INTRAGENIC VARIATION OF SYNONYMOUS SUBSTITUTION RATES

KAZUHISA TSUNOYAMA

DOCTOR OF PHILOSOPHY

DEPARTMENT OF GENETICS
SCHOOL OF LIFE SCIENCE
GRADUATED UNIVERSITY FOR
ADVANCED STUDIES

1998

Acknowledgments

I would like to express my sincere gratitude to Professor Takashi Gojobori for his continuous guidance and encouragement during all the stages of this work. I also greatly appreciate Professor Yoshio Tateno, Drs. Naruya Saitou, Kazuho Ikeo, and Tadashi Imanishi for their valuable suggestions and advice.

I would like to thank Professors Toshimichi Ikemura, Ken Nishikawa, Sadao Ishiwa, and Teruo Yasunaga for serving on my supervisory committee.

I am grateful to the fellow researchers and staff of the Center for Information Biology for their help and cooperation during various stages of this study. In particular, I thank Dr. Hidemi Watanabe for his tough but instructive guidance. I also thank Dr. Matthew I. Bellgard for his collaboration in part of this study.

I also appreciate Professor Naoto Sakamoto at University of Tsukuba for giving me a chane to study at the National Institute of Genetics.

Finally, I would like to give my special thanks to my parents for their never-changing support.

Abstract

In the protein-coding DNA sequences, nucleotide substitutions are classified into synonymous and nonsynonymous substitutions. When a nucleotide substitution does not cause an amino acid change, it is called a synonymous substitution. When the nucleotide substitution does cause an amino acid change, it is called a nonsynonymous substitution. The synonymous substitution is, by definition, free from functional constraints of a protein whereas the nonsynonymous substitution is essentially constrained by protein function. Thus, it is expected that for a given gene, the rate of synonymous substitution is constant as long as the mutation rate is constant, and synonymous substitutions take place more frequently than nonsynonymous substitutions. It follows that the difference between the numbers of synonymous and nonsynonymous substitutions reflects the degree of functional importance for a protein, meaning that the larger the degree, the deeper the difference.

I found that these properties of synonymous and nonsynonymous substitutions could be utilized for evaluating the functional importance for subunits as well as domains of a protein. Moreover, I successfully showed that the rate of synonymous substitution is variable not only among genes but also within a gene. I also found that for mammals, the intragenic variation of synonymous substitutions is mainly caused by the nonrandom

mutation due to methylation of CpG dinucleotides.

In chapter I, I described the outline of the present thesis, placing particular emphasis on the motivation and purpose of my study. In chapter II, taking nicotinic acetylcholine receptor subunit genes as an example, I examined the degree of functional importance of subunits by conducting the comparison analysis of the numbers of synonymous and nonsynonymous substitutions. In particular, calculating the ratio (*f*) of the number of nonsynonymous substitutions to that of synonymous substitutions, I showed that the α_1 and α_7 subunits had the lowest *f* values in the muscle and nervous systems, respectively. These results suggested that very strong functional constraints work on these subunits. These findings are consistent with the fact that these subunits have crucial functions for the receptor; the α_1 subunit has binding sites to the ligand and the α_7-containing receptor regulates the release of the transmitter. Moreover, the window analysis of the *f* values showed that strong functional constraints work on the so-called M2 region in all 5 types of the muscle subunits. Note that the M2 region corresponds to a hole of the ion channel in the receptor molecule. Thus, calculation of the *f* values is useful for evaluating the degree of functional importance of not only a gene but also the subregion within a gene. In chapter III, I conducted a statistical test to examine whether the rate of synonymous substitution varies within a gene, by using 418 homologous gene pairs between *Rattus norvegicus* and *Mus musculus*,

as well as 84 orthologous gene pairs between the whole bacterial genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. I found that more than 90% of gene pairs for both comparisons showed significant intragenic variation of synonymous substitution rates. By examining all conceivable possibilities for the cause of the intragenic variation of synonymous substitution rates, I found a significant correlation between synonymous substitution rates and the frequency of CpG dinucleotides in rodents. These findings suggest that intragenic variation of synonymous substitutions in mammals is caused mainly by a nonrandom mutation due to the methylation of CpG dinucleotides. In chapter IV, I described the summary and conclusion of the present study, and I also discussed the future development of this line of study.

Table of Contents

ACKNOWLEDGMENTS II

ABSTRACT III

TABLE OF CONTENTS VI

LIST OF TABLES VIII

LIST OF FIGURES IX

CHAPTER I: INTRODUCTION 1

1.1 Nucleotide substitution mutation and nucleotide substitution 2

1.2 Synonymous and nonsynonymous substitutions 3

1.3 Synonymous and nonsynonymous substitutions, and gene function 4

1.4 Intergenic variation of synonymous substitutions 5

1.5 Methods for estimating the numbers of synonymous and nonsynonymous substitutions 6

1.6 The purposes of the present study 8

CHAPTER II: ANALYSIS OF SYNONYMOUS AND NONSYNONYMOUS SUBSTITUTION RATES WITH SPECIAL REFERENCE TO EVOLUTION OF NICOTINIC ACETYLCHOLINE RECEPTOR SUBUNITS 11

2.1 Introduction 12

2.2 Materials and Methods 15

2.2.1 Sequence analysis 15
2.2.2 Calculating the f ratio 18
2.2.3 Window analysis for the f values 19

2.3 Results 20
2.3.1 Phylogenetic analysis of nAChR subunits 20
2.3.2 The controversy over evolution of subunits 24
2.3.3 Evaluation of the degree of functional constraints 25
2.3.4 Window analysis 28

2.4 Discussion 32

2.5 Questions to be addressed 36

CHAPTER III: INTRAGENIC VARIATION OF SYNONYMOUS SUBSTITUTION RATES 38

3.1 Introduction 39

3.2 Materials and Methods 43
3.2.1 Data extraction 43
3.2.2 Statistical test for intragenic variation of synonymous substitutions 45
3.2.3 Methods for examining the cause of intragenic variation of synonymous substitutions 46

3.3 Results 49
3.3.1 A significant intragenic variation of synonymous substitution rates 49
3.3.2 Possible causes of intragenic variation 53

3.4 Discussion 66

CHAPTER IV: SUMMARY 69

REFERENCES 74

List of Tables

2.1: 84 sequences of nAChR subunits and 1 sequence as an outgroup used in this study. 16

3.1: The number of gene pairs which showed statistically significant variation of the Ps values within a gene. 50

3.2: Examination of possible causes of intragenic variation of synonymous substitutions. 62

List of Figures

2.1: The Phylogenetic tree for the nAChR subunits. 21

2.2: The evolutionary mechanism of all the nAChR subunits. 23

2.3: Evaluation for the degree of functional importance for each subunit type in both nervous and muscle systems. 27

2.4: The window analysis for the subunits in the muscle system. 30

2.5: The cross section of single nAChR in the muscle system. 31

2.6: Intragenic variation in the number (ds) of synonymous substitutions, the number (dn) of nonsynonymous substitutions, and the ratio (f= dn/ds) for nAChR α_1 subunit. 37

3.1: Four examples for the intragenic variation of Ps values. 62

3.2: Estimated mRNA secondary structures of histone subunit 1 genes of both mouse and rat. 63

3.3: The frequency distribution for the probabilities of 316 correlation coefficients in each correlation analysis. 64

3.4: An example of a pairwise alignment of a interleukin-1 receptor accessory protein. 65

CHAPTER I:

INTRODUCTION

1.1 Nucleotide substitution mutation and nucleotide substitution

Mutation is a source of genetic variation and one of the most fundamental processes of molecular evolution. There are many types of mutations in the DNA sequences: They are nucleotide substitution, insertion and deletion, recombination, inversion, transposition, and so forth. In the study of molecular evolution, nucleotide substitution mutation is often focused on because its molecular mechanism has been relatively better known compared with other types of mutations.

Once a nucleotide substitution mutation takes place in a gene sequence, the mutant of having such a nucleotide change may increase or decrease in its frequency within a population, depending upon whether the mutant is selectively advantageous or deleterious. Of course, genetic drift also plays an important role to change the mutant frequency in a population particularly when the population size is small. If the mutant is selectively neutral, only the genetic drift contributes to the change of the mutant frequency (Kimura 1983). Anyway, the mutant frequency will eventually become 1 or 0, meaning that the nucleotide substitution mutation will spread the entire population or will disappear from the population. When a mutant frequency becomes 1, we call it "fixation of the mutant in a population." In particular, we say that "nucleotide substitution" takes place when a nucleotide substitution mutation is fixed in a population. Because "nucleotide substitution" is likely to be confused with "nucleotide substitution mutation," I would call nucleotide substitution mutation simply as mutation in the present study. This would be helpful to avoid possible confusions, because the type of mutations which I would deal with in the present study is mostly nucleotide substitution mutation.

1.2 Synonymous and nonsynonymous substitutions

In protein-coding regions of DNA sequences, there are two types of nucleotide substitutions. One is a synonymous substitution, which does not cause any amino acid change. The other is a nonsynonymous substitution, which alters an amino acid.

The synonymous substitution is, by definition, free from functional constraints of a protein. If no constraint other than protein function is imposed on synonymous substitutions, the rate of synonymous substitution is equal to the mutation rate because the synonymous substitution should be selectively neutral. Thus, it is expected that for a given gene, the rate of synonymous substitution is constant over time as well as space as long as the mutation rate is constant. On the other hand, the nonsynonymous substitution is constrained directly by protein function. Because most amino acid-altering nucleotide substitutions have, in general, deleterious effects on the function and structure of a protein, most nonsynonymous substitutions are selected out from a population. Thus, synonymous substitutions take place more frequently than nonsynonymous substitutions. It follows that the difference between the numbers of synonymous and nonsynonymous substitutions reflects the degree of functional importance for a protein, meaning that the larger the degree, the deeper the difference. This is exactly what the neutral theory of molecular evolution has predicted (Kimura 1968 and 1983).

1.3 Synonymous and nonsynonymous substitutions, and gene function

Comparisons of the numbers between synonymous and nonsynonymous substitutions have been conducted for evaluating the functional importance for genes. Under the neutral theory of molecular evolution, it is expected that purifying selection works on most of genes (Kimura 1968). The purifying selection is also called negative selection by which deleterious mutations are eliminated from a population. Because most of nonsynonymous substitutions are actually deleterious, it follows that for most of genes, the number of synonymous substitutions is larger than that of nonsynonymous. If a gene has more important function, the difference in the numbers between synonymous and nonsynonymous substitutions becomes larger. Thus, the comparison between the number of synonymous and nonsynonymous substitutions can be used for evaluation of the degree of functional constraints among genes.

It has been also used for detection of genes on which positive selection operates. This is because the genes on which positive selection operates are considered to have an evolutionary characteristic where the number of nonsynonymous substitutions is larger than that of synonymous substitutions. By adopting this characteristic as the criterion, several genes have been studied as candidate genes on which positive selection may operate (for review, see Endo et al. 1995). However, the number of candidate genes on which positive selection may operate is very limited. Only approximately 5% of all examined genes showed a possibility of positive selection (Endo et al. 1995).

1.4 Intergenic variation of synonymous substitutions

If the rate of synonymous substitutions is fairly uniform among different genes, the examination of only nonsynonymous substitution rates is sufficient to evaluate the relative degree of functional importance and constraints among genes. In reality, however, many studies have recently reported that the rate of synonymous substitution is not actually constant over time and genes. In particular, it is quite variable among examined genes (Graur 1985; Li et al. 1985; Wolfe et al. 1989; Bernardi et al. 1993; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995). The cause of such intergenic variation of synonymous substitutions is still unclear. It could be due to intergenic differences in codon usage bias, base composition, and the mutation rate. Anyway, because of intergenic variation of synonymous substitutions, it is essential to compare the number of synonymous substitutions with that of nonsynonymous substitutions, in order to evaluate the degree of functional importance, or constraints of a gene product.

1.5 Methods for estimating the numbers of synonymous and nonsynonymous substitutions

When the number of nucleotide substitutions between a given pair of genes is very small, the number of synonymous and nonsynonymous substitutions can be obtained simply by counting silent and amino acid-altering nucleotide differences, respectively. However, it is, in general, difficult to make a distinction between synonymous and nonsynonymous substitutions for more distantly diverged gene pairs. For this reason, many methods have been proposed to estimate the number of synonymous and nonsynonymous substitutions (for review, see Nei 1987).

Among them, Miyata and Yasunaga's (MY) method (1980), Li, Wu, and Luo's (LWL) method (1985), and Nei and Gojobori's (NG) method (1986) have been widely used. In particular, the NG method has been used most frequently because of its simplicity. These three methods should be used with some caution, because these methods may give overestimates of the number of synonymous substitutions and underestimates of the number of nonsynonymous substitutions, particularly when the bias of transition versus transversion is stronger (Ina 1995).

The so-called PBL method (Pamilo and Bianchi 1993; Li 1993) and Ina's method (1995) were developed, because these two methods give better estimates than the MY, LWL, and NG methods unless there are strong transition/transversion and nucleotide frequency biases (Ina 1995). In order to solve the problem for transition/transversion bias, Comeron (1995) modified the PBL method. Moreover, Moriyama and Powell (1997) developed new method taking base composition into consideration. However, the authors conceded that the estimates obtained from these new methods were very close to Comeron's method. Since PBL and Comeron's methods include nonsense mutation in the category of nonsynonymous substitutions, estimates sometimes become negative when sequence divergence is low, whereas Ina's method is often inapplicable when closely related sequences are used (Zhang et al. 1998). Thus, these newly developed methods have a limitation for available data sets.

More recently, Zhang et al. (1998) modified the NG method. This modified NG method corrects transition/transversion bias in the estimation. Therefore, the modified NG method is the best choice when analyzing large data sets. Thus, we used the modified NG method in the present study.

1.6 The purposes of the present study

As I have already mentioned earlier, the comparisons between the numbers of synonymous and nonsynonymous substitutions have been used for evaluating relative importance among different genes. In this study, I found that these comparisons of the numbers of synonymous and nonsynonymous substitutions could be also utilized for evaluating the functional importance for subunits as well as domains of a protein. Moreover, I successfully showed that the rate of synonymous substitution is really variable not only among genes but also within a gene. The intragenic variation has been somewhat controversial because the previous studies used only specific genes or lacked statistical reliability (Clark and Kao 1991; Lawrence et al. 1991; Eyre-Walker and Bulmer 1993; Ina et al. 1994; Cacciò et al. 1995; Zoubak et al. 1995; Comeron and Aguadé 1996). Thus, it is essential to compare the number of synonymous substitutions with that of nonsynonymous substitutions in order to evaluate the functional constraints even within a gene. I also pointed out that for mammals, the intragenic variation of synonymous substitutions is mainly caused by the nonrandom mutation due to methylation of CpG dinucleotides.

The purposes of the present study are to show these original findings by presenting substantial amount of evidence and to discuss their biological implications with special reference to molecular evolution of genes.

In chapter II, taking nicotinic acetylcholine receptor subunit genes as an example, I examined the degree of functional importance of subunits by conducting the comparison analysis of the numbers of synonymous and nonsynonymous substitutions. I used nicotinic acetylcholine receptor subunit genes because these receptor subunits have diverged from the common ancestor to 16 types of subunits and the function of subunits is very different from each other. By calculating the ratio (f) of the number of nonsynonymous substitutions to that of synonymous substitutions, I showed that α_1 and α_7 subunits had the lowest f values in the muscle and nervous systems, respectively. Thus, I showed that calculation of the f values is useful for evaluating the degree of functional importance of not only among genes but also within a gene.

In chapter III, I conducted a statistical test to examine whether the rate of synonymous substitution really varies within a gene. For this analysis, I used 418 homologous gene pairs between *Rattus norvegicus* and *Mus musculus*, as well as 84 orthologous gene pairs between the whole bacterial genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. As a result, I found that more than 90% of gene pairs for both comparisons showed statistical significance in intragenic variation of synonymous substitution rates. By examining all conceivable possibilities for the cause of the intragenic variation of synonymous substitution rates, I found a significant correlation between synonymous substitution rates and the frequency of CpG dinucleotides in rodents. I discussed the biological implication of this finding.

Finally, in chapter IV, I described the summary and conclusion of the present studies, and I also discussed the future development of this line of study.

CHAPTER II:

ANALYSIS OF SYNONYMOUS AND NONSYNONYMOUS SUBSTITUTION RATES WITH SPECIAL REFERENCE TO EVOLUTION OF NICOTINIC ACETYLCHOLINE RECEPTOR SUBUNITS

2.1 Introduction

Acetylcholine (ACh) has long been recognized as a neurotransmitter active in nervous and muscle systems of Bilateria. The nicotinic acetylcholine receptors (nAChR), focused on this study, belong to the superfamily of receptors involved in ligand-gated ion channels in both nervous and muscle systems. It is known that ACh binding causes the ionic channel open and cations to pass through, resulting in changes of the electrical state of the target cell (Changeux 1990). Because the nAChR in the muscle system was the first receptor to be isolated, the nAChR of striated muscles is the best-characterized member of the super family (Changeux 1990). In particular, it is known that the nAChR in the muscle system is a hetero-oligomer composed of five subunits, each with four transmembrane domains (Karlin 1993; Galzi and Changeux 1994). For subunits in the muscle subunits, α_1, β_1, γ, and δ, were named according to differences in the molecular weight and the ε subunit was named according to the Greek alphabet.

Although the function and structure of nAChR in the nervous system are not understood as clearly as in the muscle system, it is known that there are eight α-type (named $\alpha_2 - \alpha_9$) and three β-type (classified as $\beta_2 - \beta_4$) subunits. They are produced in the whole nervous system and even in non-neuronal tissues such as striated muscle, lymphocytes, granulocytes, skin, bones, and so forth. Because the α_9 subunit is expressed in the outer hair cells of the cochlear, it is proposed that this subunit is responsible for the signal transduction at the synapse between efferent neuronal terminals and cochlear hair cells (Elgoyhen et al. 1994).

It is of particular interest to understand how all 16 subunits, α_1, β_1, γ, δ, and ε in the muscle system and α_2 - α_9 and β_2 - β_4 in the nervous systems, are evolutionarily related to each other. In spite of the intensive previous works, the controversy over the evolutionary history of these subunits has not been settled yet (Ortells and Lunt 1995; Le Novère and Changeux 1995; Gundelfinger 1995). As for the α_1 subunit in the muscle system, it has not been clear whether this subunit has evolved from the neural α-type subunit. Moreover, it has been controversial whether the β_2 and β_4 subunits in the nervous system are evolutionarily close to the other neural subunits. In order to resolve the controversy and to elucidate the evolutionary mechanisms of these subunits, I constructed a phylogenetic tree using amino acid sequences.

I also estimated the ratio, *f*, of the number of nonsynonymous substitutions to that of synonymous substitutions through comparisons between sequences of subunits to infer a degree of functional importance. Moreover, I could successfully identify the sequence regions where strong functional constraints were imposed, using the so-called window analysis.

Thus, this method is very useful for evaluating the degree of functional constraints.

2.2 Materials and Methods

2.2.1 Sequence analysis

I collected the data from the DDBJ/EMBL/GenBank nucleotide sequence database by keyword search. I excluded the redundant data of subunit types for each species from the analysis and obtained 84 nucleotide sequences from 18 different species (See table 1.1). First, the amino acid sequences translated from the nucleotide sequences were aligned with each other by CLUSTAL W version 1.6 (Thompson et al. 1994). In this alignment, I used BLOSUM series for a protein weight matrix, 10 for a gap opening penalty, and 0.05 for a gap extension penalty. To estimate the number of amino acid substitutions per site, I used Kimura's method (Kimura 1983) for only sites commonly shared by all sequences. I took the sequence of serotonin-gated ion channel (5HT3) as an outgroup. I then constructed a phylogenetic tree by the neighbor-joining (N-J) method (Saitou and Nei 1987). The bootstrap resampling tests were performed 1000 times to confirm reliability of the constructed tree.

Table 2.1:

84 sequences of nAChR subunits and 1 sequence as an outgroup used in this study.

Species	Subunit	abbreviation	Acc. no.	References
Homo sapiens	α1	a1_human	X02502	Noda et al. 1983
	α2	a2_human	U62431	Elliott et al. 1996
	α3	a3_human	M37981	Mihovilovic and Roses 1991
	α4	a4_human	L35901	Monteggia et al. 1995
	α5	a5_human	M83712	Chini et al. 1992
	α6	a6_human	U62435	Elliott et al. 1996
	α7	a7_human	X70297	Peng et al. 1994
	β1	b1_human	X14830	Beeson et al. 1989
	β2	b2_human	X53179	Anand and Lindstrom 1990
	β3	b3_human	X67513	Willoughby et al. 1993
	β4	b4_human	X68275	Tarroni et al. 1992
	δ	d_human	X55019	Luther et al. 1989
	ε	e_human	X66403	Beeson et al. 1993
	γ	g_human	X01715	Shibahara et al. 1985
Bos taurus	α1	a1_cow	X02509	Noda et al. 1983
	α7	a7_cow	X93604	Garcia-Guzman et al. 1995
	β1	b1_cow	X00962	Tanabe et al. 1984
	δ	d_cow	X02473	Kubo et al. 1985
	ε	e_cow	X02597	Takai et al. 1985
	γ	g_cow	M28307	Takai et al. 1984
Rattus norvegicus	α2	a2_ratn	L10077	Wada et al. 1988
	α3	a3_ratn	X03440	Boulter et al. 1986
	α4	a4_ratn	L31620	Goldman et al. 1987
	α5	a5_ratn	J05231	Boulter et al. 1990
	α6	a6_ratn	L08227	Boulter Unpublished
	α7	a7_ratn	L31619	Boulter Unpublished
	β3	b3_ratn	J04636	Deneris et al. 1989
	β4	b4_ratn	M33953	Boulter et al. 1990
	ε	e_ratn	X13252	Criado et al. 1988
Mus musculus	α1	a1_mouse	X03986	Isenberg et al. 1986
	α7	a7_mouse	L37663	Orr-Urtreger et al. 1995
	β1	b1_mouse	M14537	Buonanno et al. 1986
	δ	d_mouse	L10076	Boulter Unpublished
	ε	e_mouse	X55718	Gardner 1990

	γ	g_mouse	X03818	Yu et al. 1986
Rattus rattus	α1	a1_ratnr	X74832	Witzemann et al. 1990
	α3	a3_ratr	L31621	Boulter et al. 1987
	α9	a9_ratr	U12336	Elgoyhen et al. 1994
	β1	b1_ratn	X74833	Witzemann et al. 1990
	β2	b2_ratr	L31622	Boulter et al. 1987
	δ	d_ratr	X74835	Witzemann et al. 1990
	γ	g_ratn	X74834	Witzemann et al. 1990
Gallus gallus	α1	a1_chick	X07330	Nef et al. 1988
	α2	a2_chick	X07340	Nef et al. 1988
	α3	a3_chick	M37336	Couturier et al. 1990
	α4	a4_chick	X07348	Nef et al. 1988
	α5	a5_chick	J05642	Couturier et al. 1990
	α6	a6_chick	U48860	Gerzanich et al. Unpublished
	α7	a7_chick	X68586	Couturier et al. 1990
	α8	a8_chick	X52296	Schoepfer et al. 1990
	β2	b2_chick	X53092	Schoepfer et al. 1988
	β3	b3_chick	X83739	Hernandez et al. 1995
	δ	d_chick	K02903	Nef et al. 1984
	γ	g_chick	K02904	Nef et al. 1984
Xenopus laevis	α1	a1_frog	X07067	Noda et al. 1982
	β1	b1_frog	U04618	Kullberg et al. 1994
	δ	d_frog	X07069	Baldwin et al. 1988
	ε	e_frog	U19612	Murray et al. 1995
	γ	g_frog	X07068	Baldwin et al. 1988
Danio rerio	α1	a1_zebraf	U70438	Sepich et al. Unpublished
Carassius auratus	α3	a3_gfish	X54051	Hieber et al. 1990
	β2	b2_gfish	X54052	Hieber et al. 1990
	GFn α2	na-2_gfish	X14786	Cauley 1989
	GFn α3	na-3_gfish	M29529	Cauley et al. 1990
Torpedo marmorata	α1	a1_torpedom	M25893	Devillers-Thiery et al. 1983 Devillers-Thiery et al. 1984

Torpedo californica	α1	a1_torpedoc	J00963	Noda et al. 1982;
				Sumikawa et al. 1982;
				Numa 1983; Devillers-Thiery 1983

	β1	b1_torpedoc	J00964	Numa et al. 1983; Noda et al. 1983
	δ	d_torpedoc	J00965	Numa et al. 1983; Noda et al. 1983
	γ	g_torpedoc	J00966	Ballivet et al. 1982; Numa et al. 1983; Noda et al. 1983; Claudio et al. 1983
Caenorhabditis elegans	α	a_Caenorhabditis	X98600	Fleming et al. Unpublished
	β1	b1_Caenorhabditis	X83888	Alliod and Ballivet Unpublished

	non-α	non-a_Caenorhabditis	X86403	Squire et al. 1995
	lev-1 gene	lev1_Caenorhabditis	X98246	Fleming et al. Unpublished
	ACR-3	acr-3_Caenorhabditis	Y08637	Baylis Unpublished
Haemonchus contortus	α	a_Haemonchus	U72490	Hoekstra et al. 1997
Onchocerca volvulus	unknown	?_Onchocerca	L20465	Ajuh and Egwang 1994
Drosophila melanogaster	α2	a2_fruitfly	X53583	Sawruk et al. 1990
	α-like	a-like_fruitfly	X07194	Bossy et al. 1988
	β2	b2_fruitfly	X55676	Sawruk et al. 1990
	β 64B	b-64B_fruitfly	M20316	Wadsworth et al. 1988
Myzus persicae	α1	a1_Myzus	X81887	Sgard et al. Unpublished
	α2	a2_Myzus	X81888	Sgard et al. Unpublished
Manduca sexta	α-like	a-like_Manduca	Y09795	Eastham et al. Unpublished
Schistocerca gregaria	αL1	a-L1_Schistocerca	X55439	Marshall et al. 1990
Mus Musculus	5HT3	5HT3_mouse	M74425	Maricq et al. 1991

2.2.2 Calculating the *f* ratio

Based on all available pairs between nucleotide sequences of different subunits, I estimated the number of synonymous and nonsynonymous substitutions using the method of Nei and Gojobori (Nei and Gojobori 1986). In this calculation, I considered only sites shared by all sequences. The ratio (f) of the number (d_n) of nonsynonymous substitutions to that (d_s) of synonymous substitutions, $f = d_n / d_s$, was then calculated in order to evaluate the degree of functional constraint. These f values were calculated as the average of all pairwise comparisons between sequences of subunit types from different species. When a given pair of sequences compared showed the saturated number of synonymous substitutions, I excluded this pair from the comparison.

2.2.3 Window analysis for the *f* values

In order to identify particular regions where functional constraints are imposed, I conducted a window analysis by modifying Endo et al.'s original method (Endo et al. 1996), using human and rat preprotein sequences of each subunit type in the muscle system, i.e. α_1, β_1, γ, δ, and ε. I used human and rat sequences because the saturation effect of synonymous substitutions between these two sequences can be much smaller than that between more distantly related species. In my modified window analysis, I calculated the *f* value for each window along the nucleotide sites codon by codon, although Endo et al.'s original method was developed for computing the inverse of *f*. In my modified widow analysis, a window size was defined as a sequence region with a 46-codon length on an alignment, because this length is the minimum to avoid inapplicable cases where one cannot obtain the number of synonymous substitutions. These inapplicable cases are due to Jukes and Cantor's correction of multiple substitutions (Jukes and Cantor 1969) in the method of Nei and Gojobori (1986).

2.3 Results

2.3.1 Phylogenetic analysis of nAChR subunits

Figure 2.1 shows a phylogenetic tree for 84 amino acid sequences from 18 different species in the muscle and nervous systems (See table 2.1). My phylogenetic tree shows that the α_7, α_8, and α_9 subunits are direct descendants of the ancestral subunit. These subunits can yield functional receptors as homo-oligomers, whereas the other α-type subunits need β-type subunits to yield a single nAChR (Couturier et al. 1990; Elgoyhen et al. 1994). All the subunits in the muscle system except α_1 are evolutionarily close to each other, suggesting that these subunits in the muscle system may have come from the same origin. Interestingly enough, these muscle subunits are close to the β_2 and β_4 subunits in the nervous system. Moreover, the α_1 subunit in the muscle system is closer to all the subunits in the nervous system except α_7-α_9, β_2, and β_4. In other words, the phylogenetic tree did not show clear divergence of subunits between the muscle and nervous systems.

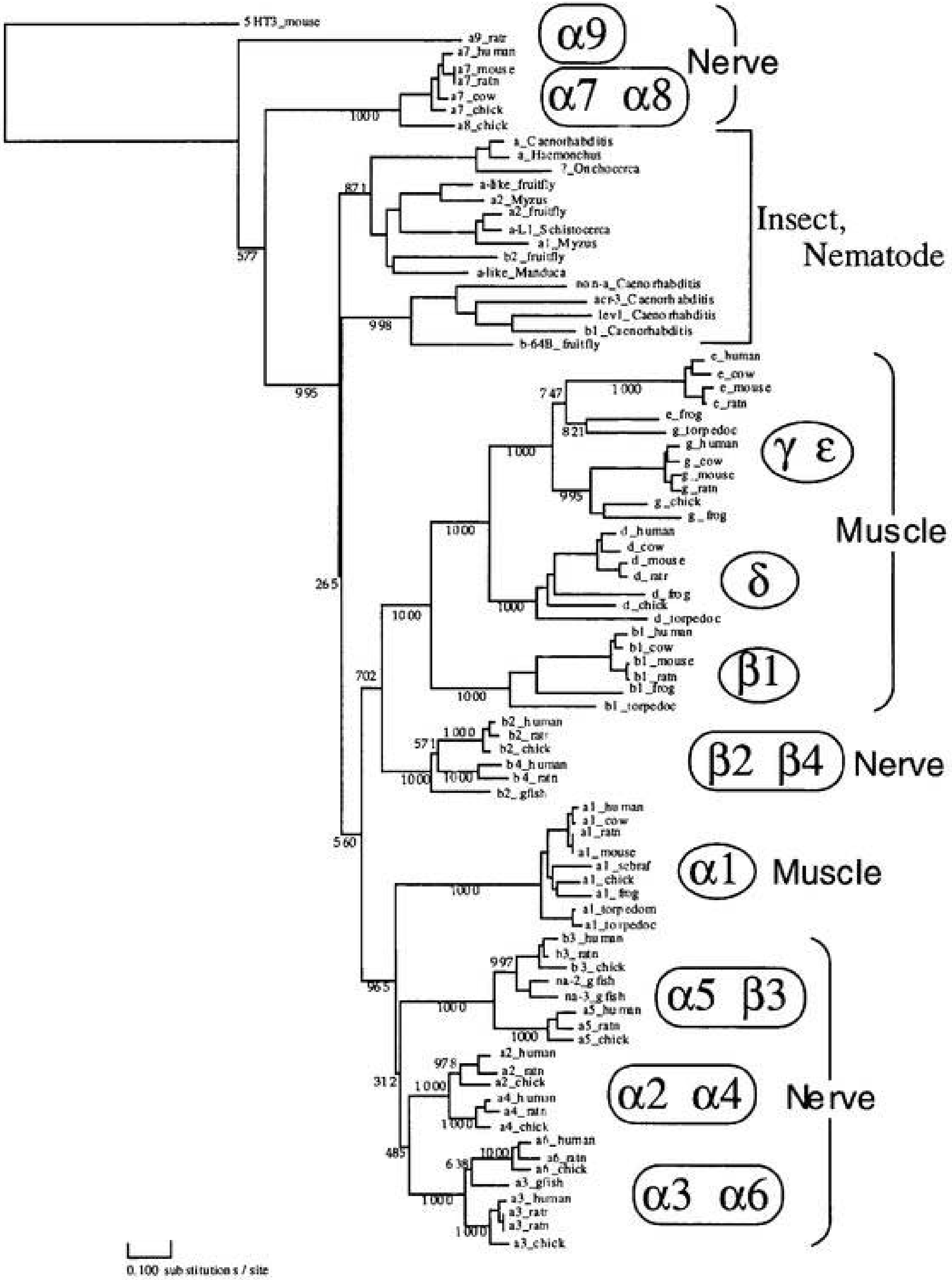


Figure 2.1: **The phylogenetic tree for the nAChR subunits. The bootstrap values are indicated at the corresponding nodes.**

My phylogenetic tree leads to the following interpretation: The ancestral subunit which had a ligand-binding site appeared first in the nervous system. This ancestral subunit may have functioned as homo-oligomers in the primitive Bilateria. This is because the α_7, α_8, and α_9 subunits in the nervous system, all of which function as homo-oligomers, have diverged first from the common ancestor. Before appearance of Deuterostomia, the subunits used in insects and nematodes may have diverged from the subunits used in Deuterostomia. The subsequent duplication may have occurred during the evolution of Deuterostomia, and may have yielded two types of subunits (fig. 2.2). These two types were the ancestral α-type subunits which maintained the function of binding to ligands and the ancestral β-type subunits which have lost the function. Present-day α-type subunits, α_1 - α_6 and β_3 emerged from the ancestral α-type subunit, and the present-day β_1, β_2, β_4, γ, δ, and ε subunits have also diverged from the ancestral β-type subunit. The ancestral β-type subunit have lost the function of binding to ligands and may have changed the function into complementing the binding sites of the α-type subunit. Switching the tissue in which subunits are expressed from the nerve to the muscle and vice versa produced the present-day combination of subunits.

I compared my phylogenetic tree with those in the previous works (Ortells and Lunt 1995; Le Novère and Changeux 1995; Gundelfinger

1995).

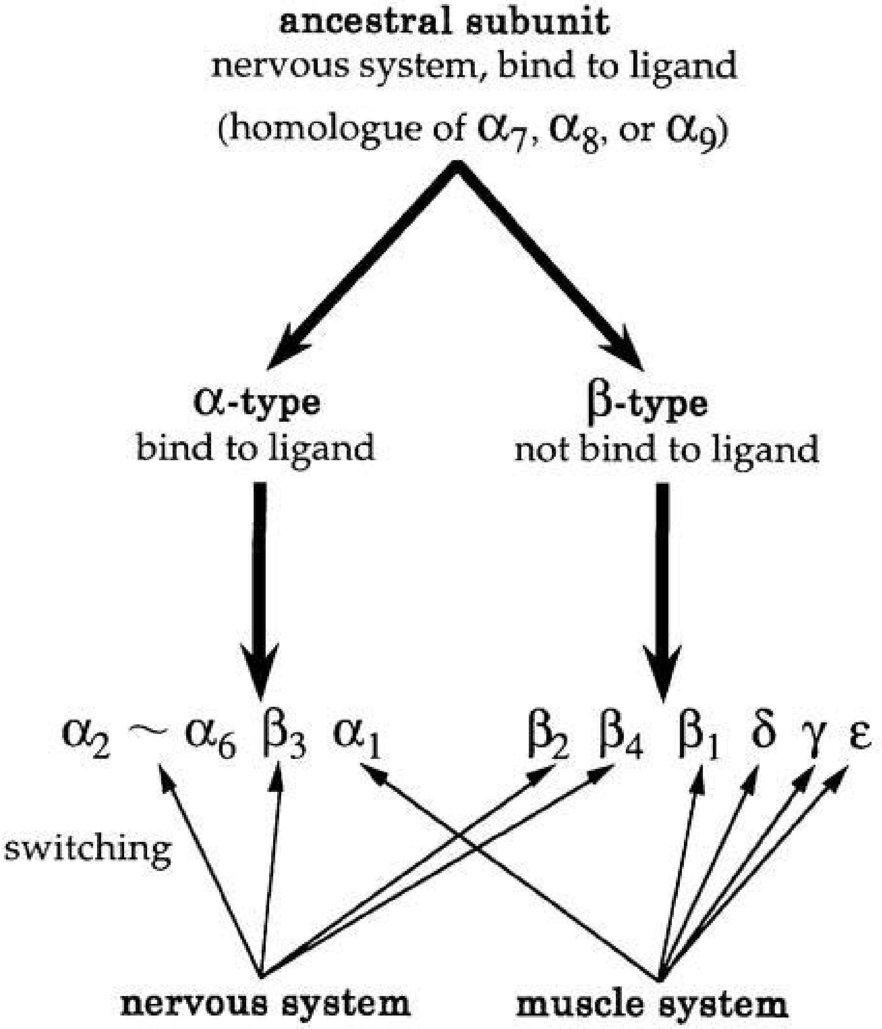

ancestral subunit
nervous system, bind to ligand
(homologue of α_7, α_8, or α_9)
α-type
bind to ligand
β-type
not bind to ligand
$\alpha_2 \sim \alpha_6$ β_3 α_1
β_2 β_4 β_1 δ γ ε
switching
nervous system
muscle system

Figure 2.2: The evolutionary mechanism of all the nAChR subunits.

My tree contains more number of subunits for invertebrates and the α_9 subunit of rat. My interpretation that the common ancestor of all subunits may have appeared first in the nervous system, is supported by previous studies (Ortells and Lunt 1995; Le Novère and Changeux 1995).

In fact, my phylogenetic tree suggests that the α_7, α_8, and α_9 subunits are direct descendants of the ancestral subunit. In particular, the α_9 subunit diverged from the ancestor at the earliest time. This observation substantiates the previous speculations by Sivilotti and Colquhoun (1995) and Changeux et al. (1996). Moreover, my phylogenetic tree clearly indicated that all subunits of insects and nematodes emerged after the divergence of the α_7, α_8, and α_9 subunits from the common ancestor.

2.3.2 The controversy over evolution of subunits

There has been a heated debate over the evolution of vertebrate subunits in both nervous and muscle systems (Ortells and Lunt 1995; Le Novère and Changeux 1995; Gundelfinger 1995). Although there is no question about that two subunit groups separated after the emergence of the insect and nematode subunits, there was sharp disagreement among different research groups on the constituting members of the two groups. Ortells and Lunt (1995) postulated that the α_1 - α_6 and β_3 subunits belong to one group whereas the β_1, β_2, β_4, γ, δ, and ε subunits belong to the other group. However, Le Novère and Changeux (1995) insisted that the muscle α_1 subunit and all the nervous subunits (α_2 - α_6 and β_2 - β_4) constitute one group and all the muscle subunits except α_1 (β_1, γ, δ, and ε) constitute the other group. Being different from these two research groups, Gundelfinger

(1995) hypothesized that one of two groups consists of all the muscle subunits (α_1, β_1, γ, δ, and ε).

My phylogenetic tree showed that the separation took place between the α_1 - α_6 and β_3 subunits and the β_1, β_2, β_4, γ, δ, and ε subunits, suggesting that the α_1 subunit in the muscle system has evolved from an ancestral α type subunit in the nervous system. Thus, it is consistent with the view of Ortells and Lunt (1995), but it disagrees with Le Novère and Changeux (1995) and Gundelfinger (1995). Indeed, Le Novère and Changeux's topology of the phylogenetic tree differs from mine. In constructing a phylogenetic tree, I employed the N-J method using the amino acid sequences. Though I also constructed a phylogenetic tree using the number of nucleotide substitutions at the 1st and 2nd positions of codons, the topology obtained was virtually the same as before. Note that I could not use the 3rd positions of codons because of the saturation effects in some comparisons. Moreover, I constructed a phylogenetic tree by the maximum likelihood method using amino acid sequences, and the tree obtained had the same topology as the one presented in figure 2.1. Thus, my phylogenetic tree is considered to be the most reliable so far.

2.3.3 Evaluation of the degree of functional constraint

The numbers of synonymous and nonsynonymous substitutions per nucleotide site, d_s and d_n, were estimated from all pairwise comparisons between nucleotide sequences from different species for each subunit type. If a subunit is important for functions, the amino acid sequence does rarely change because functional constraints are imposed on the subunit. When strong functional constraints work on a subunit, the number (d_s) of synonymous substitutions is greater than that (d_n) of nonsynonymous substitutions, resulting in the ratio, *f*, being close to zero. Thus, calculating the *f* value for each subunit type allows me to evaluate the degree of functional importance of the nAChR subunits.

Figure 2.3 shows the *f* value for each type of subunits. All subunits are conserved well, but I can still evaluate the degree of functional importance for each subunit by the differences of the *f* value. In the muscle system, the α_1 subunit is the lowest value for the *f* value. The value was about one third the highest value which was for the δ subunit. The confidence interval with reliability of 95% is from 0.03 to 0.09 for the α_1 subunit and from 0.10 to 0.19 for the δ subunit. This implies that the strongest functional constraint is imposed on the α_1 subunit in the muscle system. This is possibly because this subunit has binding sites to the ligand. The ε subunit had the second lowest *f* value among subunits expressed in the muscle system. Thus, stronger functional constraints are also working on the ε subunit, which suggests that the ε subunit is

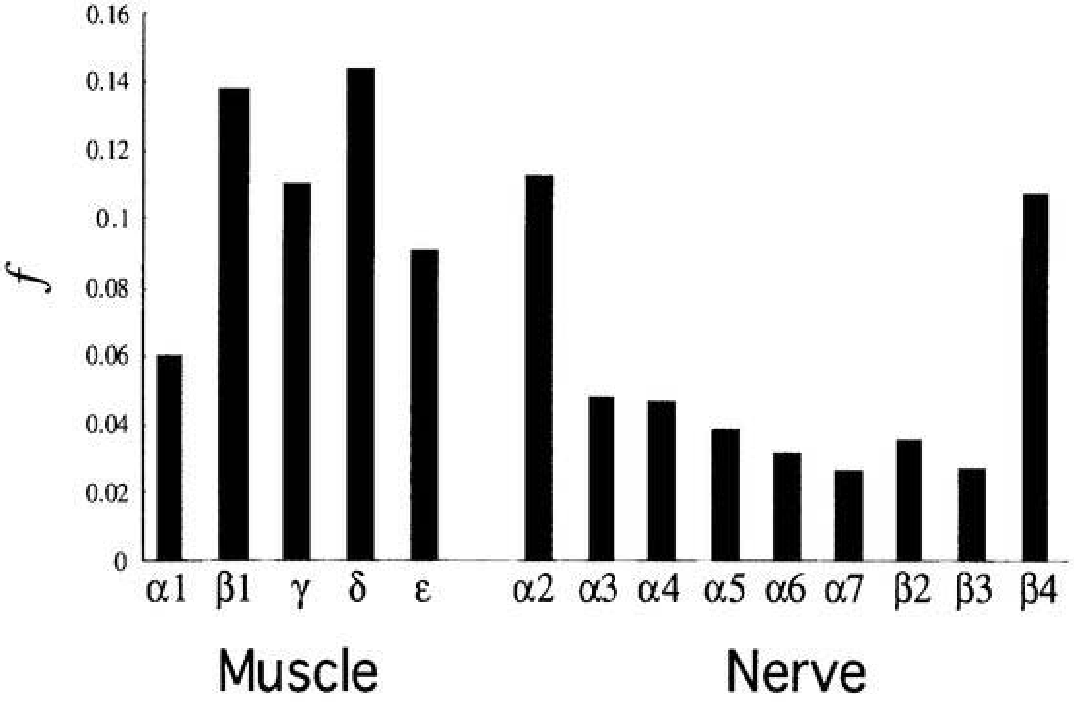


Figure 2.3: Evaluation for the degree of functional importance for each subunit type in both nervous and muscle systems. A vertical line shows the *f* value for each subunit type.

important for the function of the nAChR in the muscle system. This is consistent with the fact that the ε subunit can change kinetics of binding with the ligands through single mutations (Ohno et al. 1995). Moreover, it is known that the ε subunit has a complementary part of the ligand binding sites. Therefore, stronger functional constraints may be also imposed on the ε subunit.

As for the nervous system, in spite of many kinds of experiments, the functions of most subunits (α_2 - α_9, β_2 - β_4) are not known. The lowest value in the nervous system was for α_7 subunit. The value was about one fifth the highest value of the α_2 subunit. The confidence interval with reliability of 95% is from 0.02 to 0.03 for the α_7 subunit and from 0.07 to 0.15 for the α_2 subunit. This low value for the α_7 subunit suggests that the α_7 subunit has a crucial function for the nAChR in the nervous system. This is supported by experimental results showing that the α_7 subunit regulates the release of neurotransmitters in central nervous system (McGehee et al. 1995). Because this subunit is most conserved, the low *f* value of the α_7 subunit is reasonable (See figs. 2.1, 2.2, and 2.3).

2.3.4 Window analysis

As shown above, I can recognize functionally important subunits for the nAChR by evaluating the degree of functional constraint imposed on the subunits. In general, however, I do not know where the functional constraints work on. To answer this question, I conducted my modified window analysis to identify particular sequence regions where stronger functional constraints work on. In practice, I used the nucleotide sequences of α_1, β_1, γ, δ, and ε subunit sequences in the muscle system of human and rat. I defined the window size as a sequence region with a 46-

codon length on an alignment of the same subunit type. This length is the minimum to avoid inapplicable cases where one cannot obtain the number of nonsynonymous or synonymous substitutions. For each window shifted along the nucleotide sites codon by codon, the *f* value was calculated.

Figure 2.4 shows the results of my modified window analysis for subunits in the muscle system. Some sequence regions show the lower *f* value, implying that the functional constraints are at work strongly on the regions. In particular, two regions of α_1 subunits (dashed arrow in fig. 2.4) show the lower values, suggesting that these regions have crucial functions. I found that these regions correspond to acetylcholine binding sites. Moreover, all the subunits of α_1, β_1, γ, δ, and ε in the muscle system showed a common sequence region where strong functional constraints work on (black arrow in fig. 2.4). This region may have a specific function in common to these five subunits, because these subunits are used to construct a single nAChR in the muscle system (fig. 2.5). The existence of this region is consistent with the fact that for all subunits in the muscle system, this region is used to construct a hole in the ion channel. This region is called the M2 region (Changeux 1990).

The result of the window analysis revealed a sequence region which had a specific function for nAChR. I could also find some regions where strong functional constraints are at work, but their functions are not clear.

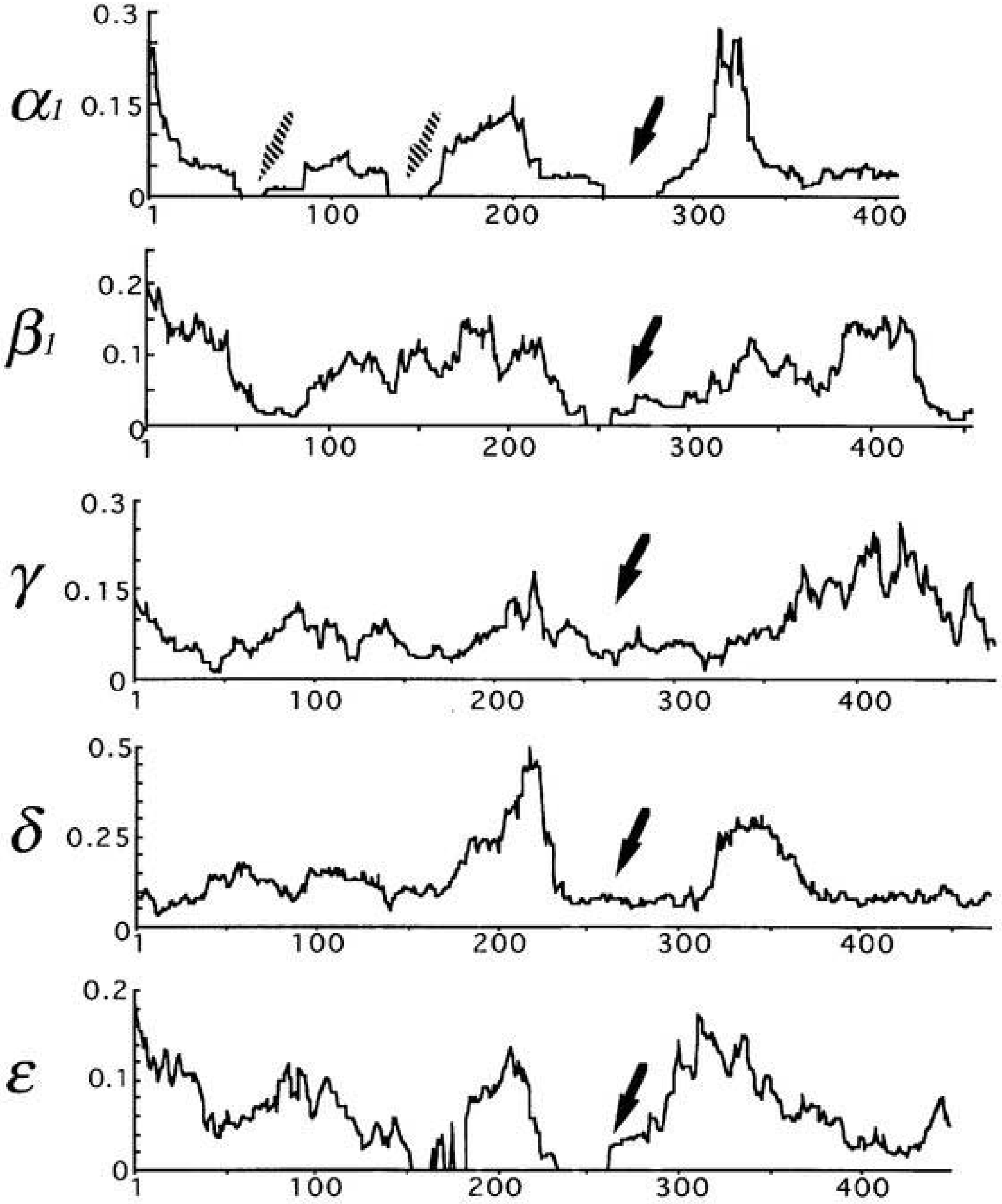


Figure 2.4: The window analysis for the subunits in the muscle system. The vertical line shows the *f* value and the horizontal line shows the amino acid site number of aligned sequences. Black and dashed arrows indicate the sequence regions which show the lower *f* values.

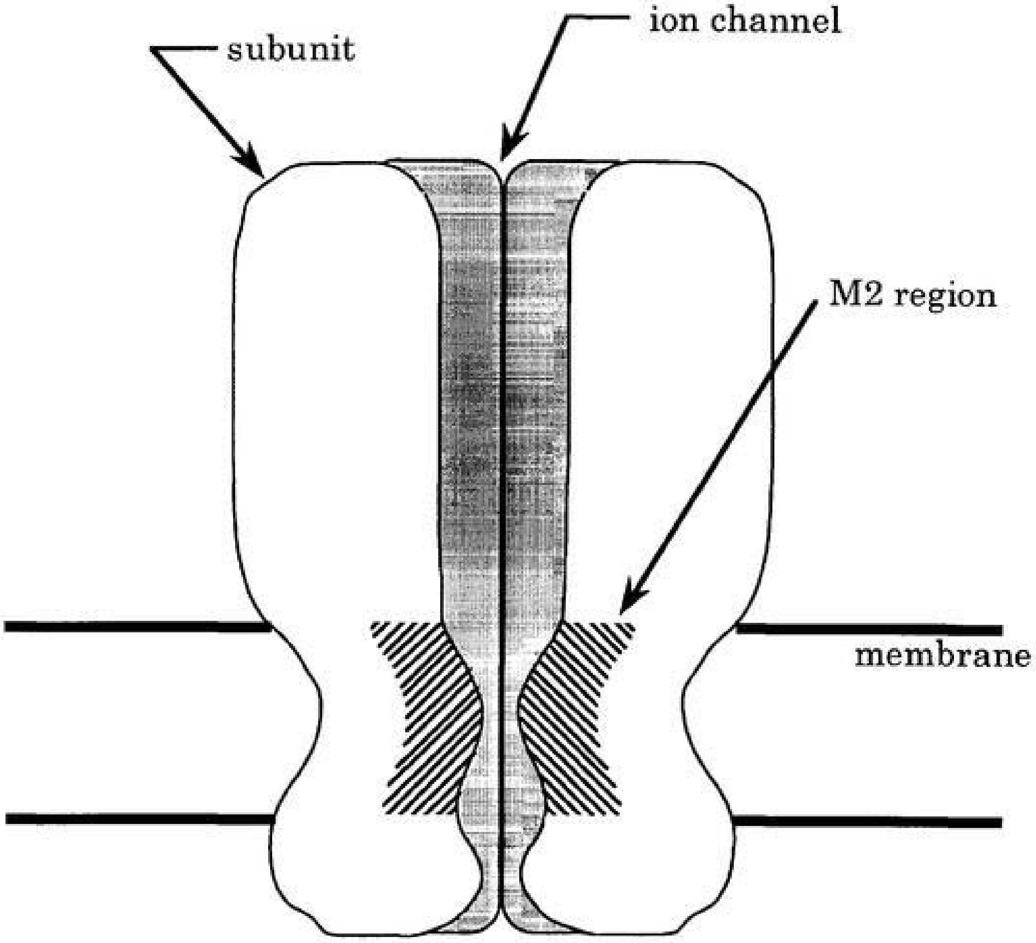


Figure 2.5: The cross section of single nAChR in the muscle system. The M2 region is marked by diagonal lines.

2.4 Discussion

I studied the evolution of nicotinic acetylcholine subunits by constructing phylogenetic trees, and showed that the α_1 subunit in the muscle system have evolved from an ancestral subunit in the nervous system. I also showed that stronger functional constraints work on the subunit which has crucial functions, and that the functional constraints work on a particular sequence region which has a specific function.

The genomic structure of all subunits showed a general tendency that the exon-intron structures of the former half of genes appeared similar to each other whereas those of the latter half seemed quite different. I, then, divided each sequence into two fragments: one fragment having 140 amino acids from the N-terminal and the other fragment for the rest of sequence. The former fragment contains exons 1-4 of almost all the subunits, whose genomic structure is well conserved. I constructed phylogenetic trees for the two fragments, separately, by the N-J method. The phylogenetic trees for both fragments showed that the α_1 subunit is evolutionarily close to the α_2 - α_6 and β_3 subunits. This observation does not support the proposal by Gundelfinger (1995). Moreover, there is the possibility that the β_2 and β_4 subunits in the nervous system have undergone recombination to become closer to muscle subunits. This is because in the phylogenetic tree for the latter fragment, the β_2 and β_4 subunits are close to the neural subunits. In fact, the former fragment of the β_2 and β_4 subunit genes may have come from the former fragment of muscle subunit genes by recombination. On the other hand, the α_1 subunits may have functioned in the muscle system by switching the tissue.

In my phylogenetic tree of subunits, I have paraphyletic cases for some subunit types. For instance, there are such cases for the α_3 and β_2 subunits from a gold fish and the γ subunit from a ray. I, however, cannot discriminate whether these cases occurred by the divergence or by the choice of data used even if the corresponding bootstrap value is very high, because I have only a few sequences of Pisces.

In the analysis of a degree of functional importance, I calculated the *f* value. For subunits in the muscle system, fig. 2.3 shows that the α_1 subunit has the lowest *f* value in the muscle system. The α_1 subunit may be the most conserved subunit in the muscle system, because the α_1 subunit diverged first in the muscle system and binds to the ligand (See fig. 2.1). The ε subunit also has a lower *f* value. Though the ε subunit does not have the same degree of functional importance as the α subunit, it has a complementary part of the binding sites of the α_1 subunit and its change can affect binding to the ligand. Thus, stronger functional constraints also work on this subunit. The γ and δ subunits have also a complementary part of the binding sites, but my result does not show the same degree of functional importance as the ε subunit. Although the β_1 subunit does not bind to the ligand and not have a complementary part of the binding sites, this subunit may be more important for the structure of nAChR than the other subunits in the muscle system.

For the subunits in the nervous system, on the other hand, I could show that the α_7 subunit had the lowest *f* value. This is supported by the fact that the α_7 subunit is responsible for the crucial function to regulate the release of neurotransmitters. The actual functions of the other subunits in the nervous system are unclear. However, the α_3 - α_6 and β_2 - β_3 subunits may have crucial functions because they also have lower *f* values. As for the subunits in the nervous system, the coexpression of subunits has been reported (for reviews see Role and Berg 1996). For instance, the coexpression of the α_3 subunit along with the β_2 or β_4 subunit is known to produce a functional receptor *in vivo,* and a recent report showed that the α_5 subunit can also produce a functional receptor through coexpression with the α_4 and β_2 subunits (Ramirez-Latorre et al. 1996). The β_3 subunit, however, failed to produce a functional receptor with coexpression with any one of the other subunits in the nervous system. Thus, I showed that the *f* value could evaluate the degree of functional importance of the nAChR subunits and that the window analysis could predict the regions where important functions exist. These analyses are useful for giving an insight to further experimental studies for elucidating the actual functions of these subunits.

In the window analysis, subunits in the muscle system have a common sequence region (called the M2 region) which produces a hole in the ion channel (fig. 2.5). When subunits produce one receptor, these subunits may share a common region with strong functional constraints. Under this assumption, if I perform the window analysis for subunits in the nervous system when the nucleotide sequences of all the subunits involved become known, I may be able to predict the combination of subunits necessary to produce one nAChR.

Further studies will be needed to elucidate clearer relationships between function and evolution of all the subunits in both muscle and nervous systems.

2.5 Questions to be addressed

In the window analysis, the ratio of the number of nonsynonymous substitutions to that of synonymous substitutions was used for identification of particular regions where functional constraints are imposed on.

It is expected that since the synonymous substitution is free from functional constraints of a protein, the rate of synonymous substitution is expected to be constant within a given gene. In this analysis, however, I found that that the number of synonymous substitutions may vary within a subunit gene, as shown in figure 2.6. Indeed, previous studies reported that the synonymous substitution rates may not be uniform within a gene (Lawrence et al. 1991; Comeron and Aguadé 1996; Eyre-Walker and Bulmer 1993; Ina et al. 1994; Cacciò et al. 1995; Zoubak et al. 1995). However, several previous studies suggested that the number of synonymous substitutions was homogeneous within a gene (Clark and Kao 1991; Cacciò et al. 1995; Zoubak et al. 1995). Thus, the intragenic variation of synonymous substitution rates is somewhat controversial. This may be because in the previous studies, only specific genes were used or statistical reliability was lacked. In the following chapter, chapter III, I would like to examine whether synonymous substitution rates vary within a gene by using a larger amount of data set and by employing rigorous statistical methods.

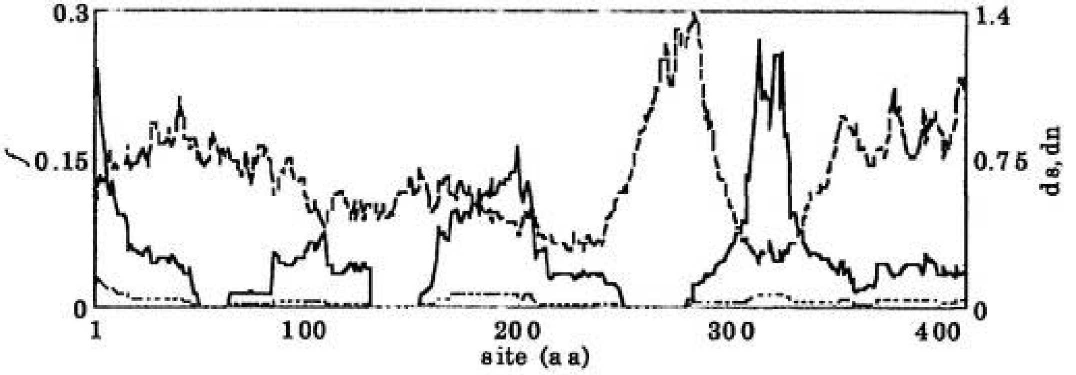


Figure 2.6: Intragenic variation in the number (ds) of synonymous substitutions, the number (dn) of nonsynonymous substitutions, and the ratio (*f* = dn/ds) for nAChR α_1 subunit. The ds, dn, and *f* values are shown with black dashed, gray dashed, and black lines, respectively.

CHAPTER III:

INTRAGENIC VARIATION OF SYNONYMOUS SUBSTITUTION RATES

3.1 Introduction

As synonymous nucleotide substitutions do not affect the primary structure of a protein, it has been commonly thought that functional constraints for synonymous changes were either very weak or non-existent (Kimura 1968; King and Jukes 1969). As a result, the rates of synonymous substitutions were expected to be fairly uniform among different genes as well as within a genes (Kimura 1983).

In reality, the intergenic variation of synonymous substitution rates has been observed for most organisms (Graur 1985; Li et al. 1985; Wolfe et al. 1989; Bernardi et al. 1993; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995). However, the intragenic variation has been somewhat controversial. Because the underlying spontaneous mutation rate is considered to be more constant within a gene than among genes, the rate of synonymous substitution is thought to be constant within a gene if synonymous substitutions are exempted from the functional constraints at the DNA or mRNA levels. However, several studies using the window analysis have suggested that the synonymous rate is not uniform within a gene.

For example, the glyceraldehyde-3-phosphate dehydrogenase and the outer membrane protein 3A in enteric bacteria displays intragenic variation of synonymous substitution rates (Lawrence et al. 1991). Synonymous rates within the Xdh gene of Drosophila are also shown not to be constant (Comeron and Aguadé 1996). In addition, Ina et al. (1994) suggests reduction of synonymous substitutions in the core protein of hepatitis C virus. Eyre-Walker and Bulmer (1993) used a sizable gene set of 138 homologous gene pairs between *E. coli* and *Salmonella typhimurium* and suggested that the rate of synonymous substitution is reduced near the start of genes, although they did not quantify this analysis statistically. Moreover, Cacciò et al. (1995) and Zoubak et al. (1995) used 69 homologous genes of four mammalian orders and they suggested that the synonymous substitution process is nonrandom. However, when they examined statistically homologous gene pairs of the same mammalian orders, they could not find significant nonrandomness of synonymous substitutions. Furthermore, the self-incompatibility locus in plants does not display intragenic variation (Clark and Kao 1991). Thus, these previous results were somewhat controversial for intragenic variation of synonymous substitution rates. This may be because the previous studies used only specific genes or lacked statistical reliability.

In order to solve this controversial issue, I examined whether there is intragenic variation of synonymous substitution rates by using a substantial number of gene pairs and by employing rigorous statistical methods. In fact, I compared 418 homologous gene pairs from *Rattus norvegicus* and *Mus musculus* as well as 84 orthologous gene pairs from the recently sequenced whole bacterial genomes of *Mycoplasma genitalium* and

Mycoplasma pneumoniae (Fraser et al. 1995; Himmelreich et al. 1996). These comparisons were made because a large number of genes could be used and the divergence between two species, for both rodents and *Mycoplasmas*, was not too large to confront the saturation effect of synonymous substitutions and not too small to suffer from shortage of substitutions. For each gene pair, I then estimated the proportion (Ps) of synonymous differences within a gene by the window analysis. The Ps values obtained were statistically examined to elucidate whether these values show the intragenic variation.

I found that there is a significant variation of synonymous substitution rates within a gene. Indeed, 92% of compared gene pairs between mouse and rat and 95% of the gene pairs between the two species of *Mycoplasma* showed the intragenic variation of the Ps values under the 5% level of significance. Therefore, synonymous substitution rates actually vary within genes of both mammals and bacteria.

I examined all conceivable possibilities that may cause the intragenic variation of synonymous substitutions. In particular, I examined whether the rate of synonymous substitution are correlated with that of nonsynonymous substitution, the degree of codon usage bias, the stem or loop regions of the mRNA structures, base content, and the frequency of CpG dinucleotides. Among these possibilities, I finally found a significant correlation between synonymous substitutions and the frequency of CpG dinucleotides in rodents. Since a methylated C at CpG dinucleotides is known to be a mutable site, my observation suggests that intragenic variation of synonymous substitutions is caused mainly by a nonrandom mutation due to the methylation of CpG dinucleotides.

3.2 Materials and Methods

3.2.1 Data extraction

The gene pairs between *R. norvegicus* and *M. musculus* were extracted from the SODHO database (Tateno et al. 1997) which was constructed with the DDBJ database release 30. These two species provide the largest number of homologous gene pairs currently available in the public nucleotide sequence databases. The gene pairs between *M. genitalium* and *M. pneumoniae* were selected from the complete genome sequence data of Himmelreich et al. (1996) where orthologous gene pairs have been determined. These two bacteria are most closely related phylogenetically among genomes completely sequenced so far. These comparisons were made for the following reasons. First, a large number of genes could be compared. Second, the divergence between two species, for both rodents and two species of *Mycoplasma*, was not too large to confront the saturation effect of synonymous substitutions. Third, the divergence is not too small that we may suffer from shortage of substitutions.

The following criteria were used to further select gene pairs. First, I extracted gene pairs sharing the same function to ensure that the pair was orthologous. Second, I eliminated from the analysis gene pairs whose gene lengths were less than twice the window size, in order to avoid statistical fluctuation due to the window size. Third, I used gene pairs which had no gaps in the pairwise alignment, guaranteeing that gene pairs were of the same length. I am fully aware that two serious problems on the window analysis can arise if I use pairwise alignments with gaps: (1) if I omit the gapped regions in a given gene sequence and then conduct the window analysis, some windows will contain consecutive regions that are artificially connected to each other and thereby have no biological meaning; (2) if one ignores gaps in the calculation for a window, then the estimated values will depend on the number of codons in the window. These problems are more serious when the window size is relatively small.

To conduct the window analysis, a window was set on the first codon of the pairwise alignment and shifted one codon at a time. This process was repeated by shifting the window codon by codon until it reached the last codon of the alignment. The window size was chosen to be 60 bases (20 codons) unless mentioned otherwise. I then estimated the proportion (Ps) of synonymous differences for each window. Use of the Ps value is sufficient as it is free from the "saturation" effect of synonymous substitutions. The modified Nei and Gojobori method (Zhang et al. 1998) was used for this estimation. In the present study, I did not use the original Nei and Gojobori (1986) method as it has been shown that it may underestimate the Ps value when there is a strong transition/transversion bias (Ina 1995). The other published estimation methods were not used as they sometimes return inapplicable values when closely related sequences are used (Zhang, et al. 1998). Furthermore, I eliminated gene pairs which had implausibly high Ps values (Ps >1) due to statistical fluctuation from the sampling errors.

In this way, I finally obtained 418 gene pairs from rat and mouse and 84 gene pairs from the two species of *Mycoplasma*.

3.2.2 Statistical test for intragenic variation of synonymous substitutions

In order to examine, with statistical validity, whether Ps values vary within the gene, I generated random nucleotide sequences reflecting codon usage of the gene pair and compared them with actual sequences. The statistical test for each gene pair was conducted as follows. First, I computed the frequency of each codon pair of the two gene sequences aligned. Second, using these frequencies, I generated random nucleotide sequences such that they have the same length as the window size and they reflect the codon usage of the actual pairwise alignment. Third, I generated 10,000 pairs of random sequences and estimated the Ps value for each pair of random sequences. Thus, a random distribution of 10,000 Ps values is obtained. Finally, I computed the probability of the Ps value observed for each window on the actual pairwise alignment by using the distribution of 10,000 Ps values.

3.2.3 Methods for examining the cause of intragenic variation of synonymous substitutions

I investigated the causes of intragenic variation of synonymous substitutions by examining possible correlations between Ps and other measures; the proportion (Pn) of nonsynonymous differences, the codon usage bias, the mRNA structures, the base composition, and a frequency of CpG dinucleotides. These were also calculated for each window. Whenever the correlation analysis was conducted, I used the windows which were not overlapped to each other, in order to ensure independence of calculated measures. For each window, I computed the average of the codon usage bias, base contents, and frequencies of CpG dinucleotides between a pair of genes.

Because the biological background is considerably different between rodents and *Mycoplasma*, I paid my attention of the correlation analysis only to rodents. First, I calculated Peason's correlation coefficient between Ps values and one of these measures for each gene pair. Because I selected only gene pairs having at least nine non-overlapping windows, the number of gene pairs that I could use was 316. The remaining 102 gene pairs (= 418 – 316) were not used for the correlation analysis. Second, I conducted the t-test for each correlation coefficient by setting the significance level at 5%. I then calculated the probability that the observed number of gene pairs having correlation coefficients with statistical significance was expected by chance with the binomial distribution. Third, I then used the reduced 0.0158% (= 5%/316) level of significance for each correlation coefficients in accordance with the Bonferroni method. In the Bonferroni method, the overall significance level divided by the number of comparisons is used as a significance level for each comparison. This is because the overall significance level may become larger than 5% if I use the 5% of significance for each correlation coefficient. If I find at least one significant correlation coefficient by this method, I can exclude a possibility that significance of the overall correlation between Ps and the measure examined took place by chance.

I calculated the GC1%, GC2%, and GC3% (the GC content at the 1st, 2nd, and 3rd positions of the codon, respectively) as base compositions. When I calculated these measures, I excluded codons having no synonymous codons because they do not contribute to synonymous substitutions. Such codons were Met and Trp in the comparison between mouse and rat.

I used ENC (an effective number of codons) as a measure of codon usage bias (Wright 1990). This measure quantifies how far the codon usage of a gene departs from equal usage of synonymous codons. Note that when the short length of windows is used, the biased value of ENC is likely to be obtained (Comeron and Aguadé 1998). Thus, for the comparison between the Ps and ENC values, I used 300 bp of the window length instead of a regular window size of 60 bp.

3.3 Results

3.3.1 A significant intragenic variation of synonymous substitution rates

Table 3.1 summarizes the results of the statistical test that was conducted to check significance of intragenic variation of the Ps values. Interestingly enough, these results indicate that almost all gene pairs examined have at least one intragenic region where synonymous variation is statistically significant. In this region, the Ps value is high or low at the 5% level of significance. In the comparison between mouse and rat, 92% of 418 gene pairs showed statistically significant variation of the Ps values within a gene. In the comparison of two species of *Mycoplasma*, 94% of 84 gene pairs showed the significant variation.

These results were not overly affected by a window size. When I used a window size of 540 bases (180 codons) instead of 60 bases, 54% of the gene pairs of rodents compared showed statistically significance in intragenic variation of the Ps values. Therefore, in spite of the relatively large window size of 540 bp, which is about half of the average gene length over all 418 gene pairs of rodents, more than half of the compared gene pairs still showed statistically significant variation of synonymous substitution rates within a gene.

Although a longer length gene was expected to show significant variation, I did not observe any notable correlation between the gene length and the number of intragenic regions where the Ps value is significantly high or low (data not shown).

Figures 3.1 (a)–(d) show the intragenic variation of Ps values for four gene pairs, as an example. The p value for each of Ps values is shown in these figures, demonstrating that I can identify the intragenic regions where the Ps value is significantly high or low.

Comparison		Total # of pair^{s}	# of sign pairs[1]
M. genitalium	vs *M. pneumoniae*	84	79(94%)
R. norvegicus	vs *M. musculus*	418	384(92%)

Table 3.1 The number of gene pairs which showed statistically significant variation of the Ps values within a gene. [1] I used the 5% level of significance for each comparison of gene pair.

Figure 3.1: Four examples for the intragenic variation of Ps values.

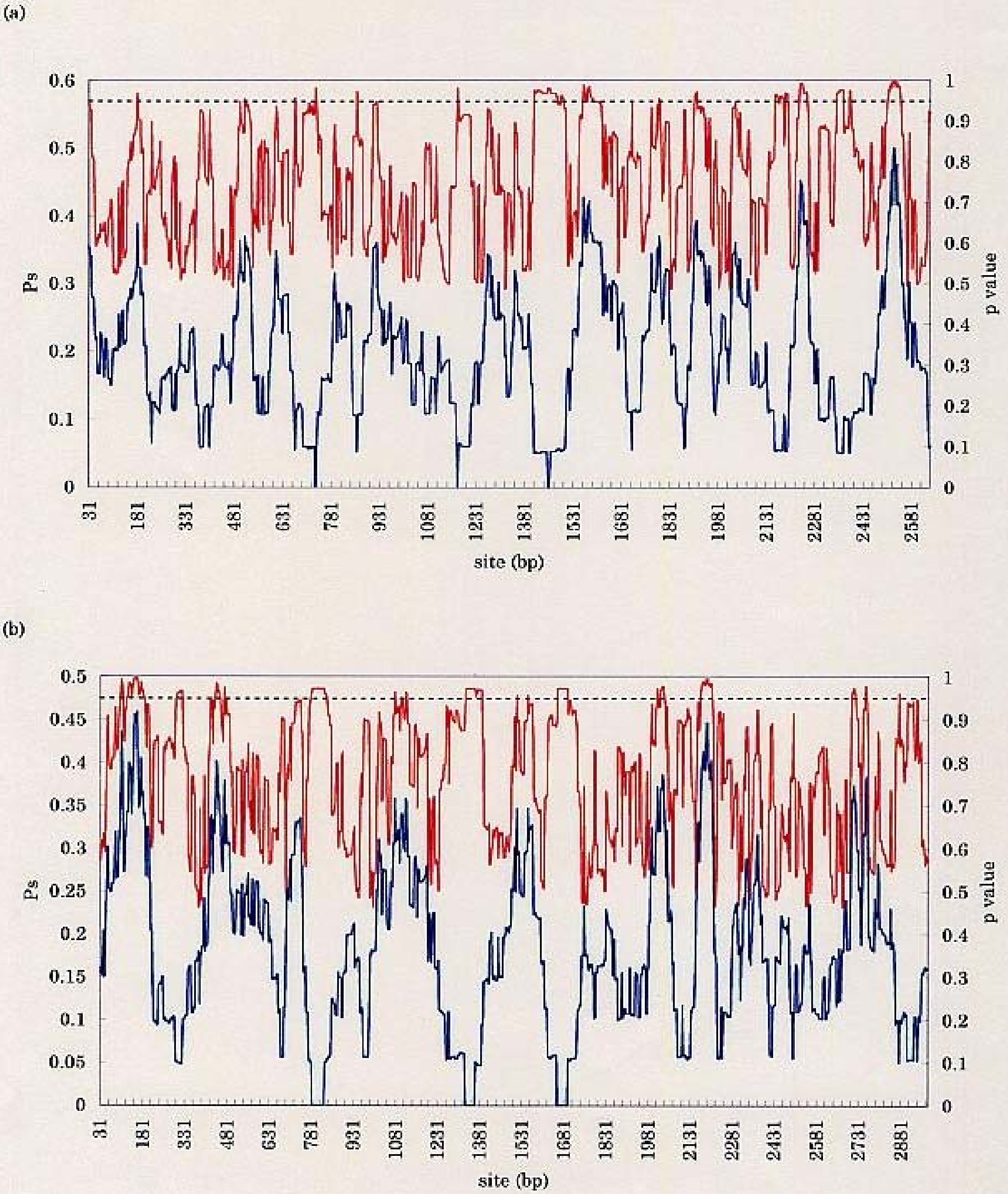
(a) iron responsive element binding factor of mouse and rat

(b) thiazide-sensitive sedium-chloride cotransporter of mouse and rat

(c) ATP-dependent protease in *Mycoplasmas*

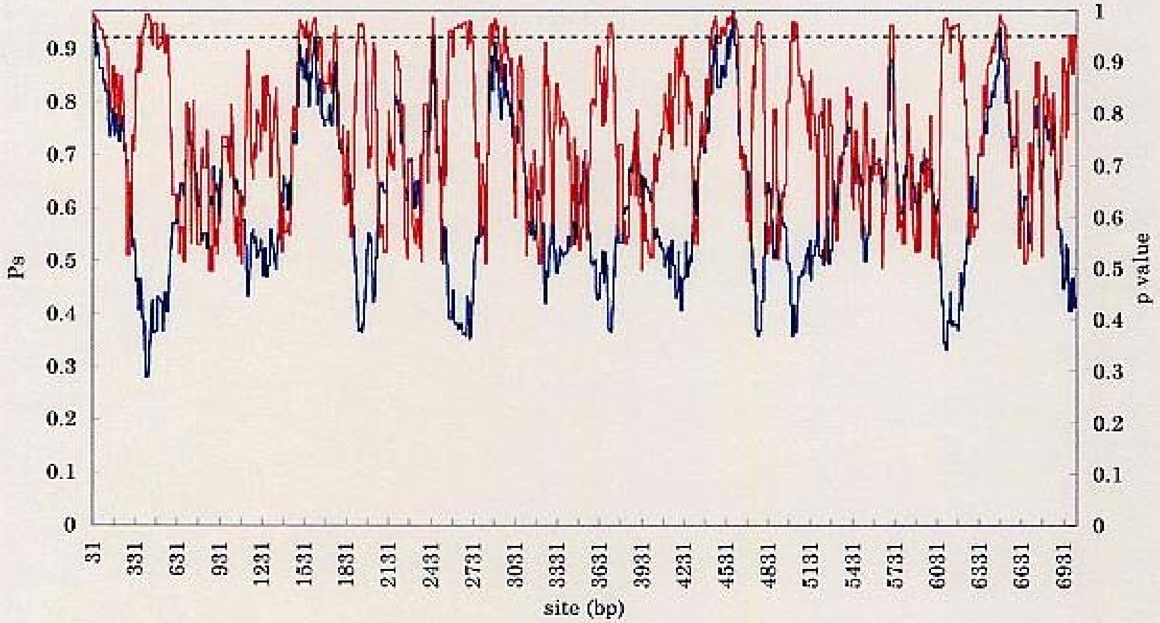
(d) ribonucleoside-diphosphate reductase in *Mycoplasmas*

red line: Ps values, blue line: the p values for each of Ps values


(a)

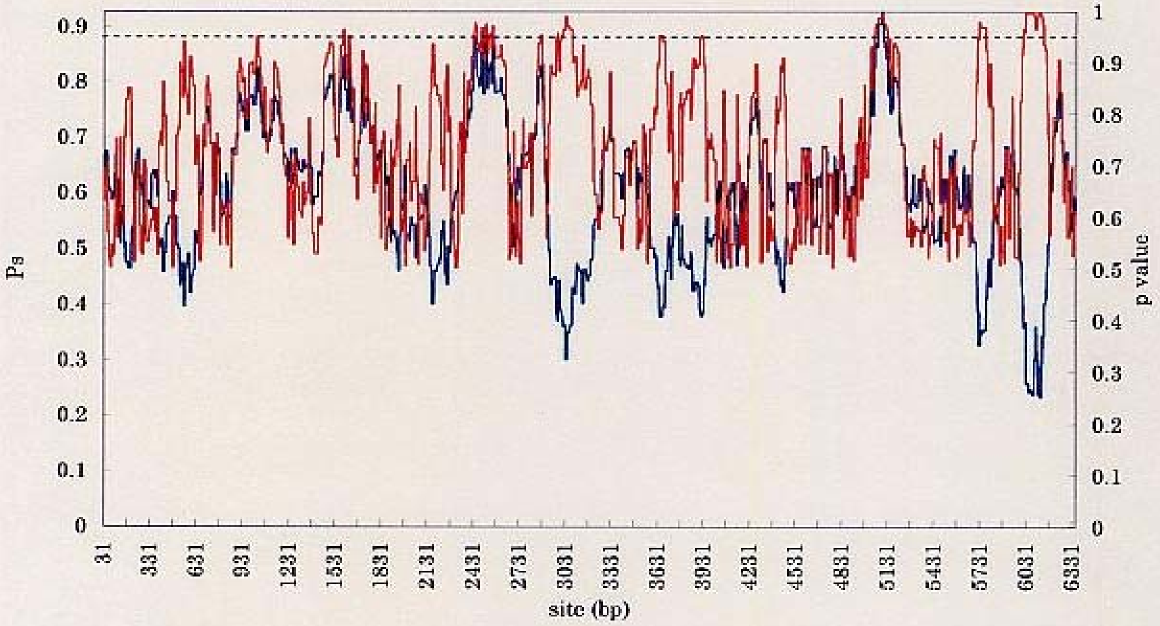
(b)

(c)



Ps
p value
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
31
331
631
931
1231
1531
1831
2131
2431
2731
3031
3331
3631
3931
4231
4531
4831
5131
5431
5731
6031
6331
6631
6931
site (bp)

(d)



Ps
p value
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
31
331
631
931
1231
1531
1831
2131
2431
2731
3031
3331
3631
3931
4231
4531
4831
5131
5431
5731
6031
6331
site (bp)

3.3.2 Possible causes of intragenic variation

Three possible explanations for intergenic variation have been proposed so far by previous studies. These explanations can be also applied for intragenic variation. Adding two new explanations, I examined a total of five possible causes of intragenic variation of synonymous substitutions.

First, functional constraints at the protein level, which typically affect nonsynonymous substitutions, could also operate on synonymous substitutions. Second, synonymous substitutions could be constrained by the codon usage bias. Third, the intragenic variation of synonymous substitutions may be due to the secondary structure of mRNA. Fourth, the intragenic variation may be caused by heterogeneity of base composition. Fifth and finally, there is a possibility that an underlying mutation rate varies within a gene. In the followings, these possibilities were examined gene by gene, by using gene pairs of rodents only, because the biological background is considerably different between rodents and *Mycoplasmas.*

(1) *Functional constraints at the protein level*

I examined the possibility that functional constraints against nonsynonymous substitutions work even on synonymous substitutions. This possibility was deduced from the observation that the gene having a low rate of synonymous substitution also manifests a low rate of nonsynonymous substitution. In fact, this has been observed in bacteria, Drosophila, and mammals (Graur 1985; Li et al. 1985; Wolfe et al. 1989; Bernardi et al. 1993; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995). In order to examine this possibility, I conducted a window analysis for computing the proportions (Pn) of nonsynonymous differences. Table 3.2 shows the number of gene pairs in which correlations between the Ps and Pn were found to be significant. As shown in table 3.2, 49 out of 316 gene pairs of rodents showed the significant correlation at the 5% level of significance. When I used a lower level of significance, that is, 0.0158% according to Bonferroni method, only one gene pair showed significance in a correlation between Ps and Pn. Thus, the intragenic variation of synonymous substitution rates may be, to some extent, caused by functional constraints of proteins. One such possible constraint may originate from translational efficiency that depends upon the occurrence frequency of rare codons, because amino acids encoded by the rare codons eventually affect protein function.

(2) *Bias of codon usages*

I examined codon usage bias, though it is also related to the above-mentioned possibility. Numerous studies of Drosophila genes have shown that the degree of codon usage bias is negatively correlated with the rate of synonymous substitution (Shields et al. 1988; Sharp and Li 1989; Moriyama and Gojobori 1992). In bacteria and yeast, the degree of codon usage bias is correlated with the level of gene expression, and the codon used most frequently in each synonymous codon family shows a clear relationship with tRNA-abundance (Ikemura 1981; 1982; Sharp and Li 1986). Moreover, it has been shown that genes having a strong bias of codon usage have evolved with a slower rate of synonymous substitution (Sharp and Li 1986; Powell and Moriyama 1997). It has also been suggested that selection for translation accuracy works on synonymous substitutions (Akashi 1994).

To test a possible relationship between the intragenic variation of synonymous substitution rates and the degree of codon usage bias, I calculated the ENC value by the window analysis. The ENC values quantify how far the codon usage of a gene departs from equal usage of synonymous codons (Wright 1990). Because I used 300 bp of the window length instead of a regular window size of 60 bp as described in materials and methods, the number of gene pairs compared reduced to 9. As shown in table 3.2, only 2 out of 9 gene pairs showed significance in a correlation between Ps and ENC. This number of gene pairs showing significant correlations was not enough to conclude statistical significance of the overall correlation between Ps and ENC. This is because the probability that the observed number of gene pairs having significant correlation coefficients was expected by chance was not less than the 5% of significance level. Indeed, when the lower level of significance is adopted, I could not find any gene pair showing significant correlation between Ps and ENC.

Two previous studies, which focused on the intragenic variation, also observed no correlation between the rate of synonymous substitutions and the degree of codon usage bias (Lawrence et al. 1991; Comeron and Aguadé 1996).

(3) The secondary structure of mRNAs.

The observation of no correlation between the intragenic variation of synonymous substitution rates and the codon usage bias would be understandable if selection was acting only on mRNA secondary structure (Eyre-Walker and Bulmer 1993). In other words, it suggests that functional constraints working on synonymous substitutions are at the mRNA level, not at the protein level. Indeed, several studies have reported that there is a relationship between mRNA secondary structure and synonymous substitutions in the genes of bacteria and hepatitis C virus (Lawrence et al. 1991; Comeron and Aguadé 1996; Smith and Simmonds 1997). I then investigated possible functional constraints for maintenance of mRNA secondary structure affecting synonymous substitutions. Unfortunately, only three gene pairs in my data set have descriptions of their mRNA sequences in the entries of the

EMBL/GenBank/DDBJ database. This is because although I could use the cDNA data for prediction of a mRNA secondary structure, the lack of 5'- and 3' untranslated regions affect the result of prediction. Thus, I used three gene pairs of the histone subunit 1, glycoprotein hormone alpha subunit, and regenerating protein I. However, only the histone subunit 1 among three gene pairs showed statistical significance in intragenic variation at the 5% level. Although the number of data is very limited, I estimated a mRNA secondary structure for each of histone subunit 1 genes of mouse and rat by the mfold software version 2.3 (Zuker 1989). As shown in figure 3.2, the intragenic region where Ps is significantly high, could be observed in both the stem and loop regions. On the other hand, the intragenic region where Ps is significantly low tended to be observed in the loop region of mRNA structures of both mouse and rat. However, the structural features corresponding specifically to these regions are quite different from each other. Thus, the possibility of functional constraints at the mRNA level is deniable, at present, as a cause of the intragenic variation of synonymous substitution rates.

(4) *Base composition*

I considered the possibility of functional constraints acting at the DNA level (Ticher and Graur 1989; Wolfe and Sharp 1993). It has been recently suggested that there are functional constraints which maintain a particular base composition (Alvarez-Valin et al. 1998). To test this possibility, I calculated the GC1%, GC2%, and GC3% and examined their correlations with Ps. The GC1%, GC2%, and GC3% are the GC content at the 1st, 2nd, and 3rd positions of the codon, respectively. As a result, for GC1%, GC2%, and GC3%, 61, 64, and 71 gene pairs out of 316 showed significant correlations with Ps at the 5% level of significance, respectively. These results are shown in table 3.2. Among these three measures, only GC3% showed overall significance of its correlation with Ps because 2 gene pairs showed significant correlations when I used a lower significance level.

(5) Spontaneous mutation rate

Finally, I investigated the remaining possibility that the intragenic variation of synonymous substitutions reflect heterogeneity of the mutation rate within a gene. Such nonrandomness of mutation has been known to be 'hotspots' of mutation. In order to examine this possibility for nonrandomness of mutation, I calculated the average frequency of CpG dinucleotides, because almost all regions of vertebrate genomes are subject to methylation and it is generally accepted that the methylcytosine, which is known as a mutable site, exists primarily in the CpG dinucleotide (for review, see Bird 1993; Holliday and Grigg 1993). As shown in table 3.2, 65 gene pairs out of 316 showed significant correlations between Ps and C1G2 (CG dinucleotides of the first and second codon positions), and 67 gene pairs showed significant correlation between Ps and C2G3 (CG dinucleotides of the second and third codon positions). Moreover, C3G1 (CG dinucleotides of the third and first codon positions spanning two codons) showed statistically significant correlations with the Ps values for 90 gene pairs. When I used a lower level of significance, only one gene pair showed a significant correlation between Ps and C1G2. Interestingly enough, a larger number of 5 gene pairs showed significant correlations between Ps and C2G3 at a lower level of significance. Moreover, for C3G1, a larger number of 4 gene pairs showed significant correlations with Ps at a lower level of significance. Thus, gene pairs having significant correlations with Ps were more frequently observed in the correlation analysis for C2G3 and C3G1 than in the other correlation analysis. These results lead me to the possibility that intragenic variation of synonymous substitutions reflects heterogeneity of the mutation rate within a gene.

Among the above-mentioned possibilities (1)–(5), when I calculated the probability that the observed number of significant correlation coefficients is expected by chance, all measures except ENC showed that correlations with the Ps are statistically significant at the 5% level. Thus, at this stage, a possibility for the degree of codon usage bias was rejected. Moreover, when I used a lower level of significance, GC1% and GC2% did not show significant correlation with Ps at all. Therefore, I eliminated the possibilities for base compositions at the 1st and 2nd positions of the codon.

On the other hand, the possibility for base compositions (GC3%) at the 3rd position of the codon remained because 2 gene pairs showed significant correlations between Ps and GC3%. Thus, including this possibility, the remaining possibilities were (1) functional constraints at the protein level, (2) base composition at the 3rd position of the codon, and (3) spontaneous mutation rate. These possibilities were examined by five measures of Pn, GC3%, C1G2, C2G3, and C3G1.

Then, I investigated the frequency distribution for the probabilities of correlation coefficients, as shown in figure 3.3. For each correlation analysis, I computed the probability for each of 316 correlation coefficients. As shown in figure 3.3, a probability lower than 1% was most frequently observed in the correlation analysis for all of these five measures. However, C3G1 showed the largest number of correlation coefficients having the probability lower than 1% when compared with the other measures. The second largest number of correlation coefficients having a probability lower than 1% was observed in the correlation analysis of C2G3. Therefore, I considered that the frequencies of C2G3 and C3G1 dinucleotides may be related to the cause of intragenic variation of synonymous substitutions.

Figure 3.4 shows one example of a gene pair of interleukin-1 receptor accessory protein. Out of 69 codon pairs where synonymous substitutions can be observed, 23 codons have dinucleotide C2G3 or C3G1 in either one of a codon pair. At these codon sites having synonymous changes, the most frequently observed substitution for C3G1 is a substitution from C to T at T3G1, whereas the one for C2G3 is a substitution from G to A, resulting in C2A3. This observation is consistent with mutation at a CpG dinucleotide producing a TpG and its complementary CpA dinucleotide.

My results, including the correlation analysis described above, always showed the strong correlation between synonymous substitution rates and frequencies of CpG dinucleotides. Since methylated CpG dinucleotide has been known as a mutable site in mammals, I finally concluded that at least in mammals, intragenic variation of synonymous substitutions is caused mainly by a nonrandom mutation due to the methylation of CpG dinucleotides.

	# of sig. pair[1] (5% level)	Prob.[2]	+/-[3]	# of sig. pair[1] (0.0158% level)	+/-[3]
Pn	50	2.74×10^{-12}	40/9	1	1/0
ENC	2	0.071	0/2	0	-/-
GC1%	61	1.39×10^{-19}	32/29	0	-/-
GC2%	64	1.31×10^{-21}	39/25	0	-/-
GC3%	71	1.26×10^{-26}	31/40	2	0/2
C1G2	65	2.66×10^{-22}	40/25	1	1/0
C2G3	67	1.04×10^{-23}	59/8	5	5/0
C3G1	90	4.17×10^{-42}	81/9	4	4/0

Table 3.2 Examination of possible causes of intragenic variation of synonymous substitutions. [1] Total number of gene pairs which show significant correlations with Ps. [2] The probability that the observed number of gene pairs having significant correlation coefficients was expected by chance. [3] The number of gene pairs which show significant positive and negative correlations with Ps, respectively.

Figure 3.2: Estimated mRNA secondary structures of histone subunit 1 genes of both mouse and rat. Light red and blue lines correspond to the region where significantly high and low Ps values were shown, respectively. Red and blue regions contain sites in which synonymous substitutions were observed.

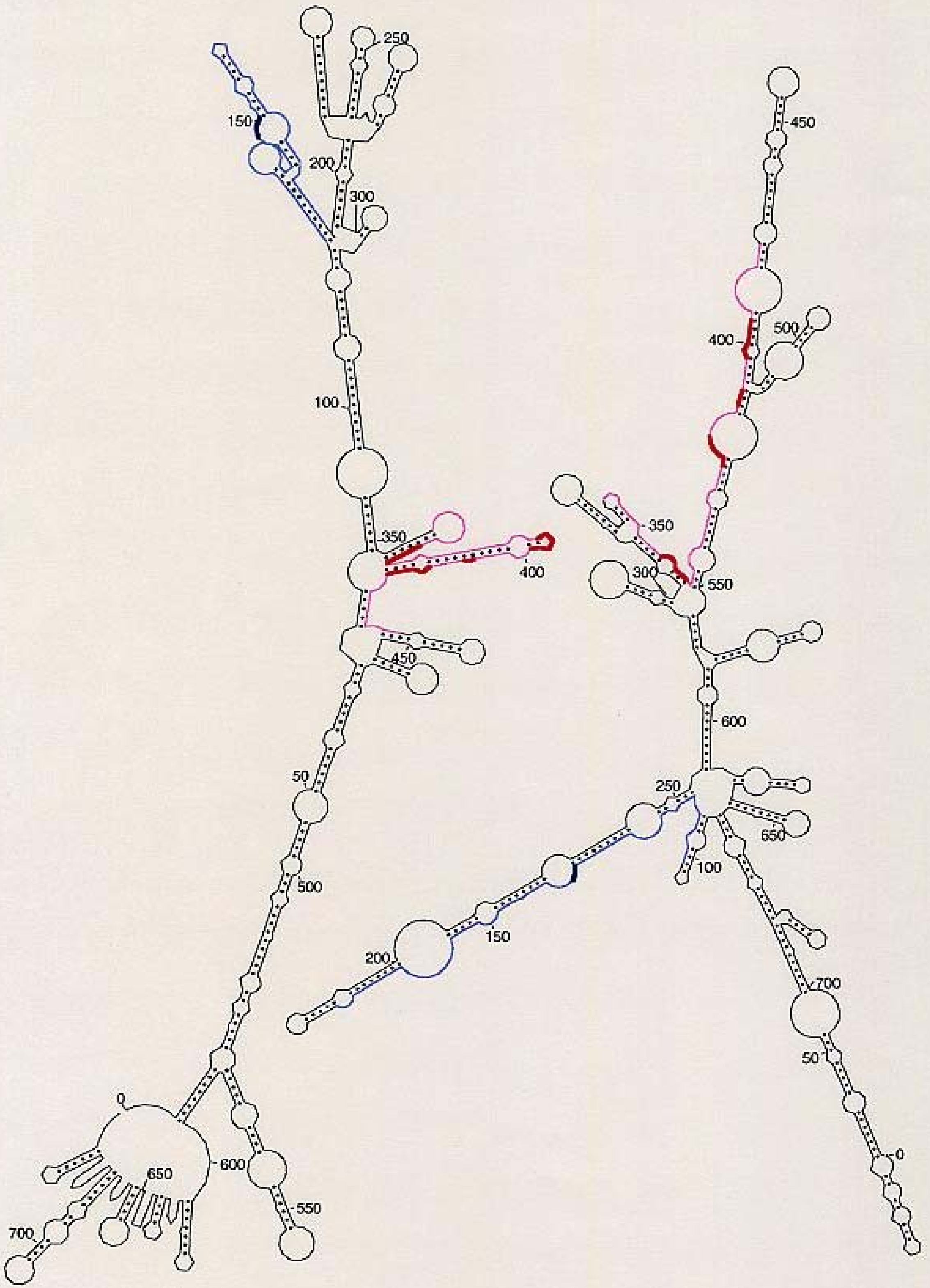


Figure 3.3: The frequency distribution for the p values of 316 correlation coefficients in each correlation analysis.

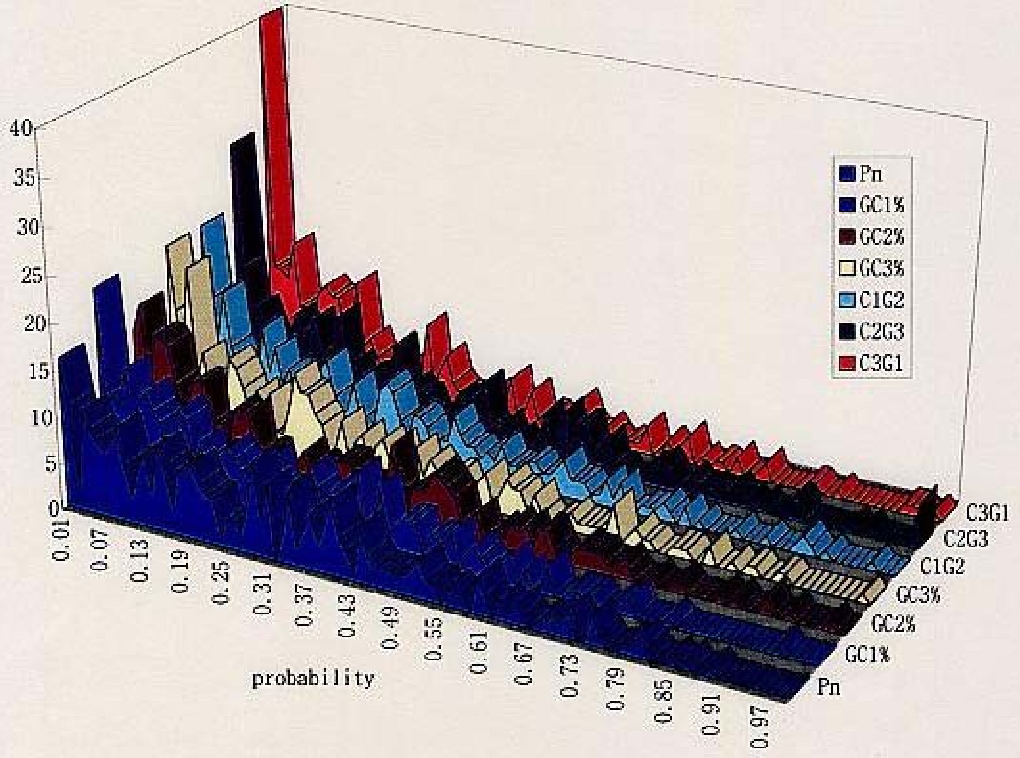


Figure 3.4: An example of a pairwise alignment of a interleukin-1 receptor accessory protein. The lines indicate in order: codon site number (bp), amino acid sequence for mouse, DNA sequence for mouse, DNA sequence for rat, and amino acid sequence for rat, respectively. Codon sites with synonymous and nonsynonymous changes are shown. The neighboring codons are also shown. Dots indicate identical codons. The number for codon sites (colored red) indicates codon sites with synonymous changes. Blue and green boxes show codon sites with synonymous changes at C3G1 and C2G3 in one of genes, respectively. Note that out of 69 codon pairs where synonymous substitutions can be observed, 13 codons have C3G1 and 10 codons have C2G3 in one of the species.

```
. 10   13   16   . 31   . 37   . 55   58   61   64   67   . 124  127  130  133  . 157  160  . 193  196  . 250  253  . 259
. L    W    Y    . S    . Y    . H    A    S    E    R    . P    A    R    I    . F    L    . G    L    . N    F    . L
. cTG  tgg  tAt  . tCc  . tAt  . caT  gct  tcG  gaG  cgc  . ccG  gct  cgA  atc  . ttc  Ctg  . ggC  ctt  . aaC  ttc  . ctC
. cCT  tgg  tGt  . tTc  . tGt  . caC  gct  tcC  gaA  cgc  . ccA  gct  cgC  atc  . ttc  Ttg  . ggA  ctt  . aaT  ttc  . ctT
. P    W    C    . F    . C    . H    A    S    E    R    . P    A    R    I    . F    L    . G    L    . N    F    . L

262  . 355  358  . 418  421  . 427  430  . 442  445  . 457  460  463  466  469  472  475  478  . 490  493  496  499  . 553  556
P    . T    T    . S    A    . R    F    . K    M    . H    G    I    H    K    I    T    C    . D    G    Y    F    . V    D
cca  . acA  act  . tcT  Gcc  . aga  Ttc  . aAg  Atg  . caT  ggC  att  caT  aaG  atc  acA  tgt  . gaC  gga  taC  ttt  . gtg  GaC
cca  . acG  act  . tcC  Ccc  . aga  Ctc  . aGg  Ctg  . caG  ggT  att  caC  aaT  atc  acG  tgt  . gaT  gga  taT  ttt  . gtg  AaT
P    . T    T    . S    P    . R    L    . R    L    . Q    G    I    H    N    I    T    C    . D    G    Y    F    . V    N

559  . 568  571  574  577  . 592  595  . 610  613  . 640  643  646  649  652  . 661  664  . 682  685  . 700  703  . 718  721
F    . V    L    P    E    . S    F    . V    S    . V    T    Y    P    E    . R    L    . T    V    . V    G    . A    L
ttt  . gtA  cTa  ccc  Gag  . agC  ttt  . gtT  tca  . gtT  acA  tat  cCT  gaa  . cgT  ctc  . act  Gtg  . gtG  ggc  . gca  TtG
ttt  . gtT  cAa  ccc  Aag  . agT  ttt  . gtA  tca  . gtG  acG  tat  cTG  gaa  . cgG  ctc  . act  Atg  . gtA  ggc  . gca  GtA
F    . V    Q    P    K    . S    F    . V    S    . V    T    Y    L    E    . R    L    . T    M    . V    G    . A    V

724  727  730  733  736  739  742  . 751  754  . 766  769  . 781  784  . 796  799  . 826  829  . 835  838  . 874  877  880  883
P    P    Q    I    Y    S    P    . R    V    . K    E    . E    L    . C    K    . S    H    . E    V    . V    T    V    D
cca  ccC  caG  atc  taT  tcT  cca  . cgT  gtt  . aaA  gaa  . gaA  Ctg  . tgC  aaa  . tcC  cac  . gag  Gtc  . gtc  AcA  gtC  gac
cca  ccT  caT  atc  taC  tcG  cca  . cgC  gtt  . aaG  gaa  . gaG  Ttg  . tgT  aaa  . tcT  cac  . gag  Atc  . gtc  CcT  gtT  gac
P    P    H    I    Y    S    P    . R    V    . K    E    . E    L    . C    K    . S    H    . E    I    . V    P    V    D

. 895  . 901  904  . 928  931  . 967  970  . 976  979  . 997  1000 . 1006 1009 1012 . 1021 1024 1027 1030 1033 . 1057 1060
. N    . S    V    . E    T    . P    E    . L    R    . H    A    . N    T    K    . A    E    Q    A    A    . P    P
. aAt  . agT  gta  . gaA  aca  . ccG  gag  . ctC  aGg  . caT  gct  . aat  Acc  Aaa  . gcT  gag  caG  gcT  gcc  . ccA  cca
. aTt  . agC  gta  . gaG  aca  . ccT  gag  . ctG  aAg  . caC  gct  . aat  Gcc  Gaa  . gcC  gag  caA  gcA  gcc  . ccG  cca
. I    . S    V    . E    T    . P    E    . L    K    . H    A    . N    A    E    . A    E    Q    A    A    . P    P

. 1099 1102 . 1111 1114 . 1123 1126 . 1135 1138 . 1168 1171 1174 1177 . 1183 1186 . 1201 1204 1207 1210 1213 . 1243 1246 .
. T    V    . V    V    . I    V    . H    V    . R    A    H    F    . T    D    . D    G    K    E    Y    . V    E    .
. acG  gtc  . gtA  gtg  . atT  gtg  . caT  gtt  . cgA  gct  caC  ttt  . acA  gat  . gaT  gga  aaG  gaG  tat  . gTG  gaa  .
. acA  gtc  . gtG  gtg  . atC  gtg  . caC  gtt  . cgG  gct  caT  ttt  . acG  gat  . gaC  gga  aaA  gaA  tat  . gCA  gaa  .
. T    V    . V    V    . I    V    . H    V    . R    A    H    F    . T    D    . D    G    K    E    Y    . A    E    .

1252 1255 1258 . 1270 1273 1276 . 1297 1300 . 1324 1327 1330 1333 1336 . 1405 1408 . 1522 1525 . 1552 1555 . 1564 1567 . 1582
E    E    F    . T    L    R    . F    G    . R    D    S    L    P    . S    P    . D    M    . K    T    . T    V    . G
gaG  gaA  ttt  . acG  ctG  cgt  . ttT  gga  . agA  gac  agc  CtG  cct  . agT  ccc  . gac  Atg  . aag  Acg  . acG  gtc  . ggA
gaA  gaG  ttt  . acA  ctA  cgt  . ttC  gga  . agG  gac  agc  TtC  cct  . agC  ccc  . gac  Ttg  . aag  Tcg  . acC  gtc  . ggG
E    E    F    . T    L    R    . F    G    . R    D    S    F    P    . S    P    . D    L    . K    S    . T    V    . G

1585 . 1621 1624 . 1669 .
E    . Q    L    . N    .
gag  . cag  Ttg  . aAt  .
gag  . cag  Ctg  . aGt  .
E    . Q    L    . S    .
```

3.4 Discussion

In this paper, I have shown that synonymous substitutions significantly vary within a gene, by using a substantial number of data set and employing an appropriate statistical test. Moreover, I found that intragenic variation of synonymous substitutions is mainly due to an nonrandom mutation within a gene.

Although I concluded that a nonrandom mutation due to the methylation of CpG dinucleotides was the main cause of intragenic variation of synonymous substitutions, it is also possible that functional constraints of the base composition cause intragenic variation of synonymous substitution rates. This is because 2 gene pairs showed significant correlations with Ps and GC3% when I used a lower significance level. However, I think that these correlations can be explained by a nonrandom mutation due to the methylated CpG dinucleotides in the following observations. I first observed that 2 gene pairs having significant correlations between Ps and GC3% always showed 'negative' correlations (Table 3.2). On the other hand, C1G2, C2G3, and C3G1 were always shown to have 'positive' correlations for the gene pairs having statistically significant correlations with Ps at a lower level (Table 3.2). Therefore, these opposite correlations of these measures with Ps lead to the possibility that synonymous substitutions can be frequently observed at the codon sites having CpG dinucleotides and, at the same time, GC3% is reduced at the codon sites. Indeed, I observed this possibility in the gene pair in figure 3.4. Among 13 C3G1 codon pairs having synonymous changes, the most frequently observed substitution is from C3G1 to T3G1. Moreover, the most frequently observed substitution is from C2G3 to C2A3 among 10 C2G3 codons pairs with synonymous changes. Thus, synonymous substitutions at the codon pairs having CpG dinucleotides make the cause of reduction of GC3%. Therefore, intragenic variation of synonymous substitutions is mainly caused by a nonrandom mutation due to the methylation of CpG dinucleotides rather than by functional constraints of the base composition. Since DNA methylation in vertebrates could control gene activities, one may infer that the mutation from a CpG dinucleotide to a TpG/CpA dinucleotide influences gene regulation by methylation. However, synonymous substitutions at CpG dinucleotides may not affect the gene regulation because other substitutions could supply CpG dinucleotides.

In this discussion, I focused my study on possible cause of intragenic variation of synonymous substitutions in mouse and rat. I did not observe similar correlations between synonymous substitution rates and CpG in the two *Mycoplasmas*. This may be because biological backgrounds are very much different between prokaryotes and eukaryotes. In particular, the two species of *Mycoplasma* have features distinctive from rodents in the following points. The divergence in GC% between the two species of *Mycoplasma* is quite different when compared with that of rodents. This difference leads me to the possible confrontation that it is difficult to explain the intragenic variation by the GC content. As mentioned earlier, the codon usage bias in bacteria directly affects the level of gene expression. This may imply that the codon usage bias within a gene of bacteria can cause the intragenic variation of synonymous substitutions more severely than that of mammals. Moreover, not only the methylation pattern but also the role of methylation in bacterial genomes may be quite different from vertebrate genomes having CpG islands.

However, my findings that for rodents the intragenic variation of synonymous substitutions may be caused by nonrandom mutation may also apply to *Mycoplasma*. At any rate, more detailed analysis is needed to identify the cause of the intragenic variation in the bacteria.

CHAPTER IV:

SUMMARY

A nucleotide substitution in the protein-coding gene sequences is classified into synonymous or nonsynonymous substitution. The synonymous substitution, which does not cause an amino acid change, is free from functional constraints of a protein whereas the nonsynonymous substitution, which does cause an amino acid change, is essentially constrained by protein function. Thus, the rate of synonymous substitution is expected to be constant within a given gene because the underlying spontaneous mutation rate is considered to be more constant within a gene than among genes. Moreover, it is expected that synonymous substitutions take place more frequently than nonsynonymous substitutions. Since the synonymous substitution is exempted from functional constraints of a protein whereas the nonsynonymous substitution is essentially constrained by protein function, the difference between the numbers of synonymous and nonsynonymous substitutions is thought to reflect the degree of functional importance for a protein.

In chapter I, I first described the history for estimation methods of synonymous and nonsynonymous substitution rates and the outline of the present thesis, placing particular emphasis on the motivation and purpose of my study.

In chapter II, I found the difference between the numbers of synonymous and nonsynonymous substitutions could be utilized for evaluating the functional importance for genes as well as intragenic regions with special reference to nicotinic acetylcholine receptor (nAChR) subunit genes. nAChR is composed of 16 types of subunits, which expressed in both nervous and muscle systems. There are five types of α_1, β_1, γ, δ, and ε subunits in the muscle system, whereas it is known that there are eight α-type (named $\alpha_2 - \alpha_9$) and three β-type (classified as $\beta_2 - \beta_4$) subunits in the nervous system. I first examined an evolutionary relationship among these subunits by constructing the phylogenetic tree.

By using 84 nucleotide sequences of receptor subunits from 18 different species, I showed that the common ancestor of all subunits may have appeared first in the nervous system. Moreover, I suggested that the α_1 subunits in the muscle system originated from the common ancestor of α_2, α_3, α_4, α_5, α_6, and β_3 in the nervous system, whereas the β_1, γ, δ, and ε subunits in the muscle system shared the common ancestor with the β_2 and β_4 subunits in the nervous system. Next, on the basis of the evolutionary relationship among these subunits, I examined the degree of functional importance for these subunit genes as well as intragenic regions of a subunit. Calculation of the ratio (f) of the number of nonsynonymous substitutions to that of synonymous substitutions suggested that very strong functional constraints work on the α_1 subunit among 5 types of subunits in the muscle system and the α_7 subunits among 11 types of the nervous subunits.

These findings are consistent with the fact that these subunits have crucial functions for the receptor; the α_1 subunit has binding sites to the ligand and the α_7-containing receptor regulates the release of the transmitter. Moreover, by applying calculation of the *f* value to the intragenic regions, I found that strong functional constraints work on binding sites in the α_1 subunit and the so-called M2 region, which constructs a hole in the ion channel, in all 5 types of the muscle subunits. Therefore, I concluded that comparison between the number of synonymous and nonsynonymous substitutions can be useful for evaluation of functional importance of subunits as well as intragenic regions of a subunit.

In chapter III, I successfully showed that the rate of synonymous substitution is variable not only among genes but also within a gene, by using a substantial number of data set and by employing rigorous statistical methods. To avoid the saturation effect of synonymous substitutions and to use a large number of gene pairs, we used 418 homologous gene pairs from *Rattus norvegicus* and *Mus musculus* as well as 84 orthologous gene pairs from the whole bacterial genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae* . I found that 92% of gene pairs of rodents showed the significant variation of synonymous substitution rates within a gene. Moreover, 94% of gene pairs of *Mycoplasmas* showed the significant variation. Therefore, synonymous substitution rates actually vary within genes for both of mammals and bacteria. Furthermore, in this chapter, I examined all conceivable possibilities that may cause the intragenic variation of synonymous substitutions. In particular, I examined whether the rate of synonymous substitution are correlated with that of nonsynonymous substitution, the degree of codon usage bias, mRNA secondary structures, base content, and the frequency of CpG dinucleotides. I finally found a significant correlation between synonymous substitutions and the frequency of CpG dinucleotides in rodents. Since a methylated C at CpG dinucleotides is known to be a mutable site, our observation suggests that intragenic variation of synonymous substitutions is caused mainly by a nonrandom mutation due to the methylation of CpG dinucleotides.

As the future development of this line of study, I would think that more data of comparable sequences, the accumulation of which can be expected particularly by the genome projects of various organisms, can be used for the studies of evaluating of functional importance of the intragenic regions, along with the elucidation of more detailed molecular mechanisms of intragenic variation of synonymous substitutions. These studies will become powerful tools for predicting and identifying a particular region of having important function within a gene. Moreover, these studies will give deep insight into the evolutionary process of functional differentiation of genes.

References

Ajuh, P. M. and T. G. Egwang. 1994. Cloning of a cDNA encoding a putative nicotinic acetylcholine receptor subunit of the human filarial parasite Onchocerca volvulus. Gene **144**:127-129.

Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics **136**:927-935.

Alvarez-Valin, F., K. Jabbari, and G. Bernardi. 1998. Synonymous and nonsynonymous substitutions in mammalian genes: Intragenic correlations. J. Mol. Evol. **46**:37-44.

Anand, R. and J. Lindstrom. 1990. Nucleotide sequence of the human nicotinic acetylcholine receptor beta 2 subunit gene. Nucleic Acids Res. **18**:4272.

Baldwin, T. J., C. M. Yoshihara, K. Blackmer, C. R. Kintner, and S. J. Burden. 1988. Regulation of acetylcholine receptor transcript expression during development in Xenopus laevis. J. Cell Biol. **106**:469-478.

Ballivet, M., J. Patrick, J. S. Lee, and S. Heinemann. 1982. Molecular cloning of cDNA coding for the gamma subunit of Torpedo acetylcholine receptor. Proc. Natl. Acad. Sci. U.S.A. **79**: 4466-4470.

Beeson, D., M. Brydson, M. Betty, S. Jeremiah, S. Povey, A. Vincent, and J. Newsom-Davis. 1993. Primary structure of the human muscle acetylcholine receptor. cDNA cloning of the gamma and epsilon subunits. Eur. J. Biochem. **215**:229-238.

Beeson, D., M. Brydson, and J. Newsom-Davis. 1989. Nucleotide sequence of human muscle acetylcholine receptor beta-subunit. Nucleic Acids Res. **17**:4391.

Bernardi, G., D. Mouchiroud, and C. Gautier. 1993. Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Evol. **37**:583-389.

Bird, A. P. 1993. Functions for DNA methylation in vertebrates. Cold Spring Harbor Symp. Quant. Biol. **58**:281-285.

Bossy, B., M. Ballivet, and P. Spierer. 1988. Conservation of neural nicotinic acetylcholine receptors from Drosophila to vertebrate central nervous systems. EMBO J. **7**:611-618.

Boulter, J., A. O'Shea-Greenfield, R. M. Duvoisin, J. G Connolly, W. Wada, A. Jensen, P. D. Gardner, M. Ballivet, E. S. Deneris, D. McKinnon, S. Heinemann, and J. Patrick. 1990. Alpha-3, alpha-5, and beta-4: three members of the rat neuronal nicotinic acetylocholine receptor-related gene family form a gene cluster. J. Biol. Chem. **265**:4472-4482.

Boulter, J., J. Connolly, E. S. Deneris, D. J. Goldman, S. F. Heinemann, and J. Patrick. 1987. Functional expression of two neuronal nicotinic acetylcholine receptors from cDNA clones identifies a gene family. Proc. Natl. Acad. Sci. U.S.A. **84**:7763-7767.

Boulter, J., K. Evans, D. Goldman, G. Martin, D. Treco, S. Heinemann, and J. Patrick. 1986. Isolation of a cDNA clone coding for a possible neural nicotinic acetylcholine receptor alpha-subunit. Nature **319**:368-374.

Buonanno, A., J. Mudd, V. Shah, and J. P. Merlie. 1986. A universal oligonucleotide probe for acetylcholine receptor genes: Selection and sequencing of cDNA clones for the mouse muscle beta subunit. J. Biol. Chem. **261**:16451-16458.

Cacciò, S., S. Zoubak., G. D'Onofrio, and G. Bernardi. 1995. Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. J. Mol. Evol. **40**:280-292.

Cauley, E. A., B. W. Agranoff, and D. Goldman. 1990. Multiple nicotinic acetylcholine receptor genes are expressed in goldfish retina and tectum. J. Neurosci. **10**:670-683.

Cauley, K., B. W. Agranoff, and D. Goldman. 1989. Identification of a novel nicotinic acetylcholine receptor structural subunit expressed in goldfish retina. J. Cell Biol. **108**:637-645.

Changeux, J. P. 1990. Functional architecture and Dynamics of the nicotinic acetylcholine receptor: An allosteric ligand-gated ion channel. Raven Press Ltd., New York, 21-168.

Changeux, J. P., A. Bessis, J. P. Bourgeois, P. P. Corringer, A. Devillers-Thiery, J. L. EiselÉ, M. Kerszberg, C. LÉna, N. L. NovÈre, M. Poicciotto, and M. Zori. 1996. Nicotinic receptors and brian plasticity. Cold Spring Harbor Symp Quant Biol **61**:343-362.

Chini, B., F. Clementi, N. Hukovic, and E. Sher. 1992. Neuronal-type alpha-bungarotoxin receptors and the alpha 5-nicotinic receptor subunit gene are expressed in neuronal and nonneuronal human cell lines. Proc. Natl. Acad. Sci. U.S.A. **89**:1572-1576.

Clark, A. G. and TH. Kao. 1991. Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc. Natl. Acad. Sci. USA **88**:9823-9827.

Claudio, T., M. Ballivet, J. Patrick, and S. Heinemann. 1983. Nucleotide and deduced amino acid sequences of Torpedo californica acetylcholine receptor gamma subunit. Proc. Natl. Acad. Sci. U.S.A. **80**:1111-1115.

Comeron, J. M. and M. Aguadé. 1996. Synonymous substitutions in the *Xdh* gene of Drosophila: Heterogeneous distribution along the coding region. Genetics **144**:1053-1062.

Comeron, J. M. and M. Aguadé. 1998. An evaluation of measures of synonymous codon usage bias. J. Mol. Evol. **47**:268-274.

Couturier, S., D. Bertrand, J. M. Matter, M. C. Hernandez, S. Bertrand, N. Mille, S. Valera, T. Barkas, and V. Ballivet. 1990. A neuronal nicotinic acetylcholine receptor subunit (α7) is developmentally regulated and forms a homo-oligomeric channel blocked by α-BTX. Neuron **5**:847-856.

Couturier, S., L. Erkman, S. Valera, D. Rungger, S. Bertrand, J. Boulter, M. Ballivet, and D. Bertrand. 1990. Alpha 5, alpha 3, and non-alpha 3. Three clustered avian genes encoding neuronal nicotinic acetylcholine receptor-related subunits. J. Biol. Chem. **265**:17560-17567.

Criado, M., V. Witzemann, M. Koenen, and B. Sakmann. 1988. Nucleotide sequence of the rat muscle acetylcholine receptor epsilon-subunit. Nucleic Acids Res. **16**:10920.

Deneris, E. S., J. Boulter, L. W. Swanson, J. Patrick, and S. Heinemann. 1989. β3: a new member of nicotinic acetylcholine receptor gene family is expressed in brain. J. Biol. Chem. **264**:6268-6272.

Devillers-Thiery, A., J. Giraudat, M. Bentaboulet, A. Klarsfeld, and J. P. Changeux. 1984. Molecular genetics of Torpedo marmorata acetylcholine receptor. Adv. Exp. Med. Biol. **181**:17-29.

Devillers-Thiery, A., J. Giraudat, M. Bentaboulet, and J. P. Changeux. 1983. Complete mRNA coding sequence of the acetylcholine binding alpha-subunit of Torpedo marmorata acetylcholine receptor: a model for the transmembrane organization of the polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. **80**:2067-2071.

Elgoyhen, A. B., D. S. Johonson, J. Boulter, D. E. Vetter, and S. Heinemann. 1994. α9: An acetylcholine receptor with novel pharmacological properties expressed in rat cochlear hair cells Cell **79**:705-715.

Elliott, K. J., S. B. Ellis, K. J. Berckhan, A. Urrutia, L. E. Chavez-Noriega, E. C. Johnson, G. Velicelebi, and M. M. Harpold. 1996. Comparative structure of human neuronal alpha 2-alpha 7 and beta 2-beta 4 nicotinic acetylcholine receptor subunits and functional expression of the alpha 2, alpha 3, alpha 4, alpha 7, beta 2, and beta 4 subunits. J. Mol. Neurosci. **7**:217-228.

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685-690.

Eyre-Walker, A. 1996. The close proximity of Escherichia coli genes: Consequences for stop codon and synonymous codon use. J. Mol. Evol. **42**:73-78.

Eyre-Walker, A. and M. Bulmer. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res.

2 1:4599-4603.

Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J-F. Tomb, B. A. Dougherty, K. F. Bott, P-C. Hu, T. S. Lucier, S. N. Perterson, H. O. Smith, C. A. Hutchison III, and J. C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science **2 7 0**:397-403.

Galzi, J. L., and J. P. Changeux. 1994. Ligand-gated channel as unconventional allosteric proteins. Curr. Opin. Struct. Biol. **4**:554-565.

Garcia-Guzman, M., F. Sala, S. Sala, A. Campos-Caro, W. Stuhmer, L. M. Gutierrez, and M. Criado. 1995. Alpha-bungarotoxin-sensitive nicotinic receptors on bovine chromaffin cells: molecular cloning, functional expression and alternative splicing of the alpha 7. Eur. J. Neurosci. **7**:647-655.

Gardner, P. D. 1990. Nucleotide sequence of the epsilon-subunit of the mouse muscle nicotinic acetylcholine receptor. Nucleic Acids Res. **1 8**:6714.

Gojobori, T., W-H. Li, and D. Graur. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol.

18:360-369.

Goldman, D. J., E. S. Deneris, W. Luyten, A. Kochhar, J. Patrick, and S. F. Heinemann. 1987. Members of a nicotinic acetylcholine receptor gene family are expressed in different regions of the mammalian central nervous systems. Cell **48**:965-973.

Graur, D. 1985. Amino acid composition and the evolutionary rates of protein-coding genes. J. Mol. Evol. **22**:53-62.

Gudelfinger, E. D. 1995. Evolution and desensitization of LGIC receptors. Trends Neurosci **18**:297.

Hernandez, M. C., L. Erkman, L. Matter-Sadzinski, T. Roztocil, M. Ballivet, and J. M. Matter. 1995. Characterization of the nicotinic acetylcholine receptor beta 3 gene. Its regulation within the avian nervous system is effected by a promoter 143 base pairs in length. J. Biol. Chem. **270**:3224-3233.

Hieber, V., J. Bouchey, B. W. Agranoff, and D. Goldman. 1990. Nucleotide and deduced amino acid sequence of the goldfish neural nicotinic acetylcholine receptor beta-2 subunit. Nucleic Acids Res. **18**:5307.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B-C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. **24**:4420-4449.

Hoekstra, R., A. Visser, L. J. Wiley, A. S. Weiss, N. C. Sangster, and M. H. Roos. 1997. Characterization of an acetylcholine receptor gene of haemonchus contortus in relation to levamisole resistance. Mol. Biochem. Parasitol. **84**:179-187.

Holiday, R. and G. W. Grigg. 1993. DNA methylation and mutation. Mutation Research **285**:61-67.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codons choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389-409.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40**:190-226.

Ina, Y., M. Mizokami, K. Ohba, and T. Gojobori. 1994. Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. J. Mol. Evol. **38**:50-56.

Isenberg, K. E., J. Mudd, V. Shah, and J. P. Merlie. 1986. Nucleotide sequence of the mouse muscle nicotinic acetylcholine receptor alpha subunit. Nucleic Acids Res. **14**:5111.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. In H. N. Munro ed. Mammalian Protein Metabolism. Academic Press, NewYork. 21-132.

Karlin, A. 1993. Structure of nicotinic acetylcholine receptors. Curr. Opin. Neurobiol. **3**:299-309.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature **217**:624-626.

Kimura, M. 1983. The neutral theory of molecular evolution. Camb. Univ. Press.

King, J. L. and T. H. Jukes. 1969. Non-Darwinian evolution. Science **164**:788-798.

Kliman, R. M. and J. Hey. 1994. The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics **137**:1049-1056.

Kubo, S., M. Noda, T. Takai, T. Tanabe, T. Kayano, S. Shimizu, K. Tanaka, H. Takahashi, T. Hirose, S. Inayama, R. Kikuno, T. Miyata, and S. Numa. 1985. Primary structure of delta subunit precursor of calf muscle acetylcholine receptor deduced from cDNA sequence. Eur. J. Biochem. **149**:5-13.

Kullberg, R. W., Y. C. Zheng, W. Todt, J. L. Owens, S. E. Fraser, and G. Mandel. 1994. Structure and Expression of the nicotinic acetylcholine receptor beta subunit of Xenopus laevis. Recept. Channels **2**:23-31.

Lawrence, J. G., D. L. Hartl, and H. Ochman. 1991. Molecular considerations in the evolution of bacterial genes. J. Mol. Evol. **33**:241-250.

Li, W-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**:96-99.

Li, W-H., C-I. Wu, and C-C, Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150-174.

Luther, M. A., R. Schoepfer, P. Whiting, B. Casey, Y. Blatt, M. S. Montal, M. Montal, and J. Linstrom. 1989. A muscle acetylcholine receptor is expressed in the human cerebellar medulloblastoma cell line TE671. J. Neurosci. **9**:1082-1096.

Maricq, A. V., A. S. Peterson, A. J. Brake, R. M. Meyers, and D. Julius. 1991. Primary structure and functional expression of the 5HT3 receptor, a serotonin-gated ion channel. Science **254**:432-437.

Marshall, J., S. D. Buckingham, R. Shingai, G. G. Lunt, M. W. Goosey, M. G. Darlison, D. B. Sattelle, and E. A. Barnard. 1990. Sequence and functional expression of a single alpha subunit of an insect nicotinic acetylcholine receptor. EMBO J. **9**:4391-4398.

McGehee, D. S., M. J. S. Heath, S. Geiber, P. Devay, and L. W. Role. 1995. Nicotine enhancement of fast excitatory synaptic transmission in CNS by presynaptic receptors. Science **269**:1692-1696.

Mihovilovic, M., and A. D. Roses. 1991. Expression of mRNAs in human thymus coding for the alpha 3 subunit of a neuronal acetylcholine receptor. Exp. Neurol. **111**:175-180.

Miyata, T. and T. Yasunaga. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16**:23-26.

Monteggia, L. M., M. Gopalakrishnan, E. Touma, K. Idler, N. Nash, S. P. Arneric, J. P. Sullivan, and T. Giordano. 1995. The cloning and transient expression of the human alpha 4 and beta 2 neuronal nicotinie acetylcholine receptor (nAChR) subunits. Gene **155**:189-193.

Moriyama, E. N. and J. R. Powell. 1997. Synonymous substitution rates in *Drosophila* : mitochondrial versus nuclear genes. J. Mol. Evol. **45**:378-391.

Moriyama, E. N. and T. Gojobori. 1992. Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics **130**:855-864.

Mouchiroud, D. and C. Gautier. 1990. Codon usage changes and sequence dissimilarity between Human and Rat. J. Mol. Evol. 31:81-91.

Mouchiroud, D., C. Gautier, and G. Bernardi. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. 40:107-113.

Murray, N., Y. Zheng, G. Mandel, P. Brehm, R. Bolinger, Q. Reuer, and R. Kullberg. 1995. A single site on the epsilon subunit is responsible for the change in ACh receptor channel conductance during skeletal muscle development. Neuron **14**:865-870.

Nef, P., A. Mauron, R. Stalder, C. Alliod, and M. Ballivet. 1984. Structure, linkage, and sequence of the two genes encoding the delta and gamma subunits of the nicotinic acetylcholine receptor. Proc. Natl. Acad. Sci. U.S.A. **81**:7975-7979.

Nef, P., C. Oneyser, C. Alliod, S. Couturier, and M. Ballivet. 1988. Genes expressed in the brain define three distinct neuronal nicotinic acetylcholine receptors. EMBO J. **7**:595-601.

Nei, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

Nei, M. and T. Gojobori. 1986. Simple methods for estimating the number of synonymous and nonsynonymous substitutions. Mol. Biol. Evol. **3**:418-426.

Noda, M., H. Takahashi, T. Tanabe, M. Toyosato, Y. Furutani, T. Hirose, M. Asai, S. Inayama, T. Miyata, and S. Numa. 1982. Primary structure of alpha-subunit precursor of Torpedo californica acteylcholine receptor deduced from cDNA sequence. Nature **299**:793-797.

Noda, M., Y. Furutani, H. Takahashi, M. Toyosato, T. Tanabe, H. Shimizu, S. Kikyotani, T. Kayano, T. Hirose, S. Inayama, and S. Numa. 1983.

Cloning and sequence analysis of calf cDNA and human genomic DNA encoding alpha-subunit precursor of muscle acetylcholine receptor. Nature **305**:818-823.

Novére, N. L., and J. P. Changeux. 1995. Molecular evolution of the nicotinic acetylchopline receptor: an example of multigene family in excitable cells. J. Mol. Evol. **40**:155-172.

Numa, S., M. Noda, H. Takahashi, T. Tanabe, M. Toyosato, Y. Furutani, and S. Kikyotani. 1983. Molecular structure of the nicotinic acetylcholine receptor. Cold Spring Harb. Symp. Quant. Biol. **48**:57-69.

Ohno, K., D. O. Hutchinson, M. Milone, J. M. Brengman, C. Bouzat, S. M. Sine, and A. G. Engel. 1995. Congenital myasthenic syndrome caused by prolonged acetylcholine receptor channel openings due to a mutation in the M2 domain of the ε subunit Proc. Natl. Acad. Sci. USA **92**:758-762.

Ohta T. and Y. Ina. 1995. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J. Mol. Evol. 41:717-720.

Orr-Urtreger, A., M. F. Seldin, A. Baldini, and A. L. Beaudet. 1995. Cloning and mapping of the mouse alpha 7-neuronal nicotinicacetylcholine receptor. Genomics **26**:399-402.

Ortells, M. O., and G. G. Lunt. 1995. Evolutionary history of the ligand-gated ion-channel superfamily of receptors. Trends Neurosci. **18**:121-127.

Pamilo, P. and N. O. Bicanchi. 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. Mol. Biol. Evol. **10**:271-281.

Peng, X., M. Katz, V. Gerzanich, R. Anand, and J. Lindstrom. 1994. Human alpha 7 acetylcholine receptor: cloning of the alpha 7 subunit from the SH-SY5Y cell line and determination of pharmacological properties of native receptors and functional alpha 7 homomers expressed in Xenopus oocytes. Mol. Pharmacol. **45**:546-554.

Powell, J. R. and E. N. Moriyama. 1997. Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. USA **94**:7784-7790.

Ramierz-Latorre, J., C. Yu, X. Qu, F. Perin, A. Karlin, and L. Role. 1996. Functional contributions of α5 subunit to neuronal acetylcholine receptor channels. Nature **380**:347-351.

Role, L. W., and D. K. Berg. 1996. Nicotinic receptors in the development and modulation of CNS synapses. Neuron, **16**:1077-1085.

Saitou, N., and M. Nei. 1987. Neighbor-joining: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406-425.

Sawruk, E., C. Udri, H. Betz, and B. Schmitt. 1990. SBD, a novel structural subunit of the Drosophila nicotinic acetylcholine receptor, shares its genomic localization with two alpha-subunits. FEBS Lett. **273**:177-181.

Sawruk, E., P. Schloss, H. Betz, and B. Schmitt. 1990. Heterogeneity of Drosophila nicotinic acetylcholine receptors: SAD, a novel developmentally regulated alpha-subunit. EMBO J. **9**:2671-2677.

Schoepfer, R., P. Whiting, F. Esch, R. Blacher, S. Shimasaki, and J. Lindstrom. 1988 cDNA clones coding for the structural subunit of a chicken brain nicotinic acetylcholine receptor. Neuron **1**:241-248.

Schoepfer, R., W. G. Conroy, P. Whiting, M. Gore, and J. Lindstrom. 1990. Brain alpha-bungarotoxin binding protein cDNAs and MAbs reveal subtypes of this branch of the ligand-gated ion channel gene superfamily. Neuron **5**:35-48.

Sharp, P. M. and W-H. Li. 1986. Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. Nucleic Acids Res. **14**:7737-7749.

Sharp, P. M. and W-H. Li. 1989. On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28**:398-402.

Shibahara, S., T. Kubo, H. J. Perski, H. Takahashi, M. Noda, and S. Numa. 1985. Cloning and sequence analysis of human genomic DNA encoding gamma subunit precursor of muscle acetylcholine receptor. Eur. J. Biochem. **146**:15-22.

Shields, D. C., P. M. Sharp, D. G. Higgins, and F. Wright. 1988. "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Mol. Biol. Evol. **5**:704-716.

Sivilotti, L., and D. Colquhoun. 1995. Acetylcholine Receptors: too many channels, too few functions. Science **269**:1681-1682.

Smith, D. B. and P. Simmonds. 1997. Characteristics of nucleotide substitution in the hepatitis C virus genome: Constraints on sequence change in coding regions at both ends of the genome. J. Mol. Eovl. 45:238-246.

Squire, M. D., C. Tornoe, Baylis, H. A., Fleming, J. T., Barnard, E. A. and D. B. Sattelle. 1995. Molecular cloning and functional co-expression of a Caenorhabditis elegans nicotinic acetylcholine receptor subunit (acr-2). Recept. Channels **3**:107-115.

Stephan, W. and D. A. Kirby. 1993. RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics 135:97-103.

Sumikawa, K., M. Houghto, Smith, J. C., Bell, L., Richards, B. M. and E. A. Barnard. 1982. The molecular cloning and characterization of cDNA coding for the alpha subunit of the acetylcholine receptor. Nucleic Acids Res. **1 0**:5809-5822.

Takai, T., M. Noda, M. Mishina, S. Shimizu, Y. Furutani, T. Kayano, T. Ikeda, T. Kubo, H. Takahashi, T. Takahashi, M. Kuno, and S. Numa.

1985. Cloning, sequencing and expression of cDNA for a novel subunit of acetylcholine receptor from calf muscle. Nature **315**:761-764.

Takai, T., M. Noda, Y. Furutani, H. Takahashi, M. Notake, S. Shimizu, T. Kayano, T. Tanabe, K. I. Tanaka, T. Hirose, S. Inayama, and S. Numa. 1984. Primary structure of gamma subunit precursor of calf-muscle acetylcholine receptor deduced from the cDNA sequence. Eur. J. Biochem. **143**:109-115.

Tanabe, T., M. Noda, Y. Furutani, T. Takai, H. Takahashi, K. Tanaka, T. Hirose, S. Inayama, and S. Numa. 1984. Primary structure of beta subunit precursor of calf muscle acetylcholine receptor deduced from cDNA sequence. Eur. J. Biochem. **144**:11-17.

Tarroni, P., F. Rubboli, B. Chini, R. Zwart, M. Oortgiesen, E. Sher, and F. Clementi. 1992. Neuronal-type nicotinic receptors in human neuroblastoma and small-cell lung carcinoma cell lines. FEBS Lett. **312**:66-70.

Tateno, Y., K. Ikeo, T. Imanishi, H. Watanabe, T. Endo, Y. Yamaguchi, Y. Suzuki, K. Takahashi, K. Tsunoyama, M. Kawai, Y. Kawanishi, K. Naitou, T. Gojobori. 1997. Evolutionary motif and its biological and structual significance. J. Mol. Evol. **44(Suppl 1)**:S38-S43.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**:4673-4680.

Ticher, A. and D. Graur. 1989. Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. J. Mol. Evol. 28:286-298.

Wada, K., M. Ballivet, J. Boulter, J. Connolly, E. Wada, E.S. Deneris, L. W. Swanson, S. Heinemann, and J. Patrick. 1988. Functional expression of a new pharmacological subtype of brain nicotinic acetylcholine receptor. Science **240**:330-334.

Wadsworth, S. C., L. S. Rosenthal, K. L. Kammermeyer, M. B. Potter, and D. J. Nelson.1988. Expression of a Drosophila melanogaster acetylcholine receptor- related gene in the central nervous system. Mol. Cell. Biol. **8**:778-785.

Willoughby, J. J., N. N. Ninkina, M. M. Beech, D. S. Latchman, and J. N. Wood. 1993. Molecular cloning of a human neuronal nicotinic acetylcholine receptor beta 3-like subunit. Neurosci. Lett. **155**:136-139.

Witzemann, V., E. Stein, B. Barg, T. Konno, M. Koenen, W. Kues, M. Criado, M. Hofmann, and B. Sakmann. 1990. Primary structure and functional expression of the alpha-, beta-, gamma-, delta- and epsilon-subunits of the acetylcholine receptor from rat muscle. Eur.

J. Biochem. **194**:437-448.

Wolfe, K. H. and P. M. Sharp. 1993. Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37:441-456.

Wolfe, K. H., P. M. Sharp, and WH. Li. 1989. Mutation rates differ among regions of the mammalian genome. Nature 337:283-285.

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

Yu, L., R. J. LaPolla, and N. Davidson. 1986. Mouse muscle nicotinic acetylcholine receptor gamma subunit: cDNA sequence and gene expression. Nucleic Acids Res. **14**:3539-3555.

Zhang, J., H. F. Rosenberg, M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**:3708-3713.

Zoubak, S., G. D'Onofrio, S. Cacciò, G. Bernardi, and G. Bernardi. 1995. Specific compositional patterns of synonymous positions in homologous mammalian genes. J. Mol. Evol. 40:293-307.

Zuker, M. 1989. On finding all suboptimal foldings of an rna molecule. Science **224**:48-52.