average G + C Content along 25 *Ser*H Nucleotide Sequences



From the alignment of the 25 nucleotide sequences the nucleotide diversity (π) was calculated using the function in DNAsp. The nucleotide diversity (π) in a sample of *n* haploid individuals can be estimated by averaging the estimated numbers of nucleotide changes (*d*) over all the pairs in the sample. The greatest variability is in the first 200 nucleotides (Fig. 14). This also corresponds to an area of relatively high G + C content along the sequence (Fig. 13). This region has the highest amino acid diversity and is likely the region of the protein most exposed to the environment.

Fig. 14 Mean Nucleotide Diversity Rate along 25 *Ser*H Sequences



3.3 Tests for Selection

To determine whether there was any selection acting upon the *Ser*H locus several analyses were performed. Using sequences aligned to their deduced amino acid sequence alignments the 25 *Ser*H nucleotide sequences were found to have a mean rate of

nonsynonymous (dN) to synonymous (dS) of 0.825. A ratio ω (dN/dS) significantly > 1.0 indicates positive selection, while ω < 1.0 indicates negative, or purifying, selection (Yang and Bielawski 2000). Tajima's D test (Tajima 1989) and Fu and Li's D test (Fu and Li 1993) were also performed and showed no evidence of positive selection at the *Ser*H locus, again, indicating that the *Ser*H alleles are selectively neutral and the locus is under purifying selection.

3.4 Phylogenetic Analysis for Mode of Evolution

As a first step in analyzing evolutionary relationships among the *Ser*H alleles, phylogenetic trees were built using both the complete nucleotide sequences and the complete amino acid sequences. A neighbor joining analysis using the complete nucleotide sequences (Fig. 15) showed that 16 of the 20 wild alleles were most similar to the published *Ser*H6 or *Ser*H3 sequences and four (18253-4, 18266-2, 18267-1 and 18280-1) more closely resembled *Ser*H4 and *Ser*H5. The tree built with nucleotide sequences showed only one difference from the tree built with the deduced amino acid sequences. In the amino acid sequence derived tree the sequence of wild line 18267-1 was more closely related to the *Ser*H3/H6 group then to *Ser*H4/*Ser*H5 (Fig 16).

To assess the rates of vertical transmission and concerted evolution of *Ser*H repeats, as in Katz et al. (2006), neighbor joining trees were constructed using each full repeat of the 25 *SerH* sequences as a separate sequence. That is, each sequence was separated into its imperfect repeats as in Fig. 2 and these imperfect repeats were used as individual sequences. The half repeats were not used. The resulting trees of 77 sequences, one based upon nucleotide sequences and one base on amino acid sequences, showed a

mixed mode of evolution for the *SerH* gene. In both the tree built on nucleotide sequences and the tree built on amino acid sequences the *Ser*H6, *Ser*H3 and the *Ser*H6-like and *Ser*H3-like wild strains showed a standard vertical transmission pattern of evolution (Fig. 17 pink, green and orange sections). The only exception was the wild line 18267-1 which aligned more closely to *Ser*H4 and *Ser*H5 in the whole nucleotide sequence neighbor-joining tree but more closely to *Ser*H3/H6 in the putative amino acid sequence tree. This line showed standard vertical transmission with the *Ser*H3/H6 – like sequences.

The *Ser*H1, *Ser*H4 and *Ser*H5 and the two wild sequences with 4.5 repeats (18253-4 and 18266-2) showed a pattern consistent with a mix of concerted evolution and vertical transmission (Fig. 17, blue section). The *Ser*H1 repeats showed a different mixed transmission pattern than the *Ser*H4, *Ser*H5 and related sequences as did the wild line 18280-1, which only had 2.5 repeats. *Ser*H1's 1st repeat was not closely related to anything else, its 1st and 2nd repeats appear on their own branch, most closely related to each other, and its 3rd repeat is closely related to the 3rd repeats of the H3/H6 group (Fig. 17 yellow section). 18280-1's 1st and 2nd repeats were most closely related to each other and its final (half) repeat was most closely related to the half repeats from the other sequences.

A neighbor joining tree was drawn using each half-repeat unit as an individual nucleotide sequence (Fig. 18). The 8-cysteine repeats were split so that the second half of the repeat began with the 5th cysteine (Fig. 2). This allowed for the half repeats of each *Ser*H gene (e.g. the last half on a 3.5 repeat sequence) to be included. This tree (Fig. 18) again showed a pattern suggesting vertical transmission for H3, H6 and the 17 H3/H6

like wild lines and a mixed pattern of vertical transmission and concerted evolution for H4, H5 and the 3 wild sequences closely related to them. H1, once again, showed a mixed pattern of evolution, but did not fall on the same branches as the H4/H5 group. The half repeats, regardless of the number of full repeats in the sequence all sorted to the same branch of the tree. These half repeats appear to be most similar to the first two cysteines of the first full repeat and the final two cysteines of the final repeat (Fig. 2). That is to say that they look as if they are the remnants of a full repeat whose middle section had been removed.

Fig. 15 Neighbor Joining Tree - Complete *Ser*H Nucleotide Sequences
-Major relationships are color coded.

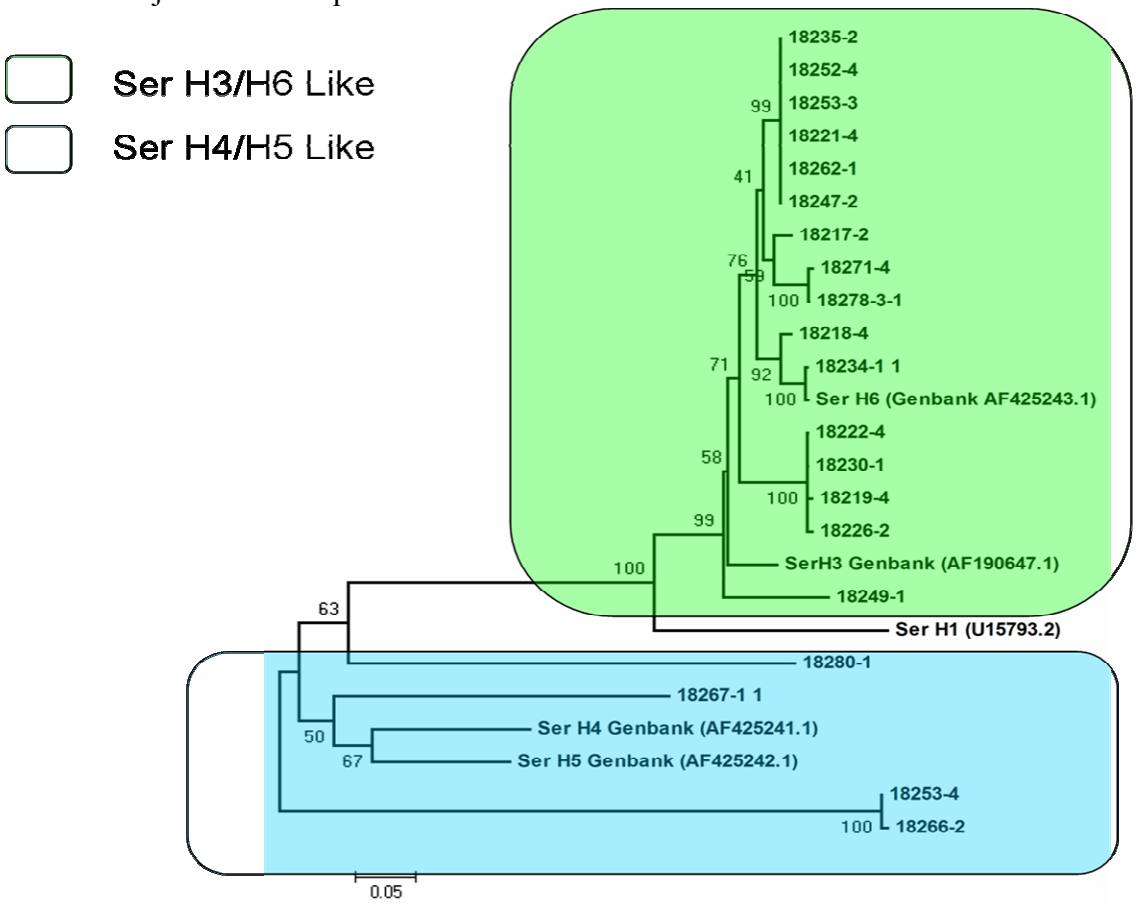


Fig. 16 Neighbor Joining Tree: Complete Deduced Amino Acid Sequences
-Major relationships are color coded.

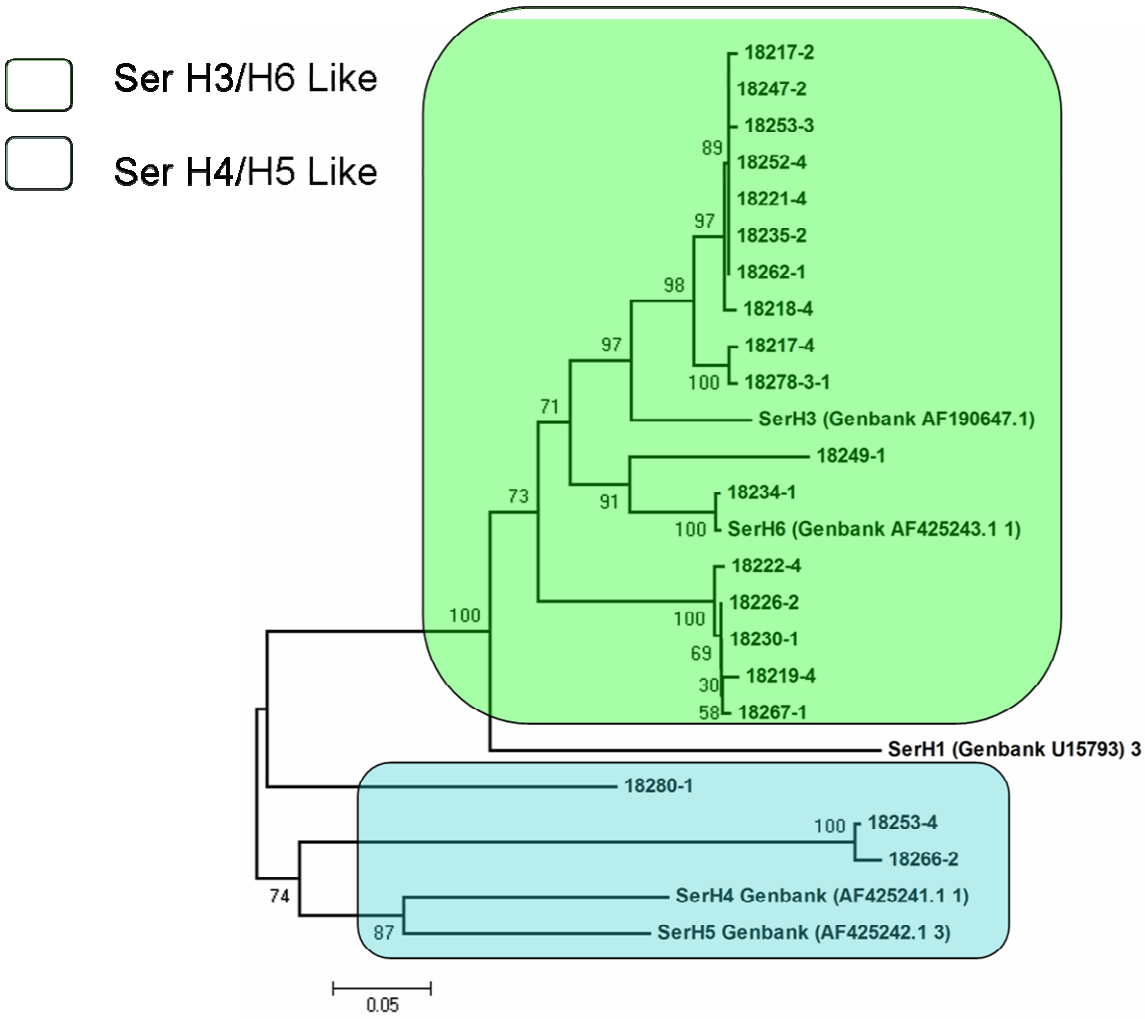
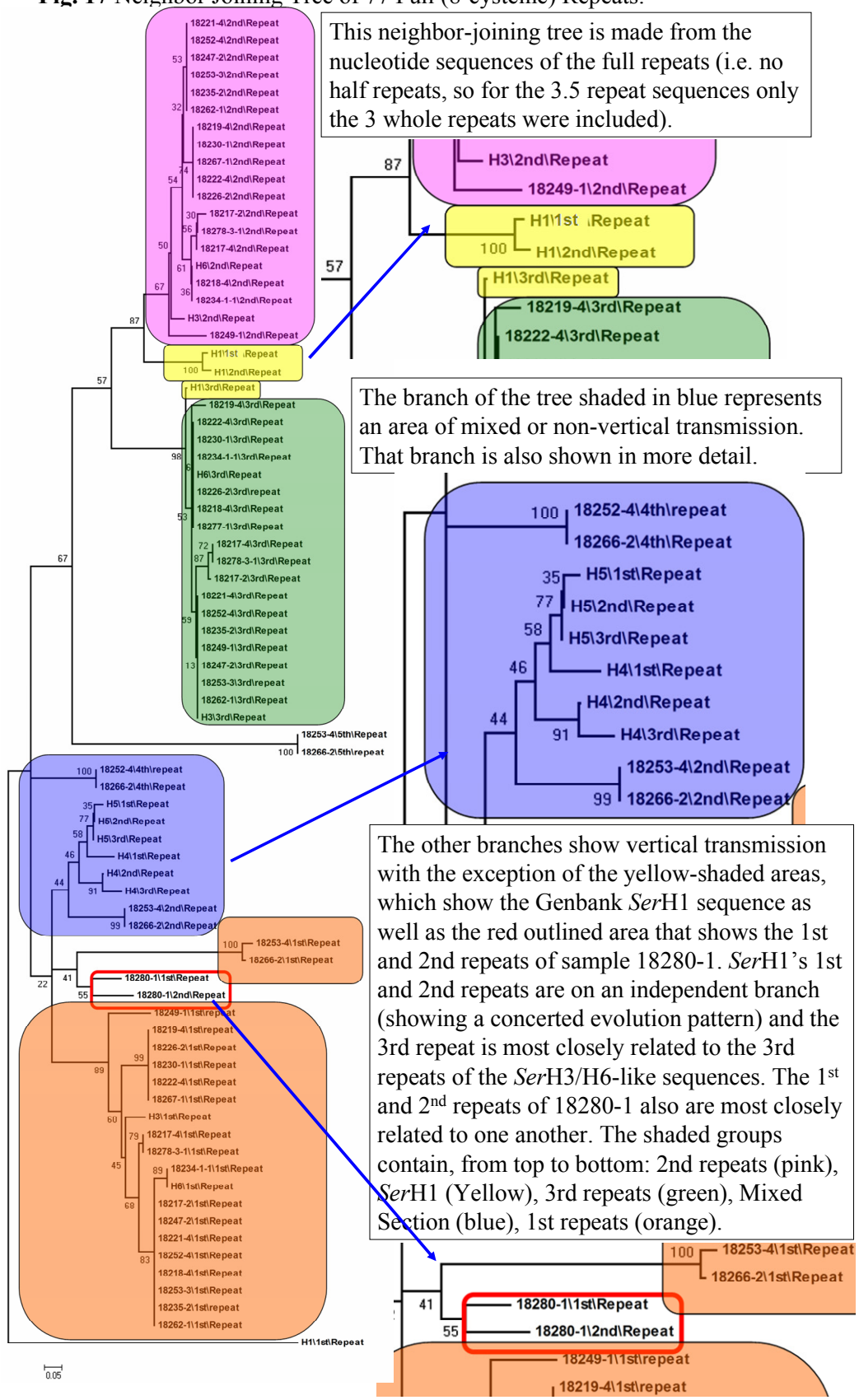


Fig. 17 Neighbor Joining Tree of 77 Full (8-cysteine) Repeats.



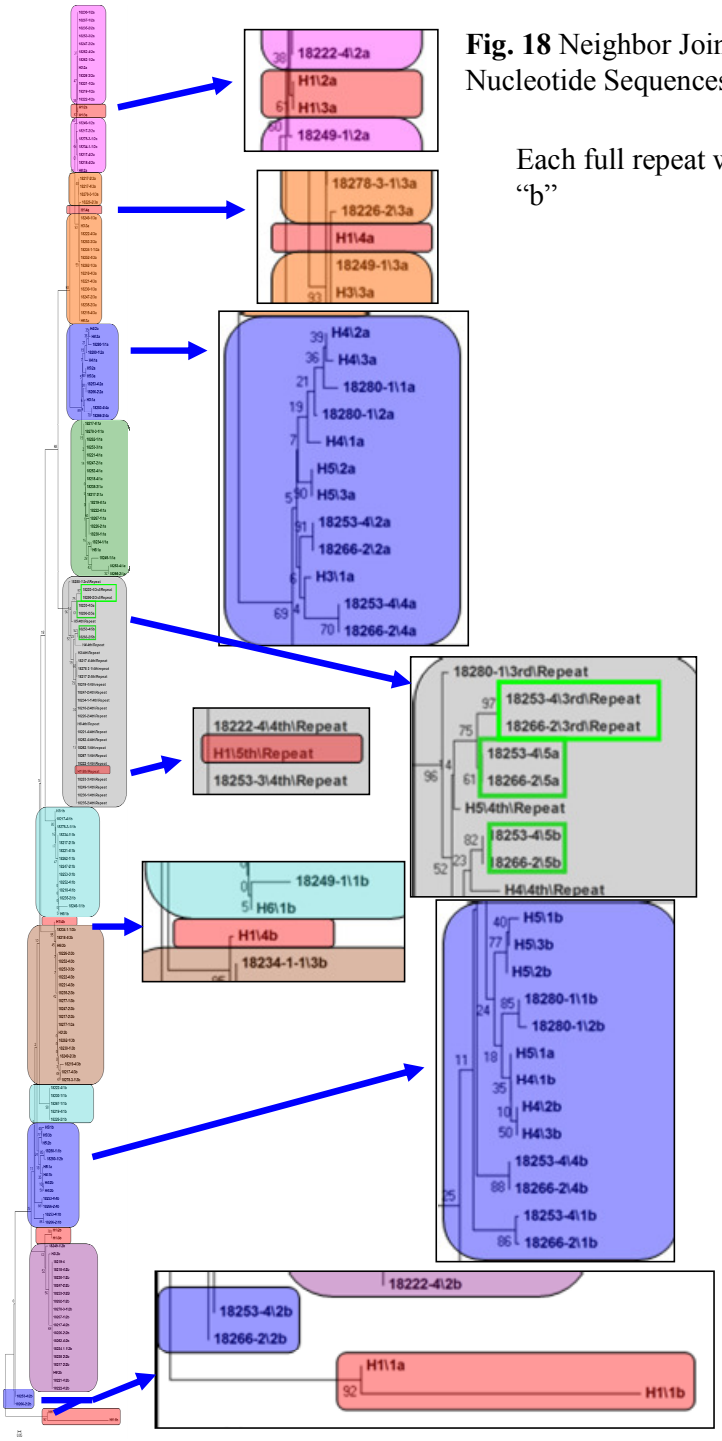


Fig. 18 Neighbor Joining Tree: All Half Repeat Nucleotide Sequences

Each full repeat was divided into halves "a" and "b"

From top to bottom the branches contain:

Repeat 2a (pink),

Repeat 3a (orange),

Mixed Section,

Repeat 1a (green),

Final (half) repeats and repeats 5a and 5b (grey),

1b (light blue),

3b (brown),

1b (blue),

Mixed Section,

2b (purple),

Mixed Section (Blue and Red).

The red shaded areas indicate *Ser*H1 which showed a mixed pattern of evolution but did not follow the same pattern as any other sequence.

The green outlined areas (within the grey) denote the half repeat and the final full of the two 4.5 repeats with the half repeat in the middle. The light green is the middle (half repeat) and the dark green are the two halves of the final (full) repeat.

CHAPTER IV

DISCUSSION

Vertical transmission of genes is the standard method of genetic inheritance. Under this model a mutation in a gene would be transferred from parent to offspring in a linear fashion. The mutation would not be associated with other sections of the gene. For repeated genes or genes that contain repeated segments, such as the *Ser*H i-antigen genes of *T. thermophila,* there is the possibility of concerted evolution. In the case of concerted evolution a mutation in one repeat of a gene would appear in the other repeats of the same gene. Therefore that single mutation would appear in locations across the gene, or in multiple genes.

A previous study by Katz et al. (2006) found evidence of a mixed mode of concerted evolution and standard vertical transmission among the alleles of the *Ser*H gene (Katz et al. 2006). The present study examined the question of concerted evolution at this locus in more detail. Here, the sequences of 20 separate variants collected in the wild were added to the 5 *Ser*H alleles available on Genbank. Most sequences used here had 3.5 repeats while the majority used in the Katz study had 4.5 In addition, these sequences were more fully characterized with respect to the structure of the *Ser*H genes. These sequences showed the overall structure expected of the *Ser*H gene. Specifically,

they had a leader sequence with an ER transmembrane signal, mostly 3.5 (range 2.5 to 4.5) imperfect, 8-cysteine containing repeats, and a carboxyl terminus area containing a GPI anchor site.

Seventeen of the wild variants showed the expected pattern of repeats, specifically, three 8-cysteine containing repeats followed by a half (4-cysteine) repeat. In all cases the half repeat appears to consist of the ends of a full repeat, as if the 4-cysteine central region were removed. The half repeats do not appear to be the building block of full repeats. Two (18253-4 and 18266-2), however, had two full repeats followed by the half repeat followed by two more full repeats while one (18208-1) only had 2.5 repeats total.

4.1 Patterns of Repeat Inheritance

Based on the unique arrangement of the repeats from 18253-4 and 18266-2 as well as the fact that these two sequences were two of the three sequences that contained glutamic acid residue and had weak GPI sites it is possible that these two sequences represent pseudogenes. In these two lines the half repeat is clearly related to the final full repeat (Fig. 12). Wild lines containing *Ser*H alleles with 4.5 repeats need to be more closely studied for expression and possession of *Ser*H paralogs.

In the sequence that had only 2.5 repeats (18280-1) both the first and second repeats, the full repeats, are very similar to one another and are most similar to the first repeats of the other sequences. The half repeat is very similar to the final half repeat of the regular 3.5 repeat containing sequences. If this sequence did lose a repeat it is impossible to tell whether that repeat was more similar to the 1st, 2nd, or 3rd repeats from

the other sequences. The fact that this sequence only has 2.5 repeats could be evidence that it is non-functional. If it is a pseudogene it is possible that *Ser*H homolog is being expressed. Predictions about gene functionality and expression are unconfirmed as the original isolates have not been tested for expression of the surface antigen by RT-PCR or an antibody test.

4.2 Composition and Evolutionary Mode of *Ser*H

The *Ser*H sequences had a G+C content of 42%, which is consistent with protein coding areas across the *T. thermophila* genome. Interestingly, the G + C content within the repeat area of the genome was higher than that in both the leader section and carboxyl terminus (Table III). This could indicate codon usage bias within the repeats, or the necessity of higher G + C to code for necessary amino acids within the repeats. The deduced amino acid composition was rich in alanine, cysteine, serine and threonine and contained no histidine or glutamate (except as mentioned above). This amino acid composition is similar to previously described *Ser*H examples, *Ser*L (Doerder and Gerber 2000) and *Ser*J (Doerder 2000). It is also is similar to that 51A surface antigen of *P. tetraurelia* except that 51A does contain glutamate and histidine. Also, aside from the conserved cysteines within the repeat sections there are conserved sites that have charged amino acids (usually aspartate immediately preceding a cysteine) and conserved sites containing large amino acids (phenylalanine, tyrosine and tryptophan) (See appendix). These conserved sites could be necessary the *Ser*H allele to have a functional structure.

The carboxyl end of the *Ser*H sequences was the most conserved area of the sequence. This was also the area that had the lowest G + C content. It is possible that this

section of the gene may play some role in RNA stability and the regulation of *Ser*H expression (Gerber et al. 2002).

The *Ser*H sequences collected for this study showed no sign of positive selection. Neither the Fu and Li nor Tajima's test found evidence of positive selection and the mean dN/dS ratio of these sequences was under 1.00, again indicating no positive selection. This is in keeping with Gerber et al. (2002) who used a smaller sample size. These results suggest that *Ser*H alleles are effectively neutral and purifying selection is removing non-functional alleles from the *Ser*H locus.

In order to test for concerted evolution, phylogenetic trees of the full repeats were analyzed. Using a smaller sample, Katz et al. (2006) observed primarily vertical transmission, with some instances of concerted evolution. Using the repeats as individual sequences, on neighbor joining trees for the this larger sample, the 1st, 2nd and 3rd repeats generally grouped together as would be seen in vertical transmission. However, some repeats (namely those of the *Ser*H1, *Ser*H4, *Ser*H5 sequences, and the three wild sequences most similar to *Ser*H4 and *Ser*H5.) grouped most closely with other repeats from the same sequence. This pattern of branching on a neighbor-joining tree is indicative of concerted evolution. In other words a mixed mode of evolution is occurring as observed by Katz et al (2006). These results confirmed the findings that *Ser*H evolved by a mixed mode of vertical transmission and concerted evolution.

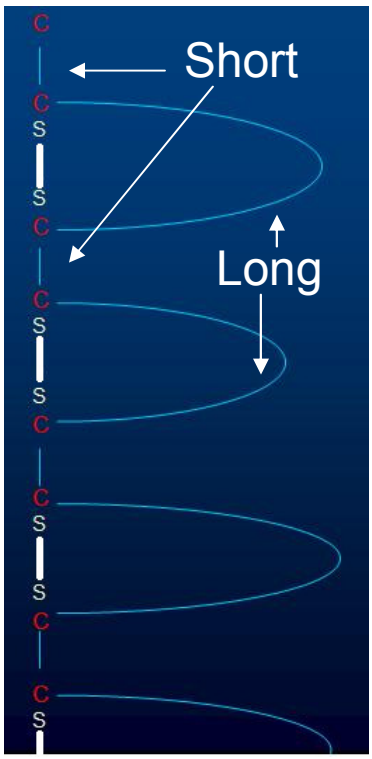
4.3 Structure of the *Ser*H Surface Protein

Lastly, this study allows more informed speculation regarding the structure of the i-antigen in relationship to the cell surface. First, for all genes examined here, the

greatest nucleotide and hence amino acid diversity was in the long amino terminal leader sequence containing 10 cysteines (which is the same number of cysteines in *Ser*J which has no such long leader). This region, opposite the carboxyl GPI linkage site, is likely the region of the molecule most exposed to the aqueous environment. The epitopes recognized by antibodies used to serotype *T. thermophila* cells could be located in that region.

Second, because all i-antigens contain periodic cysteines and because each passes through the ER for GPI attachment, it is reasonable to conclude that the cysteines form disulfide linkages while in the ER. The general pattern of the cysteines within the repeat section is (CX_{short} CX_{long}) where X_{short} is 2-7 amino acids and X_{long} is 12-18 amino acids. A speculative model is that disulfide bonds occur between CX_{short} and the CX_{long} segment then between the next CX_{short} and CX_{long}. Due to the fact that there are up to 38 cysteines in the mature *Ser*H encoded proteins there is the possibility for 703 different cysteine bonding combinations. No current disulfide bonding modeling software can accommodate anywhere near that number of possibilities. However, this pattern of periodic cysteines could produce a molecule in which the amino acids in the short group form a backbone with the disulfide bridges and the amino acids in the long group arrange as loops extending out from that backbone as shown in Fig. 19.

Fig. 19 Hypothesized model of Disulfide Bonding in the *Ser*H Protein



The white lines indicate disulfide bridges and the blue lines indicate the polypeptide chain. Note the alternating pattern of short chains and longer chains which loop.

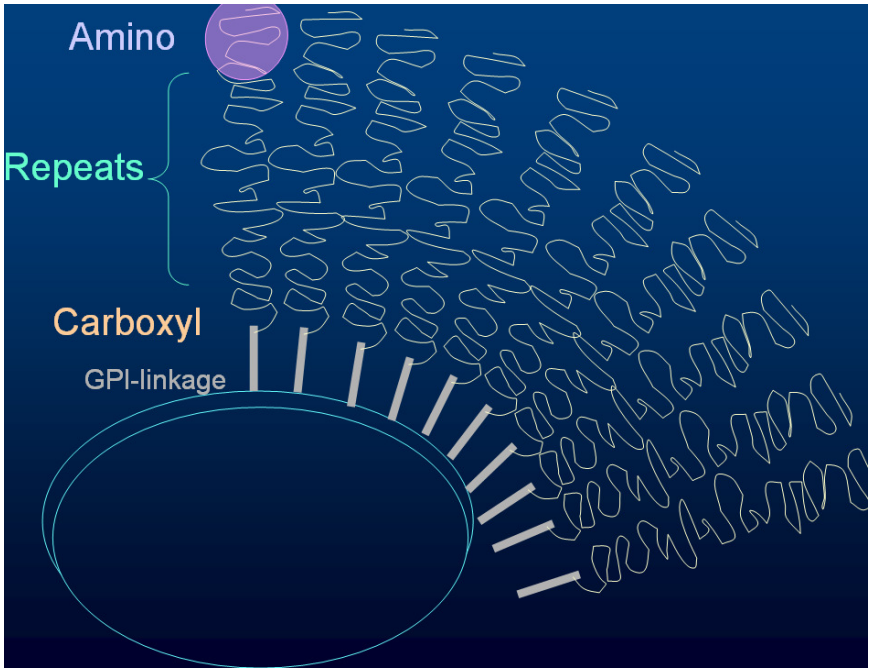
This hypothesis regarding disulfide bonding along with the well established GPI linkage allows for speculation regarding the model of the i-antigen. A previous biochemical analysis of a similar surface protein from *Paramecium aurelia* found that particular surface protein to be fibrous in nature as opposed to globular (Preer 1959). Based that, as well as the other characterizations of the *Ser*H sequence and the proposed model of disulfide bond formation a possible model of the structure of the *Ser*H protein is as follows.

The carboxyl terminus is anchored into the outer membrane via the GPI anchor, out from which the section of cysteine repeats form a fibrous extension. Finally, the amino terminus (absent the ER signal that is cleaved in the ER) extends into direct

contact with the medium (Fig 20). Because this region has more nucleotide diversity than any other section of the *Ser*H gene it is likely that these ~ 67 amino acids form the antigenic region of the protein recognized by immobilizing antibodies.

Previous speculation about the antigenic region of *Ser*H centered around the 1st repeat based on conservation of cysteine periodicity and crossreactivity of antisera against H3 with H1 (Doerder, personal communication). However, antibodies to peptide fragments of H1 and H3 first repeats both failed to immobilize and to immunoprecipitate though it is uncertain whether either peptide was indeed antigenic (Doerder, personal communication). Immobilizing antibodies to the "leader" region as well as crystallographic data on disulfide linkages would provide better tests of the proposed model.

Fig. 20 Proposed Model of *Ser*H Surface Antigen



The pink area indicates the proposed epitopic region.

This study has found evidence to confirm that the *Ser*H gene of *T. thermophila* evolves by a mix of primarily vertical transmission and occasional concerted evolution. Given the general similarity of the *Ser*H proteins to the other *T. thermophila* i-antigens (L and J are the only ones for which sequence information is published) as well as to the i-antigens of other genera (*Paramecium, Ichthyophtherius*, *Lembadion bullinum*, *Giardia lamblia* and *Pneumocystis carinii*), particularly in regards to the cysteine periodicity, it would be interesting to apply the methods used here and to determine whether these other genes show any sign of mixed evolution. Such results might inform on the evolution of genes in a system with nuclear dimorphism. It would also be helpful if future studies performed serotyping, PCR and RT-PCR on samples soon after they were brought into the lab in order to be certain of which surface antigen is being expressed and to be able to correlate that data with the macronuclear *Ser*H sequences obtained. Also, crosses should be done when samples are brought into lab to determine whether any i-antigen genes are due to recombination in macronuclear development and to verify allelic segregation.

REFERENCES

Allen, S. L., and Nanney, D. L. (1958). An analysis of nuclear differentiation in the selfers of *Tetrahymena.* Am. Nat. *92*, 139–160.

Arnheim N., Krystal M., Schmickel R., Wilson G., Ryder O., Zimmer E. (1980). Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. Proc. Natl. Acad. Sci. USA *77*, 323–7327.

Azzouz, N., Ranck, J.L. and Capdeville, Y. (1990) Purification of the temperature-specific surface antigen of *Paramecium primaurelia* with its glycosyl-phasphatidylinositol membrane anchor. Prot. Expr. Purif. *1*, 13-18.

Beale, G. H. and Kacser, H. (1957). Studies on the antigens of Paramecium aurelia with the aid of fluorescent antibodies. J. Gen. Microbiol. *17*, 68-74.

Blattner, F. R., Plunkett, G. III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides J., et al (1997). The complete genome sequence of *Escherichia coli* K-12. Science *277*, 1453–1474.

Breuer, M., G. Schulte, K. J. Schwegmann, and H. J. Schmidt. (1996). Molecular characterization of the D surface protein gene subfamily in *Paramecium tetraurelia*. J. Eukaryot. Microbiol. *43*, 14-322.

Clark, T. G. Lin, T. and Dickerson, H. W. (1995). Surface immobilization antigens of *Ichthyophthirius multifiliis*: Their role in protective immunity. Ann. Rev. Fish Diseases. *5*, 113-131.

Coen E., Strachan T., Dover G. (1982) Dynamics of concerted evolution of ribosomal DNA and histone gene families in the *melanogaster* species subgroup of *Drosophila.* JMol Biol *158*, 17–35.

Deak, J. C. and Doerder, F. P. (1995). Sequence codon usage and cysteine periodicity of the *Ser*H1 surface protein gene of *Tetrahymena thermophila*. Gene, *164*, 163–166.

Deak, J.C., and F. P. Doerder. (1998). High frequency intragenic recombination during macronuclear development in *Tetrahymena thermophila* restores the wild-type *Ser*H1 gene. Genetics *148*, 1109-1115.

Doerder F.P. and Berkowitz, M.S. (1985). Isolation and genetic analysis of the *SerH* immobilization antigen locus of *Tetrahymena thermophila*. Genetics *111*, 273-286.

Doerder, F.P. and Berkowitz, M.S. (1986). Purification and partial characterization of the H immobilization antigens of *Tetrahymena thermophila*. J. Protozool. *33*, 204-208.

Doerder, F.P. and Berkowitz, M.S. (1987). Nucleo-cytoplasmic interaction during macronuclear differentiation in ciliate protists: genetic basis for cytoplasmic control of *SerH* expression during macronuclear development in *Tetrahymena thermophila*. Genetics. *117*, 13-23.

Doerder,F.P. and Gerber,C.A. (2000). Molecular characterization of the *Ser*L paralogs of *Tetrahymena thermophila*. Biochem. Biophys. Res. Commun. *278*, 621-626.

Doerder, F. P. (2000). Sequence and expression of the *Ser*J immobilization antigen gene of *Tetrahymena thermophila* regulated by dominant epistasis. Gene. *257*, 319-326.

Doerder, F. P,. Deak, J. C. and Lief , J. H. (1991). Rate of phenotypic assortment in *Tetrahymena thermophila*. Dev. Genet. *13*,126-132.

Drummond A.J., Ashton B., Cheung M., Heled J., Kearse M., Moir R., Stones-Havas S., Thierer T. and Wilson, A. (2009). Geneious v4.6, Available from <http://www.geneious.com>.

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L, et al. (2006). Macronuclear sequence of the ciliate *Tetrahymena thermophila*. Plos Biology. *4*, 1620-1642.

Eisenhaber B., Bork P., and Eisenhaber F. (1999). Prediction of potential GPI-modification sites in proprotein sequences JMB *292*, 741-758.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J. et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae.* Science *269*, 496–512

Forney, J. D., Epstein, L. M., Preer, L. B., Rudman, B. M., Widmayer, D. J., Klein, W. H. and Preer, J. R. (1983). Structure and expression of genes for surface proteins in Paramecium. Mol. Cell Biol. *3*, 466–474.

Fu, Y.-X., and W.-H. Li,. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 639–709.

Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis Nucleic Acids Res. *31*, 3784-3788.

Gerber, C. A., Lopez, A. B. Shook, S. J. and Doerder, F. P. (2002). Polymorphism and selection at the *Ser*H immobilization antigen locus in natural populations of Tetrahymena thermophila. Genetics *160*, 1469-1479.

Gillin, F. D., Hagblom, P., Harwood, J., Aley, S. B., Reiner, D. S., Mc.Caffery, M, So, M. & Giuney, D. G. (1990). Isolation and expression of the gene for a major surface protein of Giardia lamblia. Proc. Natl. Acad. Sci. USA *87*, 4463–4467.

Goodwin, D.C., and Lee, S.B. (1993). Microwave miniprep of total genomic DNA from fungi, plants, protists and animals for PCR. Biotechniques *15*, 438-444.

Grass, F. S. (1972). An immobilization antigen in *Tetrahymena pyriformis* expressed under conditions of high salt stress. J. Protozool. *19*, 505-511.

Hansma, H. G. and Kung, C. (1975). Studies on the cell surface of Paramecium. Ciliary membrane proteins and immobilization antigens. Biochem. J. *152*, 523–528.

Hirokawa T., Boon-Chieng S., and Mitaku S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics, *14*, 378-9.

Horowitz S. and Gorovsky M. A. (1985). An unusual genetic code in nuclear genes of Tetrahymena**.** Proc. Natl. Acad. Sci. USA *82*, 2452-2455.

Jahn, C. L. and L. A. Klobutcher. (2002). Genome remodeling in ciliated protozoa. Ann. Rev. Microbiol. *56*, 489-520.

Juergensmeyer, E. B. (1969). Serotype expression and transformation in T. pyriformis. J. Protozool. *16*, 344-352.

Katz, L.A., Snoeyenbos-West, O. and Doerder, F. P. (2006). Patterns of protein evolution in *Tetrahymena thermophila*: Implications for estimates of effective population size. Mol. Biol. Evol. *23*, 608-614.

Kovacs, J. A., Powell, F., Edman, J. C., Lundgren, B., Martinez, A., Drew, B. & Angus, C. W. (1993). Multiple genes encode the major surface glycoprotein of *Pneumocystis carinii*. J. Biol. Chem., *15*, 6034–6040.

Kusch, J. and Schmidt, H. J. (2001). Genetically controlled expression of surface variant antigens in free-living protozoa. J. Membr. Biol. *180*, 101-109.

Leak, C and Forney, J. (1996). The 5' coding region of Paramecium surface antigen genes controls mutually exclusive transcription. Proc. Natl. Acad. Sci. USA. *93*, 2828 -2843.

Loefer, J. B. and Owen, R. D. (1961). Characterization and distribution of "H" serotypes in 25 °C cultures of *Tetrahymena pyriformis* variety 1. J. Protozool. *8*, 387-391.

Love, H. D. J., Allen-Nash, A., Zhao, Q. A. and Bannon, G. A. (1988). mRNA stability plays a major role in regulating the temperature-specific expression of a *Tetrahymena thermophila* surface protein. Mol. Cell Biol. *8*, 427-432.

Lynch, M., and Connery, J. S. (2003). The origins of genome complexity. Science. *302*, 1401-1404

McGrath, C. L. and L. A. Katz. (2004). Genome diversity in microbial eukaryotes. Trends Ecol. Evol. 19, 32-38.

McGrath, C. L., R. A. Zufall and L. A. Katz. (2006). Ciliate genome evolution. In Genomics and Evolution of Microbial Eukaryotes. L. A. Katz & D. Bhattacharya, eds. (Oxford University Press). pp. 64-77.

McMillan, P. J., Stanley, J. S., and Bannon, G. A. (1995). Evidence for the requirement of protein synthesis and protein kinase activity in the temperature regulated stability of a *Tetrahymena* surface protein mRNA. Nucleic Acids Research *23*, 942-953.

Nakamura, Y., Gojobori, T. and Ikemura, T. (2000). Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucl. Acids Res. *28*, 292.

Nanney D. L. (1960). The relationship between the mating type and the H serotype systems in Tetrahymena. Genetics *45*, 1351-1358.

Nei, M. and Rooney, A. (2006). Concerted and birth and death evolution of multigene families. Annu. Rev. Genet. *39*, 121 – 52.

Nielsen, E. et al. (1991). Cysteine residue periodicity is a conserved structural feature of variable surface proteins from *Paramecium tetraurelia*. J. Mol. Miol. *222*, 835-841.

Pavelitz, T., Rusché, L., Matera, A. G., Scharf, J. M. and Weiner, A. M. (1995). Concerted evolution of the tandem array encoding primate U2 snRNA occurs *in situ*, without changing the cytological context of the *RNU2* locus. EMBO *14*, 169–177.

Petersen, B., Petersen T. N., Pernille, A., Morten, N. and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Structural Biology. *9*, 51.

Peters-Regehr, T., Kusch, J. and Heckmann, K. (1997). Primary structure and origin of a predator released protein that induces defensive morphological changes in Euplotes. Eur. J. Protistol. *33*, 389–395.

Phillips, R. B. (1967). Inheritance of T serotype in *Tetrahymena*. Genetics. *56*, 667-681.

Pond, S. L. K. and Frost, S. D. W. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics *21*, 2531-2533.

Pond, S. L. K. and Frost, S. D. W. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection
Mol. Biol. Evol. *22*, 1208-1222.

Pond, S. L. K. and Frost, S. D. W. and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics. *21*, 676-679.

Prat, A., Katinka, M., Caron, F and Meyer, E. (1986). Nucleotide sequence of the *Paramecium tetraurelia* G surface protein. A huge protein with a highly periodic structure. J. Mol. Biol. *189*, 47-60.

Preer, J. R. Jr. (1959). Studies on the immobilization antigens of *Paramecium*. J. Immunol. *83*, 385-391.

Preer, J. R. Jr, Preer, L. B., Rudman, B. M. and Barnett, A. J. (1985). Deviation from the universal code shown by the gene for surface protein 51A in *Paramecium*. Nature *314*, 188–190.

Ron, A., Williams, N. E. and Doerder, F. P. (1992). The immobilization antigens of *Tetrahymena thermophila* are glycoproteins. J. Protozool. *39*, 508-510.

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X. and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics *19*, 2496-2497.

Salim, H. M.W., Ring K. L. and Cavalcanti, A. R. O. (2007). Patterns of codon usage in two ciliates that reassign the genetic code: *Tetrahymena thermophila* and *Paramecium tetraurelia*. Protist. *159*, 283—298.

Saad, Y. and Doerder, F. P. (1995). Immobilization antigen variation in natural isolates of *Tetrahymena thermophila*. Europ. J. Protistol. *31*, 45-53.

Schlotterer C, Tautz D. (1994). Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. Curr. Biol. *4*, 777–783.

Scott, J. Leeck, C. and Forney J. (1993). Molecular and genetic analyses of the B-type surface protein gene from *Paramecium tetraurelia*. Genetics. *134***,** 189-198.

Sharp, P. M., Tuohy, T. M. F. and Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. *14*, 5125–5143.

Simon, M. C. and Schmidt, H. J. (2006). Antigenic variation in ciliates: Antigen structure, function, expression. J. Euk. Microbiol. *54*, 1-7.

Smith, D. L., Berkowitz, M. S., Potoczak, D.M., Krause, C., Raab et al. (1992). Characterization of the T, L, I, S, M and P cell surface (immobilization) antigens of *Tetrahymena thermophila*: molecular weighs, molecular isoforms, and cross-reactivity of antisera. J. Protozool. *39*, 420-428.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. *24*, 1596-1599.

Tondravi, M. M., Willis, R. L., Love, H. D. Jr., and Bannon, G.A. (1990). Molecular chararacterization of *Ser*H3, a *Tetrahymena thermophila* gene encoding a temperature-regulated surface antigen. Mol. Cell Biol. *10*, 6091-6096.

Wickert, S., Nangle, L., Shevel, S., Orias, E. (2000).*Tetrahymena* macronuclear genome mapping: Colinearity of macronuclear coassortment groups and the micronuclear map on chromosome 1L. Genetics. *154*, 1155-1167.

Wright, F. (1990). The effective number of codons used in a gene. Gene *87*, 23-29.

Wuitschick, J. and Karrer, K. (1999). Analysis of Gernomic G + C Content, Codon Usage, initiator codon context and translation termination sites in *Tetrahymena thermophila*. J Euk. Mic. *46*, 239-258.

Yang, Z. and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. *15*, 496–503

Yao, M. C., Duharcourt, S. and Chalker, D. L. (2002). Genome-wide rearrangements of DNA in ciliates. in Mobile DNA II. N. L. Craig, R. Craigie, M. Gellert and Lambowitz, A. eds. (ASM Press, Washington D.C.) Pp. 730-758.

Zufall, R. A.,. McGrath, C. L., Muse, S. V. and Katz, L. A. (2006). Genome architecture drives protein evolution in ciliates. Mol. Biol. Evol. *23*, 1681-1687.

APPENDIX

Appendix A: Alignment of Deduced Amino Acid Sequences

1. 18217-2\Whole	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
2. 18217-4\	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	Q	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
3. 18218-4\	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
4. 18219-4\	M	Q	N	K	T	I	I	I	C	L	I	L	S	Q	I	L	V	S	V	F	G	-	-	-	T	G	A	P	V	T	C	T	G	-	-	-	-	T	T	C	S	T	D	C	V	T	P	T	L	T	G	-	G	S	A	T	C
5. 18221-4\Whole	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
6. 18222-4\Whole	M	Q	N	K	T	I	I	I	C	L	I	L	S	Q	I	L	V	S	V	F	G	-	-	-	T	G	A	P	V	T	C	T	G	-	-	-	-	T	T	C	S	T	D	C	V	T	P	T	L	T	G	-	G	S	A	T	C
7. 18226-2\Whole	M	Q	N	K	T	I	I	I	C	L	I	L	S	Q	I	L	V	S	V	F	G	-	-	-	T	G	A	P	V	T	C	T	G	-	-	-	-	T	T	C	S	T	D	C	V	T	P	T	L	T	G	-	G	S	A	T	C
8. 18230-1\Whole	-	Q	N	K	T	I	I	I	C	L	I	L	S	Q	I	L	V	S	V	F	G	-	-	-	T	G	A	P	V	T	C	T	G	-	-	-	-	T	T	C	S	T	D	C	V	T	P	T	L	T	G	-	G	S	A	T	C
9. 18234-1\Whole	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	S	V	I	S	-	-	-	A	G	G	A	V	T	C	T	G	-	T	T	G	Q	N	C	S	Q	A	C	A	L	P	T	V	T	G	-	G	S	T	A	C
10. 18235-2\Whole	-	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
11. 18247-2\\Whole	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
12. 18249-1\Whole	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	S	V	I	S	Q	-	F	P	G	I	E	T	T	C	T	G	-	S	P	G	Q	K	C	S	Q	A	C	A	V	P	T	V	T	G	-	G	S	G	T	C
13. 18252-4\Whole	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
14. 18253-3\Whole	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
15. 18253-4\Whole	-	Q	N	K	T	I	I	I	C	L	I	L	S	Q	F	L	V	S	V	L	S	E	-	-	Q	G	S	P	F	N	C	A	I	-	G	L	S	C	S	S	S	P	A	C	Q	S	P	Q	L	T	G	-	-	-	G	S	C
16. 18262-1\Whole	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	A	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
17. 18266-2\Whole	-	Q	N	K	T	I	I	I	C	L	I	L	S	Q	F	L	V	S	V	L	S	E	-	-	Q	G	S	P	F	N	C	A	I	-	G	L	S	C	S	S	S	P	A	C	Q	S	P	Q	L	T	G	-	-	-	G	S	C
18. 18267-1\Whole	M	Q	N	K	T	I	I	I	C	L	I	L	S	Q	I	L	V	S	V	F	G	-	-	-	T	G	A	P	V	T	C	T	G	-	-	-	-	T	T	C	S	T	D	C	V	T	P	T	L	T	G	-	G	S	A	T	C
19. 18278-3-1\Whole	-	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	A	V	F	S	T	P	A	T	G	G	S	T	T	C	T	G	-	A	S	G	S	D	C	A	N	V	C	G	V	P	T	I	A	G	T	G	T	A	T	C
20. 18280-1\Whole	M	Q	N	K	T	I	I	I	C	L	I	L	S	Q	L	L	V	S	V	F	S	-	-	A	A	G	T	Q	V	T	C	T	G	-	-	-	-	T	A	C	A	N	D	C	G	T	P	V	V	T	G	-	-	G	A	T	C
21. AF190647	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	S	V	F	S	S	-	-	A	G	G	Q	A	N	C	T	G	V	A	A	G	T	D	C	A	S	V	C	G	V	P	T	V	A	G	T	G	T	T	A	C
22. AF425241.1 1	M	Q	N	K	T	L	I	I	C	L	I	L	S	Q	L	L	V	S	V	L	S	-	-	T	A	G	A	T	I	A	C	T	A	-	P	S	T	C	A	A	S	T	D	C	T	A	P	T	L	T	G	-	G	S	T	T	C
23. AF425242.1 3	M	Q	N	K	T	L	I	I	C	L	I	L	S	N	L	L	V	S	V	F	S	T	-	T	P	G	T	N	T	P	C	T	G	-	S	T	S	C	T	A	-	-	D	C	P	K	V	T	I	G	G	-	A	T	T	A	C
24. AF425243.1 1	M	Q	N	K	T	I	I	I	C	L	I	I	S	Q	L	L	V	S	V	I	S	-	-	-	A	G	G	A	V	T	C	T	G	-	T	T	G	Q	N	C	S	Q	A	C	A	L	P	T	V	T	G	-	G	S	T	A	C
25. TTU15793 3	M	Q	N	K	T	L	I	I	C	L	I	I	S	Q	L	L	V	S	V	F	S	V	T	-	A	G	G	A	A	Q	C	P	G	T	G	A	N	C	N	V	A	A	A	C	P	V	P	T	I	Q	G	T	G	T	A	A	C

1. 18217-2\Whole	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	T	N	I	Y	A	N	G	A	G	S	A	C
2. 18217-4\	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	S	N	T	Y	A	N	G	A	G	T	A	C
3. 18218-4\	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	T	N	I	Y	A	N	G	A	G	S	A	C
4. 18219-4\	A	W	G	-	-	G	G	S	I	S	T	C	T	V	T	D	C	T	C	I	T	A	A	G	V	T	A	V	T	K	I	T	D	P	F	C	L	S	C	I	G	S	-	A	A	T	S	F	A	N	G	A	G	S	A	C
5. 18221-4\Whole	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	?	I	T	N	?	N	D	Q	F	C	T	S	C	?	G	S	-	T	T	N	I	Y	A	N	G	A	G	S	A	C
6. 18222-4\Whole	A	W	G	-	-	G	G	S	I	S	T	C	T	V	T	D	C	T	C	I	T	A	A	G	V	T	A	V	T	K	I	T	D	P	F	C	L	S	C	I	G	S	-	A	A	T	S	F	A	N	G	A	G	S	A	C
7. 18226-2\Whole	A	W	G	-	-	G	G	S	I	S	T	C	T	V	T	D	C	T	C	I	T	A	A	G	V	T	A	V	T	K	I	T	D	P	F	C	L	S	C	I	G	S	-	A	A	T	S	F	A	N	G	A	G	S	A	C
8. 18230-1\Whole	A	W	G	-	-	G	G	S	I	S	T	C	T	V	T	D	C	T	C	I	T	A	A	G	V	T	A	V	T	K	I	T	D	P	F	C	L	S	C	I	G	S	-	A	A	T	S	F	A	N	G	A	G	S	A	C
9. 18234-1\Whole	S	W	S	G	T	G	T	D	L	T	T	C	A	V	V	D	C	T	C	L	T	A	G	T	V	T	A	T	S	G	L	T	D	L	F	C	T	S	C	K	A	S	G	V	A	T	S	F	T	N	T	A	G	S	A	C
10. 18235-2\Whole	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	T	N	I	Y	A	N	G	A	G	S	A	C
11. 18247-2\\Whole	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	T	N	I	Y	A	N	G	A	G	S	A	C
12. 18249-1\Whole	S	W	L	G	T	G	N	D	L	T	T	C	F	V	N	D	C	S	C	L	N	I	G	T	I	T	-	-	-	G	L	T	D	A	F	C	Y	S	C	R	G	S	-	-	-	-	S	F	A	N	T	A	G	S	A	C
13. 18252-4\Whole	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	T	N	I	Y	A	N	G	A	G	S	A	C
14. 18253-3\Whole	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	T	N	I	Y	A	N	G	A	G	S	A	C
15. 18253-4\Whole	S	W	A	-	-	G	T	G	L	N	A	C	Y	V	V	D	C	N	C	I	S	G	P	G	K	S	G	L	T	A	V	T	D	A	F	C	N	S	C	K	G	S	-	G	P	I	T	F	A	N	I	D	G	N	A	C
16. 18262-1\Whole	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	T	N	I	Y	A	N	G	A	G	S	A	C
17. 18266-2\Whole	S	W	A	-	-	G	T	G	L	N	A	C	Y	V	V	D	C	N	C	I	S	G	P	G	K	S	G	L	T	A	V	T	D	T	F	C	N	S	C	K	G	S	-	G	P	I	T	F	A	N	I	D	G	N	A	C
18. 18267-1\Whole	A	-	G	-	-	G	G	S	I	S	T	C	T	V	T	D	C	T	C	I	T	A	A	G	V	T	A	V	T	K	I	T	G	P	F	C	L	S	C	I	G	S	-	A	A	T	S	F	A	N	G	A	G	S	A	C
19. 18278-3-1\Whole	A	W	T	-	S	G	T	T	L	T	A	C	T	V	T	D	C	N	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	S	N	T	Y	A	N	G	A	G	T	A	C
20. 18280-1\Whole	S	W	N	-	G	G	T	V	M	T	A	C	A	V	A	D	C	A	C	I	S	T	G	T	A	T	-	-	-	N	V	D	D	F	F	C	S	S	C	K	G	A	-	A	P	T	S	F	V	N	T	A	G	T	K	C
21. AF190647	S	W	V	-	S	S	S	T	L	T	T	C	T	V	T	D	C	T	C	L	T	T	G	T	V	T	G	I	T	N	L	N	D	Q	F	C	T	S	C	K	G	S	-	T	S	N	T	Y	A	N	G	A	G	T	A	C
22. AF425241.1 1	S	W	T	-	-	G	T	A	P	-	N	C	G	V	A	D	C	A	C	I	S	A	T	G	V	T	A	V	S	G	I	T	D	L	F	C	S	S	C	K	A	S	N	-	P	T	Y	F	S	N	S	A	G	T	A	C
23. AF425242.1 3	A	W	S	-	-	G	T	S	N	S	A	C	A	I	S	D	C	D	C	L	K	T	G	-	-	-	A	A	T	G	V	S	D	T	F	C	L	S	C	K	A	G	-	-	T	T	S	F	A	N	G	A	G	S	A	C
24. AF425243.1 1	S	W	S	G	T	G	T	D	L	T	T	C	A	V	V	D	C	T	C	L	T	A	G	T	V	T	A	T	S	G	L	T	D	L	F	C	T	S	C	K	A	S	G	V	A	T	S	F	T	N	A	A	G	S	A	C
25. TTU15793 3	T	W	A	-	A	G	T	D	L	T	Q	C	T	V	T	D	C	D	C	L	T	T	G	T	V	T	G	K	T	G	L	T	D	A	F	C	T	S	C	K	G	A	-	T	Q	N	L	Y	A	N	N	A	G	T	S	C

1. 18217-2\Whole	V	A	A	T	A	S	C	N	S	T	I	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
2. 18217-4\	V	A	A	S	A	S	C	N	S	T	I	R	G	-	-	-	T	T	A	W	T	V	G	D	C	T	L	C	T	P	A	T	P	A	L	V	S	G	-	-	-	-	-	-	-	?	C	K	A	C	N	T	I	S	S	A
3. 18218-4\	V	A	A	T	A	S	C	N	S	T	I	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
4. 18219-4\	V	A	A	S	A	S	C	N	S	T	T	R	G	-	-	-	S	T	Q	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	T	C	K	A	C	N	T	I	T	S	G
5. 18221-4\Whole	V	A	A	T	A	S	C	N	S	T	I	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
6. 18222-4\Whole	V	A	A	S	A	S	C	N	S	T	T	R	G	-	-	-	S	T	Q	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	T	C	K	A	C	N	T	I	T	S	G
7. 18226-2\Whole	V	A	A	S	A	S	C	N	S	T	T	R	G	-	-	-	S	T	Q	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	T	C	K	A	C	N	T	I	T	S	G
8. 18230-1\Whole	V	A	A	S	A	S	C	N	S	T	T	R	G	-	-	-	S	T	Q	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	T	C	K	A	C	N	T	I	T	S	G
9. 18234-1\Whole	V	A	S	T	A	S	C	T	A	R	P	S	-	-	-	-	G	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
10. 18235-2\Whole	V	A	A	T	A	S	C	N	S	T	I	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
11. 18247-2\\Whole	V	A	A	T	A	S	C	N	S	T	I	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
12. 18249-1\Whole	V	A	S	S	A	S	C	T	A	R	P	S	-	-	-	-	N	V	A	W	T	L	G	D	C	T	I	C	S	P	N	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	D	I	T	S	G
13. 18252-4\Whole	V	A	A	T	A	S	C	N	S	T	I	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
14. 18253-3\Whole	V	A	A	T	A	S	C	N	S	T	I	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
15. 18253-4\Whole	V	T	S	S	D	S	C	T	T	K	K	A	P	-	-	-	-	-	-	W	I	L	S	D	C	M	K	C	S	P	T	R	P	A	L	V	L	N	-	-	-	-	-	-	-	I	C	K	A	C	N	T	I	S	S	G
16. 18262-1\Whole	V	A	A	T	A	S	C	N	S	T	I	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
17. 18266-2\Whole	V	T	S	S	D	S	C	T	T	K	K	A	P	-	-	-	-	-	-	W	I	L	S	D	C	M	K	C	S	P	T	R	P	A	L	V	L	N	-	-	-	-	-	-	-	I	C	K	A	C	N	T	I	S	S	G
18. 18267-1\Whole	V	A	A	S	A	S	C	N	S	T	T	R	G	-	-	-	S	T	Q	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	T	C	K	A	C	N	T	I	T	S	G
19. 18278-3-1\Whole	V	A	A	S	A	S	C	N	S	T	I	R	G	-	-	-	T	T	A	W	T	V	G	D	C	T	L	C	T	P	A	T	P	A	L	V	S	G	-	-	-	-	-	-	-	T	C	K	A	C	N	T	I	S	S	A
20. 18280-1\Whole	V	A	S	S	A	P	C	S	N	L	Y	R	S	-	-	-	G	V	A	W	T	V	G	D	C	T	L	C	T	P	S	T	P	A	L	V	G	S	-	-	-	-	-	-	-	T	C	K	A	C	S	G	I	S	S	G
21. AF190647	V	A	A	S	A	S	C	N	S	T	I	R	G	-	-	-	T	T	A	W	T	V	G	D	C	T	V	C	T	P	T	T	P	A	L	V	G	S	-	-	-	-	-	-	-	T	C	K	A	C	N	T	I	S	S	A
22. AF425241.1 1	V	T	S	S	A	S	C	K	T	-	-	R	G	-	-	-	G	T	A	W	N	V	G	D	C	T	L	C	T	P	S	T	P	A	L	V	G	T	-	-	-	-	-	-	-	A	C	T	A	C	S	G	I	T	S	G
23. AF425242.1 3	V	A	A	S	A	S	C	N	S	T	N	R	G	-	-	-	S	T	A	W	T	V	G	D	C	T	L	C	T	P	S	T	P	A	L	V	G	T	-	-	-	-	-	-	-	T	C	T	A	C	S	G	I	S	S	S
24. AF425243.1 1	V	A	S	T	A	S	C	T	A	R	P	S	-	-	-	-	G	T	A	W	T	V	G	D	C	T	L	C	T	P	T	T	P	A	L	V	S	G	-	-	-	-	-	-	-	A	C	K	A	C	S	S	I	T	S	A
25. TTU15793 3	V	A	A	S	K	T	C	A	S	G	S	R	G	T	T	A	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	I	F	V	A	A	V	S	P	A	T	T	T	S	C	A	A	C	S	T	V	T	T	G

1. 18217-2\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	I	G	T	T	A	N	N	A	W	T	T	A	D	C	L	A
2. 18217-4\	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	?	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	S	A	?	N	A	W	T	T	A	D	C	L	A
3. 18218-4\	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	S	A	A	N	A	W	T	A	A	D	C	L	A
4. 18219-4\	W	T	D	P	N	C	N	A	C	A	T	T	A	S	P	A	T	N	T	N	F	A	N	A	A	G	T	Q	C	V	N	A	S	K	T	C	A	S	A	S	R	G	T	T	A	T	N	A	W	T	A	A	D	C	L	A
5. 18221-4\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	T	A	N	N	A	W	T	T	A	D	C	L	A
6. 18222-4\Whole	W	T	D	P	N	C	A	N	C	A	T	T	A	S	P	A	T	N	T	N	F	A	N	A	A	G	T	Q	C	V	N	A	S	K	T	C	A	S	A	S	R	G	T	T	A	T	N	A	W	T	A	A	D	C	L	A
7. 18226-2\Whole	W	T	D	P	N	C	N	A	C	A	T	T	A	S	P	A	T	N	T	N	F	A	N	A	A	G	T	Q	C	V	N	A	S	K	T	C	A	S	A	S	R	G	T	T	A	T	N	A	W	T	A	A	D	C	L	A
8. 18230-1\Whole	W	T	D	P	N	C	N	A	C	A	T	T	A	S	P	A	T	N	T	N	F	A	N	A	A	G	T	Q	C	V	N	A	S	K	T	C	A	S	A	S	R	G	T	T	A	T	N	A	W	T	A	A	D	C	L	A
9. 18234-1\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	S	A	A	N	A	W	T	A	A	D	C	L	A
10. 18235-2\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	T	A	N	N	A	W	T	T	A	D	C	L	A
11. 18247-2\\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	T	A	N	N	A	W	T	T	A	D	C	L	A
12. 18249-1\Whole	L	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	F	A	N	S	A	G	T	A	C	V	N	A	S	A	T	C	D	S	G	S	R	G	S	T	A	L	N	A	W	T	A	A	D	C	L	A
13. 18252-4\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	T	A	N	N	A	W	T	T	A	D	C	L	A
14. 18253-3\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	T	A	N	N	A	W	T	T	A	D	C	L	A
15. 18253-4\Whole	W	T	D	G	N	C	N	V	C	A	I	T	A	S	P	Q	T	R	A	I	Y	A	S	G	D	A	T	T	C	V	A	S	S	A	S	C	S	S	T	L	R	G	-	-	-	A	T	A	W	T	V	G	D	C	T	L
16. 18262-1\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	T	A	N	N	A	W	T	T	A	D	C	L	A
17. 18266-2\Whole	W	T	D	G	N	C	N	V	C	A	I	T	A	S	P	Q	T	G	A	I	Y	A	S	G	D	A	S	T	C	V	A	S	S	A	S	C	S	S	T	L	R	G	-	-	-	A	T	A	W	T	V	G	D	C	T	L
18. 18267-1\Whole	W	T	D	P	N	C	N	A	C	A	T	T	A	S	P	A	T	N	T	N	F	A	N	A	A	G	T	Q	C	V	N	A	S	K	T	C	A	S	A	S	R	G	T	T	A	T	N	A	W	T	A	A	D	C	L	A
19. 18278-3-1\Whole	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	S	A	A	N	A	W	T	T	A	D	C	L	A
20. 18280-1\Whole	W	T	D	A	N	C	N	A	C	A	K	T	A	S	T	P	T	N	N	I	Y	A	N	G	A	A	T	S	C	V	A	A	S	A	S	C	T	T	S	S	R	-	-	-	A	G	A	A	W	T	A	A	D	C	T	L
21. AF190647	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	F	A	N	S	A	G	T	A	C	V	N	A	S	A	T	C	A	S	G	S	R	G	T	T	A	A	N	A	W	T	V	A	D	C	L	A
22. AF425241.1 1	W	T	D	A	N	C	N	A	C	A	T	T	A	S	P	V	T	K	N	V	F	A	N	G	A	G	S	S	C	V	A	A	S	A	S	C	T	T	A	S	R	-	-	-	A	G	A	A	W	T	V	G	D	C	T	L
23. AF425242.1 3	W	S	D	A	N	C	N	A	C	A	T	T	A	S	P	V	T	K	N	V	F	A	N	G	A	G	S	A	C	V	A	A	S	A	S	C	N	S	T	N	R	-	-	-	G	S	T	A	W	T	V	G	D	C	T	L
24. AF425243.1 1	W	T	D	A	N	C	A	A	C	A	S	T	S	T	P	K	G	N	T	N	Y	A	N	A	A	G	T	S	C	V	N	A	S	A	T	C	A	S	A	S	R	G	T	S	A	A	N	A	W	T	A	A	D	C	L	A
25. TTU15793 3	L	T	D	S	L	C	N	A	C	G	T	N	A	S	P	A	N	N	K	I	F	A	N	A	A	G	S	A	C	V	A	S	S	L	T	C	A	S	G	S	R	G	T	T	A	G	N	A	W	T	A	A	D	C	L	A

1. 18217-2\Whole	C	T	P	A	T	P	V	?	V	P	A	A	S	G	?	?	?	T	S	C	?	A	C	S	S	A	T	S	G	L	T	D	S	L	C	A	N	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
2. 18217-4\	C	T	P	A	T	P	V	L	V	P	A	A	S	G	S	A	T	T	S	C	I	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
3. 18218-4\	C	T	P	A	T	P	V	L	V	P	A	A	S	G	S	V	T	T	S	C	I	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
4. 18219-4\	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
5. 18221-4\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
6. 18222-4\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
7. 18226-2\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
8. 18230-1\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
9. 18234-1\Whole	C	T	P	A	T	P	V	L	V	P	A	A	S	G	S	V	T	T	S	C	I	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
10. 18235-2\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
11. 18247-2\\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
12. 18249-1\Whole	C	T	P	A	T	P	I	L	V	P	A	A	Q	G	S	T	T	T	S	C	A	A	C	S	T	V	N	T	G	L	T	D	T	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
13. 18252-4\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
14. 18253-3\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
15. 18253-4\Whole	C	T	P	S	T	P	A	F	V	G	T	A	-	-	-	-	-	-	-	C	V	A	C	N	T	I	S	S	G	W	S	D	A	N	C	A	N	C	A	A	G	G	S	-	-	-	-	A	V	F	A	N	T	A	G	S
16. 18262-1\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
17. 18266-2\Whole	C	T	P	S	T	P	A	F	V	G	T	A	-	-	-	-	-	-	-	C	V	A	C	N	T	I	S	S	G	W	S	D	A	N	C	N	A	C	A	A	G	G	S	-	-	-	-	A	V	F	A	N	T	A	G	S
18. 18267-1\Whole	C	T	P	A	T	P	V	F	V	P	A	A	S	G	Q	T	A	T	S	C	V	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
19. 18278-3-1\Whole	C	T	P	A	T	P	V	L	V	P	A	A	S	G	S	A	T	T	S	C	I	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
20. 18280-1\Whole	C	T	P	S	T	P	A	L	V	S	G	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21. AF190647	C	T	P	A	T	P	V	F	V	P	A	A	S	P	A	V	T	T	S	C	V	A	C	S	A	A	T	S	G	L	N	D	A	L	C	N	A	C	A	S	S	A	S	P	A	A	K	T	T	F	A	N	T	A	G	S
22. AF425241.1 1	C	T	P	S	T	P	A	L	V	G	T	T	-	-	-	-	-	-	-	C	T	A	C	S	G	I	T	S	G	W	T	D	A	N	C	N	A	C	A	T	T	A	S	P	A	T	K	N	V	F	A	N	G	A	G	S
23. AF425242.1 3	C	T	P	S	T	P	A	L	V	G	T	T	-	-	-	-	-	-	-	C	T	A	C	S	G	I	S	S	S	W	T	D	A	N	C	N	A	C	A	T	T	A	S	P	V	T	K	N	V	F	A	N	G	A	G	S
24. AF425243.1 1	C	T	P	A	T	P	V	L	V	P	A	A	S	G	S	V	T	T	S	C	I	A	C	S	S	A	T	S	G	L	T	D	S	L	C	N	A	C	A	S	S	A	S	P	P	A	K	T	T	F	A	N	T	A	G	S
25. TTU15793 3	C	T	P	A	T	P	I	F	V	A	A	V	S	P	S	A	N	T	S	C	A	A	C	S	T	V	T	T	G	L	T	D	S	L	C	N	A	C	G	T	N	A	S	P	A	N	N	K	I	F	A	N	A	A	G	S

1. 18217-2\Whole	A	C	V	A	S	S	A	T	C	?	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
2. 18217-4\	A	C	?	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
3. 18218-4\	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
4. 18219-4\	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
5. 18221-4\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
6. 18222-4\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
7. 18226-2\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
8. 18230-1\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
9. 18234-1\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
10. 18235-2\Whole	A	S	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
11. 18247-2\\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
12. 18249-1\Whole	A	C	V	A	S	S	A	T	C	T	T	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
13. 18252-4\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
14. 18253-3\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
15. 18253-4\Whole	S	C	V	A	A	S	F	S	C	N	I	N	L	R	G	T	N	K	W	N	D	T	D	C	G	L	C	N	G	T	T	A	N	A	A	Q	Y	A	S	A	D	G	S	M	C	V	A	A	S	A	S	C	I	T	R	G
16. 18262-1\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
17. 18266-2\Whole	S	C	V	A	A	S	S	S	C	N	I	N	L	R	G	T	N	K	W	N	D	T	D	C	G	L	C	N	G	T	T	A	N	A	A	Q	Y	A	S	A	D	G	S	M	C	V	A	A	S	A	S	C	I	T	R	G
18. 18267-1\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
19. 18278-3-1\Whole	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
20. 18280-1\Whole	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21. AF190647	A	C	V	A	S	S	A	T	C	T	A	G	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	A
22. AF425241.1 1	S	C	V	A	A	S	A	S	C	T	T	A	S	R	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
23. AF425242.1 3	A	C	V	A	A	S	A	S	C	N	S	T	N	R	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
24. AF425243.1 1	A	C	V	A	S	S	A	T	C	T	A	A	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T
25. TTU15793 3	A	C	V	A	S	S	A	T	C	A	G	G	S	R	G	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	L

1. 18217-2\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	?	G	?	?	P	?	?	T	S	?	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
2. 18217-4\	A	N	A	W	T	A	A	D	C	L	A	C	?	P	A	T	P	A	V	Q	S	G	T	A	P	A	T	S	-	-	C	V	A	C	T	S	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
3. 18218-4\	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	I	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
4. 18219-4\	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	S	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	R	A	M	V	A	S	P	Q
5. 18221-4\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	F	G	A	S	P	L	T	T	S	S	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
6. 18222-4\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	S	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
7. 18226-2\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	S	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
8. 18230-1\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	S	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
9. 18234-1\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	I	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
10. 18235-2\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	F	G	A	S	P	L	T	T	S	S	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
11. 18247-2\\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	F	G	A	S	P	L	T	T	S	S	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
12. 18249-1\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	F	G	A	S	P	L	T	T	S	S	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
13. 18252-4\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	F	G	A	S	P	L	T	T	S	S	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
14. 18253-3\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	F	G	A	S	P	L	T	T	S	S	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	N	P	Q
15. 18253-4\Whole	S	A	A	W	T	L	D	D	C	A	I	C	N	P	N	S	P	A	L	V	G	-	-	-	-	-	-	-	-	N	T	C	T	A	C	T	S	I	T	S	A	W	T	D	A	N	-	-	A	C	A	I	V	S	K	A	K
16. 18262-1\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	F	G	A	S	P	L	T	T	S	S	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
17. 18266-2\Whole	S	A	A	W	T	L	D	D	C	A	I	C	N	P	N	S	P	A	L	V	G	-	-	-	-	-	-	-	-	N	T	C	T	A	C	T	S	I	T	S	A	W	T	D	A	N	C	N	A	C	A	I	V	S	K	A	K
18. 18267-1\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	S	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
19. 18278-3-1\Whole	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	S	G	T	A	P	A	T	S	-	-	C	V	A	C	T	S	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
20. 18280-1\Whole	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	A	A	C	S	G	I	S	S	G	W	T	D	D	N	C	N	A	C	A	K	T	A	S	T	P
21. AF190647	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	F	G	A	S	P	A	T	T	S	S	C	V	A	C	N	T	I	N	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
22. AF425241.1 1	G	A	A	W	T	V	S	D	C	T	L	C	T	P	S	T	P	A	L	V	G	-	-	-	-	-	-	-	-	T	T	C	T	A	C	S	G	I	T	S	G	W	T	D	A	N	C	N	A	C	A	T	T	A	S	P	A
23. AF425242.1 3	S	T	A	W	T	V	G	D	C	T	L	C	T	P	S	T	P	A	L	V	G	-	-	-	-	-	-	-	-	T	T	C	T	A	C	S	G	I	S	S	S	W	T	D	A	N	C	N	A	C	A	T	T	A	S	P	V
24. AF425243.1 1	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	I	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q
25. TTU15793 3	A	N	A	W	T	A	A	D	C	L	A	C	T	P	A	T	P	A	V	Q	L	G	A	S	P	A	T	T	S	S	C	V	A	C	N	T	I	T	S	G	W	T	D	A	N	C	N	S	C	A	M	A	A	S	P	Q

1. 18217-2\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
2. 18217-4\	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	?	D	G	S	?	C	V	A
3. 18218-4\	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
4. 18219-4\	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
5. 18221-4\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
6. 18222-4\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
7. 18226-2\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
8. 18230-1\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
9. 18234-1\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
10. 18235-2\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
11. 18247-2\\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
12. 18249-1\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
13. 18252-4\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
14. 18253-3\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
15. 18253-4\Whole	T	N	F	A	N	T	A	G	S	S	C	V	A	A	S	F	S	C	N	K	T	T	R	S	A	N	K	W	T	D	T	D	C	G	L	C	N	G	T	T	A	N	A	A	Q	Y	A	S	T	D	G	S	M	C	V	A
16. 18262-1\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
17. 18266-2\Whole	T	N	F	A	N	T	A	G	S	S	C	V	A	A	S	F	S	C	N	K	T	T	R	S	A	N	K	W	T	D	T	D	C	G	L	C	N	G	T	T	A	N	A	A	Q	Y	A	S	T	D	G	S	M	C	V	A
18. 18267-1\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
19. 18278-3-1\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
20. 18280-1\Whole	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N	N	I	Y	A	N	V	G	A	T	S	C	V	A
21. AF190647	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
22. AF425241.1 1	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	V	F	A	S	G	T	G	S	S	C	V	A
23. AF425242.1 3	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	V	F	A	N	G	A	G	S	A	C	V	A
24. AF425243.1 1	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	N	I	V	A	K	A	D	G	S	A	C	V	A
25. TTU15793 3	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	T	I	V	A	K	A	D	G	S	A	C	V	A

1. 18217-2\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2. 18217-4\	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	?	S	-	-	-	S	T	F	S	G	Q
3. 18218-4\	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
4. 18219-4\	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
5. 18221-4\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
6. 18222-4\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
7. 18226-2\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
8. 18230-1\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
9. 18234-1\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
10. 18235-2\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
11. 18247-2\\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
12. 18249-1\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
13. 18252-4\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
14. 18253-3\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
15. 18253-4\Whole	T	S	Y	S	C	N	Q	T	T	R	G	S	N	K	W	T	N	I	D	C	A	A	C	N	G	T	A	S	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	K	A	A	I	T	A	S	T	F	S	G	Q
16. 18262-1\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
17. 18266-2\Whole	T	S	Y	S	C	N	Q	T	T	R	G	S	N	K	W	T	N	I	D	C	A	A	C	N	G	T	A	S	N	A	N	Q	Y	A	S	A	D	G	S	A	C	Q	A	T	K	A	A	I	T	A	S	T	F	S	G	Q
18. 18267-1\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
19. 18278-3-1\Whole	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
20. 18280-1\Whole	A	S	A	S	C	T	Q	P	A	R	G	S	T	K	W	T	D	A	D	C	A	A	C	N	G	T	A	A	N	A	T	Q	Y	A	S	A	D	G	S	S	C	Q	A	T	K	P	S	-	-	-	S	T	F	S	G	Q
21. AF190647	A	V	F	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
22. AF425241.1 1	A	S	Y	S	C	N	Q	T	T	R	G	S	N	K	W	S	D	A	D	C	A	L	C	N	G	S	A	S	N	A	N	Q	Y	A	S	A	D	G	S	S	C	Q	A	T	K	A	S	-	-	-	S	T	F	S	G	Q
23. AF425242.1 3	A	S	Y	S	C	N	Q	T	A	R	G	S	N	K	W	T	D	A	D	C	A	L	C	N	G	T	T	S	N	A	N	Q	F	A	S	A	D	G	S	S	C	Q	S	T	K	P	S	-	-	-	S	T	F	S	G	Q
24. AF425243.1 1	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	N	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	T	C	Q	A	T	Q	A	S	-	-	-	S	T	F	S	G	Q
25. TTU15793 3	A	V	Y	S	C	T	Q	S	A	R	G	S	N	K	W	T	D	A	D	C	A	A	C	N	G	T	A	A	N	A	N	Q	Y	A	S	A	D	G	S	S	C	Q	A	T	K	A	S	G	-	S	S	T	F	S	G	Q

1. 18217-2\Whole	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2. 18217-4\	I	F	V	S	I	L	L	V	L	S	A	L	L	I
3. 18218-4\	I	F	V	S	I	L	L	V	L	S	A	L	L	I
4. 18219-4\	I	F	V	S	I	L	L	V	L	S	A	L	L	I
5. 18221-4\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
6. 18222-4\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
7. 18226-2\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
8. 18230-1\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
9. 18234-1\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
10. 18235-2\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
11. 18247-2\\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
12. 18249-1\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
13. 18252-4\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
14. 18253-3\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
15. 18253-4\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
16. 18262-1\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
17. 18266-2\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
18. 18267-1\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
19. 18278-3-1\Whole	I	F	V	S	I	L	L	V	L	S	A	L	L	I
20. 18280-1\Whole	I	F	V	S	T	L	L	V	L	S	A	L	L	I
21. AF190647	I	F	V	S	I	L	L	V	L	S	A	L	L	I
22. AF425241.1 1	I	F	V	S	T	L	L	V	L	S	A	L	L	I
23. AF425242.1 3	I	F	V	S	T	L	L	V	L	S	A	L	L	I
24. AF425243.1 1	I	F	V	S	I	L	L	V	L	S	A	L	L	I
25. TTU15793 3	I	F	V	S	I	L	L	V	L	S	A	L	L	I