Chapter 1

Introduction And Review of Literature

CHAPTER 1

INTRODUCTION AND REVIEW OF LITERATURE

1.1. Introduction

Throughout the Beagle expedition (FitzRoy, 1839) naturalist Charles Darwin gathered evidence documenting the evolutionary history of numerous organisms across diverse regions of the world. The diverse range of species and the process of adaptation shaped the novel idea of evolution by natural selection with respect to the geographical, geological, and ecological prospectives (Darwin, 1927; Darwin et al., 2009). Since biologists in the mid-1800s and early 1900s were still searching for the ultimate answer for the factors responsible for the phenotypes of organisms, Darwin's theory of natural selection provided a strong backbone for traditional core of evolutionary biology (Gadgil and Bossert, 1970; Darwin et al., 2009; Gibbs and Grant, 1987). Ronald Fisher combined the mendelian particulate inheritance (Fisher, 1919; Berry and Browne, 2022) with the concept of Darwin's natural selection and popularised Neo-darwinism (Dawkins, 1986; Esposito, 2016). Since the early 1990s, evolutionary biologists were also keen in understanding the complicated concepts such as allele frequencies in population genetics (Hardy, 1908; Stern, 1943), speciation due to geographical barrier (Mayr, 1963; Chandler, 1914), microevolution and macroevolution (Simpson, 1953; Stanley, 1975), fossil evidence study for finding homologous and analogous organs (Boyden, 1947; Zangerl, 1948), genetic drift or genetic hitchhiking (Wright, 1942; Goodhart, 1963). However, along with the discovery of the double helical structure of the DNA (Watson and Crick, 1953), the clarity on the understanding of molecular genetics took a new turn. Since, the sequences of the four nucleotides (A, T, C, G) were known to be the key factor behind the phenotypes in organisms, several molecular biologists shifted their interested into a new era of molecular evolution.

Molecular evolution came into limelight during the 1960s in the western world of science. When the conventional evolutionary biologists were least interested in this science, the world was moving faster towards innovative technologies like genetic engineering, then molecular biologists embarked on evolutionary biology from the prospects of variations in the genome sequences. A comprehensive grasp of any biological process necessitates a dual understanding of both its molecular as well as its evolutionary dimensions. The conventional understanding of evolutionary biology treated it independently before the rise of molecular biology. But gradually the gap between the duo decreased and molecular evolution took its own realm in the field of modern biology. Due to the availability of genome sequences of different organisms, evolutionary biologists were interested in constructing phylogenetic based approaches in studying the evolutionary prospects of different organisms. In the early 1960, molecular evolution pioneers such as Linus Pauling and Emile Zuckerkandl worked on hemoglobin protein sequences and constructed vertebrate phylogeny and also hypothesized that different globin genes within a single organisms could also be traced to a common ancestral protein (Zuckerkandl et al., 1960), famously known as molecular clock or evolutionary clock (Zuckerkandl and Pauling, 1962; Morgan, 1998; Li, 1997). Subsequently, Carl Woese in 1970s had extensively worked on the famous "*three domain classification*", in which he used the phylogenetic approach to differentiate between bacteria and archaea (Woese and Fox,1977). Subsequently, the biologists across the globe worked on the phylogeny of different organisms to understand molecular evolution. The availability of ample number of genome sequences of the same species in various databases has enabled modern day researchers to study molecular evolution research in a different dimension as "intra-species study". While Giorgio Bernardi's contributions such as isochores and GC content had provided valuable insights into the structural and compositional aspects of the DNA, contributing to our foundational

understanding of genome evolution across different biological domains (Bernardi, 1989; Bernardi, 2019).

Since all population contain wild types as the most common form of alleles, they also do contain variations in them, which are commonly known as variants or alternatives of the most common types (Griffiths et al., 2000). Variants are the consequences of genetic mutations. Among different types of genetic mutations or chromosomal aberrations, single nucleotide variations (SNVs) are the most common form of base variations observed in all population forms (Shendure and Aiden, 2012). Mutations are recognized as the primary instigator of variations in the DNA base sequence, playing a pivotal role not only in the evolutionary process (Hershberg, 2015) but also in the development of complications such as cancer (Prehn, 2005). Nevertheless, mutation alone is inconsiderable, the selection of mutations is an essential step, particularly concerning the fitness of organisms, making it a pivotal driving force in the process of molecular evolution. (Sukhodolets, 1986). Most of the variations are unnoticeable to us, as the variations first undergo the process of selection. The variations having an enhancing in fitness of the organism are usually selected and others are purged out of the population. Between the two types of selections (positive and negative), positive selection refers to the variant in a population providing higher fitness and reproductive success than the individuals carrying the non-variants (Gregory, 2009). Negative selection refers to the selective removal of the harmful or deleterious genetic variants from the population (Gulisija and Crow, 2007). However, the concept of mutation-selection balance explains about the introduction of new deleterious mutations and purging out of the harmful mutations through purifying section (Sarkar, 1992; Nachman and Michael, 2004). But such fundamental understanding of mutation and selection did not contemplate to the neutral theory by Motto Kimura (Kimura, 1979). As Kimura explained that majority of the variations in the populations are nearly neutral, meaning they do not confer to the fitness of the organism to a large extent, hence the variations are due

to genetic drift in smaller populations where chance events play a pivotal role in shaping the fate of the variants (Kimura, 1979; Wright, 1984; Hurst, 2009).

The journey of gene expression starts from a non-reactive DNA and ends as a functional polypeptide chain which carries out major functions in all life forms, famously known as the central dogma in molecular biology (Crick, 1953; Watson and Crick, 1958; Crick, 1970). While the language of the DNA is translated into polypeptides through the gateway of genetic code table which is known as the most complex molecular process in the cell. The central dogma of molecular biology has itself the biology of gene expression, mathematics of mutation and chemistry of protein folding. With few exceptions in the extranuclear DNA in eukaryotes, all life forms follow the standard genetic code table (Fig. 1.1) (Nirenberg, 1963). Protein synthesis is the basis of life and amino acids are the building blocks of proteins, hence the assignment of amino acids to certain codons shows the importance of amino acids ever since prebiotic earth. The 61 sense codons and their assignment among 20 amino acids is a puzzling phenomenon, popularly known as the redundancy or degeneracy of genetic codon (Crick, 1988). Therefore, in a standard genetic code table, the degree of degeneracy ranges from a minimum of zero to a maximum of six. Certain amino acids are encoded by only one codon, while others can be encoded by up to six codons. For an example, Phe is coded by UUU and UUC, hence both UUU and UUC are known as synonymous codons. Few synonymous codons in coding sequences are used more frequently than the others. This inequal usage of synonymous codons are known as codon usage bias (CUB), which is known as a species-specific phenomenon (Ikemura, 1985). There are many factors such as GC content, tRNA abundance, gene expression, transcriptional selection, RNA stability, optimal growth temperature, and size of the coding sequence (Gouy and Gautier, 1982; McInerney, 1998; Lynn et al., 2002; Paul et al., 2008; Kober, and Pogson, 2013; Seward and Kelly, 2016) that affect the CUB pattern in an organism. Researchers in the past have rigorously worked on CUB, that enabled us to

understand the pattern of CUB even in highly expressed genes (HEG) and low expressed genes (LEG) in *E. coli* (Sen et al., 2022). CUB can only be studied by base substitutions rather than insertions and deletions (indels). Transition (*ti*) and transversion (*tv*) are two types of point mutations/base substitutions studied in molecular biology. Depending upon the stereochemistry nucleotides are divided into two classes; purine (R) and pyrimidine (Y), The intra-class substitution between nucleotides is known as *ti* (R→R, Y→Y). The inter-class substitution between nucleotides is known as *tv* (R→Y, Y→R) (Collins and Jukes, 1994; Sen et al., 2022; Beura et al., 2023) (Fig. 1.2). It is already known that the two types of point mutations never occur at equal frequencies despite *tv* having more paths than the *ti*. In *E. coli* neutral regions, *ti* has been observed to be around four times more frequent than *tv* (Seplyarskiy et al., 2012, Sen et al., 2022). This rate difference between *ti* and *tv* is crucial to understand DNA Sequence evolution. Base substitutions leading to amino acid changes or retention of the original amino acid had become trending research in the field of molecular evolution. However, the assignment of codons to a certain amino acid is still a complicated phenomenon than it was thought before. Nevertheless, concepts such as codon reassignment challenge the conventional understanding of the genetic code and can have significant implications for protein synthesis and function (Osawa and Jukes, 1989).

It is popularly believed that the sequence of the DNA destines for the folding pattern of the protein. Usually, non-synonymous substitutions were thought to have a more deleterious impact on the fitness of the organism than the synonymous substitutions (Eyre-Walker, 2007; Schmidt et al., 2008). But non-synonymous substitutions also do provide an enhanced fitness level while benefitting the organism with a successful amino acid exchangeability (Rodrigue et al., 2010). In such cases the advantageous mutations play a pivotal role in demystifying the forces of selection (Yang and Neilsen, 1998). However, many recent works have shown a deviation in protein folding with the introduction of a synonymous mutation (Shabalina et al.,

2013; Liu et al., 2021). If a gene also codes for the pattern of protein folding, is there a feedback loop working here in terms of purifying selection/translational selection? We do not know; it might have provided enough hints to evolve today's genetic code table as it is. Even though these questions make sense, but the mystery of codon assignment and related amino acids are yet unreachable for researchers.

	U		C		A		G		
U	UUU	F	UCU	S	UAU	Y	UGU	C	U
	UUC		UCC		UAC		UGC		C
	UUA	L	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	W	G
C	CUU		CCU	P	CAU	H	CGU	R	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Q	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	I	ACU	T	AAU	N	AGU	S	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	K	AGA	R	A
	AUG	M	ACG		AAG		AGG		G
G	GUU	V	GCU	A	GAU	D	GGU	G	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	E	GGA		A
	GUG		GCG		GAG		GGG		G

Fig. 1.1. A standard genetic code table illustrating 64 codons, including three stop codons and 61 sense codons. The assigned amino acid to different codons can also be visualised here. The genetic code table has 8 split boxes and 8 family boxes out of the total 16 boxes. Here, different degeneracy of amino acids can be illustrated. As an example, UUU and UUC code for Phe (F), whereas GUN (U, C, A, G) code for only valine. With a few exceptions in the start codon, *E. coli* follows the standard genetic code table as it is presented here.

1.2. Objectives of the thesis

(A) Estimation of $\frac{ti}{tv}$ ratio by accounting degeneracy and pretermination nature of codons

(i) To develop the improved equation for the estimation of synonymous $\frac{ti}{tv}$ and non-synonymous $\frac{ti}{tv}$ ratio

(ii) To understand the impact of codon composition on improved and conventional $\frac{ti}{tv}$ ratio

(iii) To develop the workflow for development of software for $\frac{ti}{tv}$ estimation

(iv) To understand the impact of purifying selection on coding sequences

(B) Analysis of non-synonymous variation in genome sequences of *Escherichia coli*

(i) To normalize non-synonymous transition (*Nti*) and non-synonymous transversion (*Ntv*) values by considering codon degeneracy and pretermination nature of the codons used in Chapter 2 of the thesis.

(ii) To estimate the $\frac{Nti}{Ntv}$ ratio across all the codons by considering the rate difference

(iii) To find out the most frequent amino acid changes in the dataset used for *Escherichia coli*

(C) A comparative polymorphism spectra analysis in inter-operon IGRs and intra-operon IGRs

(i) To understand the impact of replication and transcription on base substitution in different intergenic regions (IGRs)

(ii) To study the comparative polymorphism spectra between intra operon IGRs and inter operon IGRs

(iii) To study the amino/keto & purine/pyrimidine bias between intra operon IGRs and inter operon IGRs

(D) A comparative synonymous polymorphism spectra analysis in co-transcribed gene pairs

(i) To understand the impact of replication and transcription on co-transcribed genes

(ii) To compare the mutational spectra between co-transcribed genes

(iii) To understand the impact of codon degeneracy on mutation frequency in co-transcribed genes

1.3. Review of literature

Evolution is conventionally known as an inter-species phenomenon. Since the early days researchers in the field were interested in studying the evolutionary history of different organisms and finding their common ancestors through different phylogenetic methods like maximum likelihood, maximum parsimony, and Bayesian phylogenetics (Revell et al., 2012). Among such methods, few are intuitive and depends on the prior data to create the posterior data while generating the trees. While there has been several advantages and disadvantages of these methods, the introduction of algorithms or simulations like Markov chain Monte Carlo (MCMC) brought significant improvements in Bayesian method (Brooks et al., 1998). Such phylogenetic methods also intermingled among taxa while studying modern taxonomy (Heath et al., 2008). The inter-species study implies the appearance as well as elimination of several characters during the event of speciation. As mutation is actually known as the driving force for evolution (Hershberg, 2015), though many mutations and/or selection events remain unnoticed by the nature. Subsequently, due to the emergence of information technology in the field of biological science resulted into the digital availability of ample amount of genetic

information in databases such as DDBJ and NCBI (NCBI Resource Coordinators, 2016; Mashima et al., 2017). This has resulted into the availability of high throughput digital data of certain species in several servers. Model organisms like *Escherichia coli* is used since early days due to several advantages such as rapid and high reproductivity rate and for its simpler arrangement of the genome. Now a days the availability of several strains of an organism made it easy for the intra-species analyses in terms of their evolutionary point of view. The intra-species analyses became more informative in recovering the evolutionary journey among strains. The analysis of phylogeny, or evolutionary relationships among strains within a species, has become a pivotal tool in understanding the intricate evolutionary journey and genetic diversity within populations. The wild type alleles, considered the most prevalent original variants in a population, stand in contrast to rare variants, which can be regarded as newly introduced mutations within the population.

Chargaff's first parity rule played a significant role in elucidating the proposed double helical model of DNA by Watson and Crick (Elson and Chargaff, 1952; Forsdyke and Mortimer, 2000). Unlike the first parity rule, the second parity rule comes with violations. It has also been shown that the second parity rule applies to double stranded sequences in which the pattern of substitutions and selections are similar (Lobry, 1995; Sueoka, 1995; Lobry & Lobry, 1999). The local deviations have been proposed to be observed due to replication and transcription pressure (Francino and Ochman, 1997). During cellular events like DNA replication, the leading strand (LeS) and the lagging strand (LaS) are synthesized in different mechanisms (Okazaki et al., 1968; Kornberg, 1984). The LeS is exposed more as a single strand than the LaS (Powdel et al., 2009). It is already known that cytosine deamination in a single stranded DNA occurs in a half-life of ~200 years (Lindahl and Nyberg, 1974). Due to the single-stranded exposure of the DNA strands during the replication process at the replication fork, it results in a strand dependent mutation pattern in both the strands, which is commonly

known as the asymmetric directional mutation pressure (Lobry 1996, Lobry and Sueoka, 2002). Researchers have identified cytosine deamination as a primary contributor to base substitutions in genomic DNA, coupled with guanine oxidation that results in G→T substitutions (Kino and Sugiyama, 2001; Rocha et al., 2006; Bhagwat et al., 2016). Such a similar scenario is observed during the process of transcription. The transcription bubble progresses through the separated strands of the DNA, as the template strand is being engaged in synthesis of nascent mRNA, the non-template or the gene strand is actually exposed as a single strand. It also leads to strand asymmetry mutational patterns or a cytosine deamination (C→T/G→A) (Francino and Ochman, 1997; Mugal et al., 2009). Hence C→T/G→A is a major contributor to the polymorphism spectra in coding as well as non-coding regions in the chromosomes. As the mutation is already known to be AT biased (Hershberg and Petrov, 2010), the dynamic range of GC content became a subject of debate at a point of time in different bacterial species. There exists a variable range of GC% across the genomes of different prokaryotes, for example the minimum GC% is recorded in *Zinderia insecticola* with around 13% and the maximum is recorded in *Aneromyxobacter dehalogenans* with around 75% (Zhao et al., 2007). The GC% of prokaryotes, such as the E. coli genome, is known to be 50.7% (Araújo et al., 2021). With half of its genetic sequence already balanced in terms of composition, *E. coli* emerges as an ideal organism for studying mutation and selection biases in the field. There have been different schools of thoughts regarding genome GC% variation in the prokaryotic genome. The evolutionary biology fraternity has witnessed selectionist as well as mutationists views in the past (Singer and Ames, 1970; Muto and Osawa, 1987; Wu et al., 2012). By far, most selectionist views have been rejected by researchers whereas the mutationist view has been widely accepted. The rate of base substitution from G/C to A/T or A/T to G/C can provide better interpretation of GC content in an organism, known as GC mutational pressure (Sueoka, 1964; Muto and Osawa, 1987). Most of the prokaryotic genomes are usually composed of

protein coding genes, Thus GC mutational pressure as well as GC content are also affected by the selective constraint. The weaker the selective constraint (neutral regions), the stronger the impact of GC mutational pressure on GC composition (Brocchieri, 2014).

The structural similarity between the nucleotide bases makes it possible for the selection or retention of *ti* over *tv* during the proof-reading stage right after the DNA replication. Moreover, the pairing of R: R or Y: Y is usually not allowed due to its ability in structural distortion of the DNA backbone (Wahl and Sundaralingam, 1997). It is already known that the two types of point mutations never occur at equal frequencies despite *tv* having more paths than *ti*. In *E. coli* neutral regions *ti* has been observed to be around four times more frequent than *tv* (Seplyarskiy et al., 2012, Sen et al., 2022). To quantify this dilemma of *ti* bias, several substitution models have been proposed in the past. The very first substitution model was proposed by Jukes-Cantor (JC) in 1969 (Jukes and Cantor, 1969). JC proposed a Markov-chain model with a continuous time process. It supports the equal probability of all four nucleotides to undergo point substitutions and an equal rate of substitutions between any pair of nucleotides. Further, the JC model was revised by Felsenstein in 1981 (often called as F81) and it suggested that the probability of substitution of any nucleotide by another is proportional to the substituting nucleotide (Felsenstein., 1981). Gradually, the famous Kimura two parameter model clarified the inequal probability between *ti* and *tv* in genomes, famously known as K80 model (Kimura, 1980) (K80). The K80 model clearly suggested the difference in rates of substitutions between *ti* and *tv*. The HKY (Hasegawa-Kishino-Yano) model was introduced in 1985, and it suggested about the differences in the frequencies of substitution of four nucleotides also it suggested about the rate difference between *ti* and *tv* (Hasegawa et al., 1985). Similarly, the GTR (General Time Reversible) model suggested about the most general neutral, independent, finite-sites, time-reversible model possible (Tavaré., 1986). TN93 model (Tamura and Nei 1993) model suggested that the two different types of *ti* (A←→G and

C←→T) occur at different rates in the genome. They also suggested that *tv* occur at different rates than *ti* whereas all the *tv* were assumed to occur at similar rates (Tamura and Nei., 1993). The most acceptable codon substitution model (CSM) came into limelight during 1994 by Muse and Gaut (Muse and Gaut, 1994), it suggests different rates of synonymous substitutions and non-synonymous substitutions and assumes different purifying selection on non-synonymous substitutions. Depending on the variable rates of different types of substitutions, as discussed above different substitution models provided raw materials to estimate basic parameters of the evolutionary forces. *dN*/*dS* is one such approaches to estimate the direction and strength of selection in protein coding sequences (Aziz et al., 2022). dN is known as the number of non-synonymous changes per non-synonymous sites whereas dN is known as the number of synonymous changes per synonymous sites (Nei and Gojobori, 1986). Gojobori and Nei in their *dN*/*dS* estimation approach used Jukes and Cantor's formula (Jukes and Cantor, 1969) which is already discussed above. Eventually researchers used computational simulation approaches in estimation of the *dN*/*dS* values in coding sequences among closely related species (Ina, 1995). Also, *Ka*/*Ks* ratio, a similar approach in estimation of selective force, was developed which helped researchers to measure the balance of beneficial, neutral and harmful mutations in protein sequence evolution (Hurst, 2002). In the codon substitution model, using likelihood approach the $\frac{ti}{tv}$ ratio was estimated in accordance with the codon degeneracy class which was implemented in estimating *dN*/*dS* in coding sequences (Muse and Gaut, 1994; Goldman and Yang, 1994). Recently, by accounting $\frac{ti}{tv}$ substitution rate difference in codons as well as nonsense variations in PTC, the *dN*/*dS* values have been improved (Aziz et al., 2022). Along with the *dN*/*dS,* the estimation of *ti* to *tv* ratio is also an important parameter to study the evolutionary mechanism in the genomes of all the organisms.

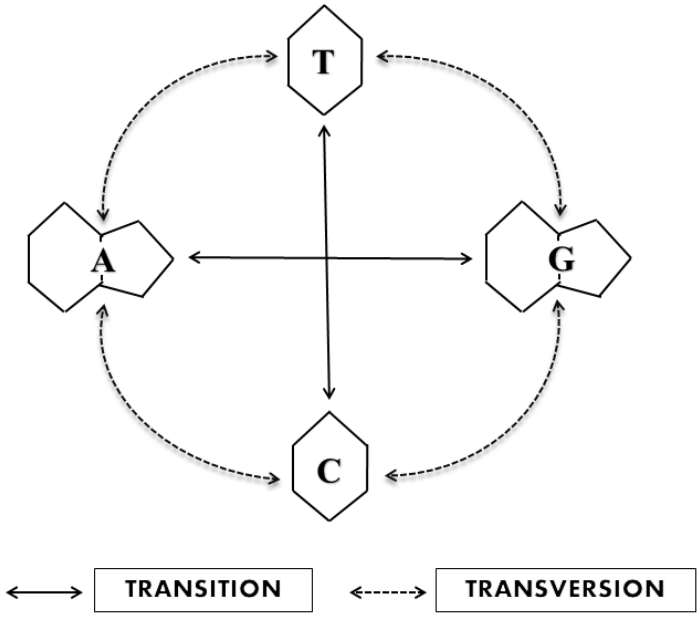

Fig. 1.2. Purine nucleotides (A & G) are shown as double-ringed structures and Pyrimidine nucleotides (T & C) are shown as single-ringed structures. When there is a base substitution between two similar classes of nucleotides, it is represented as a transition (solid lines with arrow). On the other hand, a base substitution between two different classes of nucleotides is depicted as a transversion (dotted line with arrow). In nature, 8 types of transversions and 4 types of transitions are possible.

The consequences of point mutations (*ti* or *tv*) in a coding sequence can be either synonymous or non-synonymous. Synonymous changes do not lead to alterations in amino acids, while non-synonymous changes result in modifications to the encoded amino acids in the variants (Holmes, 2003; Nei & Gojobori, 1986; Subramanian, 2013; Teng et al., 2008a). There can be different observations or theoretical calculations done in the 61 sense codons considering their SNV. Each nucleotide can be substituted by the remaining three nucleotides. Therefore, each triplet codon can give 9 different combinations of remaining codons while considering SNV (Gojobori et al., 1982; Graur and Li, 1997). In such manner, a theoretical calculation for the nine different combinations of the consequent codons can be calculated in terms of their synonymous or non-synonymous potential. For instance, in UUU, if a

substitution is implemented at the third position, the resulting codons will be UUC, UUA and UUG. Hence, only UUU→UUC will be synonymous transition (*Sti*) whereas UUU→UUA/UUG will be non-synonymous transversion (*Ntv*). Considering *ti* or *tv* as the events and synonymous or non-synonymous as the consequences, the resulting scenarios can be calculated in the remaining two positions. Overall, for UUU, synonymous transition (*Sti*), synonymous transversion (*Stv*), non-synonymous transition (*Nti*) and non-synonymous transversion (*Ntv*) can be calculated as one, zero, two and six respectively. However, this calculation of possible *Sti*, *Stv*, *Nti* and *Ntv* is not uniform across all the degenerate codons. The *two-fold degenerate* (TFD) and *four-fold degenerate* (FFD) codons exhibit different substitution possibilities (Fig. 1.3). It is noteworthy that the TFD codons have zero possibilities of Stv which contrasts with the FFD codons (Aziz et al., 2022). A table comprising all the number of possibilities of SNV, and its consequence is presented for the genetic code table (Table 1.1).

Fig. 1.3. The expected *Sti*, *Stv*, *Nti* and *Ntv* are calculated for two different degenerate codons. UUU being a TFD codons is has the expected *Sti*, *Stv*, *Nti* and *Ntv* as one, zero, two and six respectively, whereas GGU being a FFD codon has the expected *Sti*, *Stv*, *Nti* and *Ntv* as one, two, two and four respectively. Notably, the TFD codons lack the possibilities for a synonymous change through *tv*. Therefore, their composition in coding sequences is crucial while estimating transition bias.

Table 1.1. Estimated *Sti*, *Stv*, *Nti* and *Ntv* for all the codons

Codon	Sti	Stv	Nti	Ntv	Codon	Sti	Stv	Nti	Ntv	Codon	Sti	Stv	Nti	Ntv	Codon	Sti	Stv	Nti	Ntv
UUU	1	0	2	6	UCU	1	2	2	4	UAU	1	0	2	4	UGU	1	0	2	5
UUC	1	0	2	6	UCC	1	2	2	4	UAC	1	0	2	4	UGC	1	0	2	5
UUA	2	0	1	4	UCA	1	2	2	2	UAA	X	X	X	X	UGA	X	X	X	X
UUG	2	0	1	5	UCG	1	2	2	3	UAG	X	X	X	X	UGG	0	0	1	6
CUU	1	2	2	4	CCU	1	2	2	4	CAU	1	0	2	6	CGU	1	2	2	4
CUC	1	2	2	4	CCC	1	2	2	4	CAC	1	0	2	6	CGC	1	2	2	4
CUA	2	2	1	4	CCA	1	2	2	4	CAA	1	0	1	6	CGA	1	3	1	3
CUG	2	2	1	4	CCG	1	2	2	4	CAG	1	0	1	6	CGG	1	3	2	3
AUU	1	1	2	5	ACU	1	2	2	4	AAU	1	0	2	6	AGU	1	0	2	6
AUC	1	1	2	5	ACC	1	2	2	4	AAC	1	0	2	6	AGC	1	0	2	6
AUA	0	2	3	4	ACA	1	2	2	4	AAA	1	0	2	5	AGA	1	1	2	4
AUG	0	0	3	6	ACG	1	2	2	4	AAG	1	0	2	5	AGG	1	1	2	5
GUU	1	2	2	4	GCU	1	2	2	4	GAU	1	0	2	6	GGU	1	2	2	4
GUC	1	2	2	4	GCC	1	2	2	4	GAC	1	0	2	6	GGC	1	2	2	4
GUA	1	2	2	4	GCA	1	2	2	4	GAA	1	0	2	5	GGA	1	2	2	3
GUG	1	2	2	4	GCG	1	2	2	4	GAG	1	0	2	5	GGG	1	2	2	4

Robustness is a ubiquitous property among biological systems (Kitano, 2004). The genetic code is considered to be robust in the sense that it exhibits redundancy and tolerant to most harmful mutations (Maechiro and Kimura, 1998). This robustness is essential to ensure the fidelity of genetic information and preserves the biological significance. However, few codons in the genetic code table are vulnerable to non-sense substitutions which might lead to the formation of truncated proteins. To ensure the precision of synthesized proteins to be employed in different functions of the cell, the system has developed its own way of selective removal of deleterious mutations. Purifying selection is the essentiality for the existence of the species. Among the 61 sense codons, there are 18 such codons which can lead to stop codons with SNV. Hence these codons are known as pre-termination codons (PTC) (Golding and Strobeck, 1982). For instance, UGG can lead to UGA and UAG while having a *ti* at the 3rd and 2nd position respectively, hence UGG is considered as a PTC. PTC are strictly believed to be under stronger purifying selection while considering the non-synonymous substitutions. While researchers in the field are eager to unveil the random or evolved assignment of amino acids to

the codons or vice versa in the genetic code table, the impact of non-synonymous changes through *ti* or *tv* in terms of their deleterious effects remains a topic of debate. It is noteworthy that, in the split boxes, the *ti* is facilitated by the adjacent placement of R (U←→C) and Y (A←→G) nucleotides at the synonymous sites, yet the reason behind the assignment of amino acids in the nearby boxes are unknown.

Genes are essentially under selection, however the regions such as intergenic regions (IGRs) are generally devoid of selection (Shabalina et al., 2001). Hence the impact of mutations, in regions under stronger selection and weaker selection can be an essential tool in understanding the impact of different cellular processes like replication and transcription in the genome. There are several internal factors such as gene localization, strand bias, gene expression etc having an impact on the mutational spectra (Franklin and Lewontin, 1970). Apart from such factors, the role of context-dependent mutation is a trending area of research (Sung et al., 2015; Zhu et al., 2017; Aikens et al., 2019; Ling et al., 2020). Even though, the standard genetic code table is applicable universally with few exceptions in majority of the life systems, the exploration of homologous genes in prokaryotes and eukaryotes hold the potential to unveil more secrets about the genetic code table as well as the age of the proteins in terms of their essentiality.

1.4. Bibliography

- Aikens, R. C., Johnson, K. E., & Voight, B. F. (2019). Signals of variation in human mutation rate at multiple levels of sequence context. Molecular Biology and Evolution, 36(5), 955-965.
- Araújo, S., Tacão, M., Baraúna, R., Ramos, R., Silva, A. & Henriques, I. (2021). Genome analysis of two multidrug-resistant Escherichia coli O8: H9-ST48 strains isolated from lettuce. Gene, 785, 145603.

• Aziz, R., Sen, P., Beura, P. K., Das, S., Tula, D., Dash, M., ... & Ray, S. K. (2022). Incorporation of transition to transversion ratio and nonsense mutations, improves the estimation of the number of synonymous and non-synonymous sites in codons. DNA Research, 29(4), dsac023.

• Bernardi, G. (1989) The isochore organization of the human genome Annu. Rev. Genet., 23(1), 637-661.

• Bernardi, G. (2019) Chromosomal GC content and genome size in vertebrates: new insights from the modern synthesis between genomics and quantitative genetics Biol. Direct 14(1), 1-11.

• Berry, A., & Browne, J. (2022). Mendel and Darwin. Proceedings of the National Academy of Sciences, 119(30), e2122144119.

• Beura, P. K., Sen, P., Aziz, R., Satapathy, S. S., & Ray, S. K. (2023). Transcribed intergenic regions exhibit a lower frequency of nucleotide polymorphism than the untranscribed intergenic regions in the genomes of Escherichia coli and Salmonella enterica. Journal of Genetics, 102(1), 22.

• Bhagwat, A. S., Hao, W., Townes, J. P., Lee, H., Tang, H. & Foster, P. L. (2016). Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in Escherichia coli. Proceedings of the National Academy of Sciences, 113(8), 2176-2181.

• Boyden, A. (1947). Homology and analogy. A critical review of the meanings and implications of these concepts in biology. American Midland Naturalist, 648-669.

• Brocchieri, L. (2014). The GC content of bacterial genomes. J Phylogenetics Evol Biol, 2, 1-3.

• Brooks, S. (1998). Markov chain Monte Carlo method and its application. Journal of the royal statistical society: series D (the Statistician), 47(1), 69-100.

• Chandler, A. C. (1914). The effect of extent of distribution on speciation. The American Naturalist, 48(567), 129-160.

• Collins, D. W., & Jukes, T. H. (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics, 20(3), 386-396.

• Crick, F. (1953). Francis Crick. The double helix, 1951, 2-3.

• Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258), 561-563.

• Crick, F. (1988). Chapter 8: The genetic code. What mad pursuit: a personal view of scientific discovery. New York: Basic Books, 89-101.

• Darwin, C., Burrow, J. W., & Burrow, J. W. (2009). The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life (pp. 441-764). New York: AL Burt.

• Darwin, L. (1927). Natural selection. The Eugenics Review, 18(4), 285.

• Dawkins, R. (1986). The Blind Watchmaker. Norton & Company, Inc. ISBN 978-0393351491.

• Elson, D. & Chargaff, E. (1952). "On the deoxyribonucleic acid content of sea urchin gametes". Experientia. 8 (4): 143-145.

• Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. Nature Reviews Genetics, 8(8), 610-618.

• Esposito, M. (2016). From human science to biology: The second synthesis of Ronald Fisher. History of the Human Sciences, 29(3), 44-62.

• Felsenstein J (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach". *Journal of Molecular Evolution,* **17** (6): 368–76.

• Fisher, R. A. (1919). XV.—The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 52(2), 399-433.

• FitzRoy, R. (1839). Narrative of the Surveying Voyages of His Majesty's Ships Adventure and Beagle, Between the Years 1826-36, Describing Their Examination of the Southern Shores of South America, and the Beagle Circumnavigation of the Globe: In Three Volumes. Proceedings of the second expedition, 1831-1836, under the command of Captain Robert Fitz-Roy/[Robert Fitz-Roy] (Vol. 2). Colburn.

• Forsdyke, D. R. & Mortimer, J. R. (2000). Chargaff's legacy. Gene, 261(1), 127-137.

• Francino, M. P. & Ochman, H. (1997). Strand asymmetries in DNA evolution. Trends in Genetics, 13(6), 240-245.

• Franklin, I., & Lewontin, R. C. (1970). Is the gene the unit of selection?. Genetics, 65(4), 707.

• Gadgil, M., & Bossert, W. H. (1970). Life historical consequences of natural selection. The American Naturalist, 104(935), 1-24.

• Gibbs, H. L., & Grant, P. R. (1987). Oscillating selection on Darwin's finches. Nature, 327(6122), 511-513.

• Gojobori, T., Li, W. H., & Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. Journal of molecular evolution, 18, 360-369.

• Golding, G. B., & Strobeck, C. (1982). Expected frequencies of codon use as a function of mutation rates and codon fitnesses. Journal of Molecular Evolution, 18, 379-386.

• Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, *11*(5), 725-736.

• Goodhart, C. B. (1963). The Sewall Wright Effect. The American Naturalist, 97(897), 407-409.

• Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic acids research, 10(22), 7055-7074.

• Graur, D., & Li, W. H. (1997). Molecular evolution. Sinauer Associates, Sunderland, MA.

• Gregory, T.R. (2009) Understanding Natural Selection: Essential Concepts and Common Misconceptions. Evo Edu Outreach 2, 156–175.

• Griffiths, A. J. (2005). An introduction to genetic analysis. Macmillan.

• Gulisija, D. & Crow, J.F. (2007). "Inferring purging from pedigree data". Evol.; int. j. org. evol. 61 (5): 1043–51.

• Hardy, G. H. (1908). Mendelian proportions in a mixed population. Science, 28(706), 49-50.

• Hasegawa, M., Kishino, H. & Yano, Ta. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160–174.

• Heath, T. A., Hedtke, S. M. & Hillis, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. Journal of systematics and evolution, 46(3), 239.

• Hershberg, R. & Petrov, D. A. (2010). Evidence that mutation is universally biased towards AT in bacteria. PLoS genetics, 6(9), e1001115.

• Hershberg, R. (2015) Mutation--The Engine of Evolution: Studying Mutation and Its Role in the Evolution of Bacteria. Cold Spring Harb Perspect Biol. 1;7(9): a018077.

• Holmes, E. C. (2003). Patterns of intra-and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. Journal of virology, 77(20), 11296-11298.

• Hurst, L. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 2002 Sep;18(9):486.

• Hurst, L. (2009) Genetics and the understanding of selection. Nat Rev Genet 10, 83–93.

• Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Molecular biology and evolution, 2(1), 13-34.

• Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions, J. Mol. Evol., 40, 190–226.

• Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. Mammalian protein metabolism, 3(24), 21-132.

• Kimura, M. (1979). The neutral theory of molecular evolution. Cambridge University Press.

• Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution, 16, 111-120.

• Kino, K. & Sugiyama, H. (2001). Possible cause of G· C→ C· G transversion mutation by guanine oxidation product, imidazolone. Chemistry & biology, 8(4), 369-378.

• Kitano, H. (2004). Biological robustness. Nature Reviews Genetics, 5(11), 826-837.

• Kober, K. M., & Pogson, G. H. (2013). Genome-wide patterns of codon bias are shaped by natural selection in the purple sea urchin, Strongylocentrotus purpuratus. G3: Genes, Genomes, Genetics, 3(7), 1069-1083.

• Kornberg, A. (1984). DNA replication. Trends in Biochemical Sciences, 9(4), 122-124.

• Li, W. H. (1997). Molecular evolution. Sinauer Associates, Sunderland, MA.

• Li, W. H. (1997). Molecular evolution. Sinauer Associates, Sunderland, MA.

• Lindahl, T. & Nyberg, B. (1974). Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry, 13(16), 3405-3410.

• Ling, G., Miller, D., Nielsen, R., & Stern, A. (2020). A Bayesian framework for inferring the influence of sequence context on point mutations. Molecular Biology and Evolution, 37(3), 893-903.

• Liu, Y., Yang, Q., & Zhao, F. (2021). Synonymous but not silent: the codon usage code for gene expression and protein folding. Annual review of biochemistry, 90, 375-401.

• Lobry, J. R. (1995). Properties of a general model of DNA evolution under no-strand-bias conditions. Journal of molecular evolution, 40, 326-330.

• Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular biology and evolution, 13(5), 660-665.

• Lobry, J. R. & Lobry, C. (1999). Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. Molecular biology and evolution, 16(6), 719-723.

• Lobry, J. R. & Sueoka, N. (2002). Asymmetric directional mutation pressures in bacteria. Genome biology, 3, 1-14.

• Lynn, D. J., Singer, G. A., & Hickey, D. A. (2002). Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic acids research, 30(19), 4272-4277.

• Maeshiro, T., & Kimura, M. (1998). The role of robustness and changeability on the origin and evolution of genetic codes. Proceedings of the National Academy of Sciences, 95(9), 5088-5093.

• Mashima, J., Kodama, Y., Fujisawa, T., Katayama, T., Okuda, Y., Kaminuma, E. & Takagi, T. (2016). DNA data bank of Japan. Nucleic acids research, gkw1001.

• Mayr, E. (1963). Animal species and evolution. Harvard University Press.

• McInerney, J. O. (1998). Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proceedings of the National Academy of Sciences, 95(18), 10698-10703.

• Morgan, G. (1998) Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959-1965. J. Hist. Biol. pp. 155-178.

• Mugal, C. F., von Grünberg, H. H. & Peifer, M. (2009). Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Molecular biology and evolution, 26(1), 131-142.

• Muse, S. V. & Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular biology and evolution, 11(5), 715-724.

• Muto, A. & Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. Proceedings of the National Academy of Sciences, 84(1), 166-169.

• Nachman, M. W. (2004). Haldane and the first estimates of the human mutation rate. Journal of Genetics, 83, 231-233.

• NCBI Resource Coordinators. (2016). Database resources of the national center for biotechnology information. Nucleic acids research, 44(D1), D7-D19.

• Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular biology and evolution, 3(5), 418-426.

• Nirenberg, M. W. (1963). The genetic code. Scientific American, 208(3), 80-95.

• Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K. & Sugino, A. (1968). Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. Proceedings of the National Academy of Sciences, 59(2), 598-605.

• Osawa, S., & Jukes, T. H. (1989). Codon reassignment (codon capture) in evolution. Journal of molecular evolution, 28, 271-278.

• Paul, S., Bag, S. K., Das, S., Harvill, E. T., & Dutta, C. (2008). Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome biology, 9, 1-19.

• Powdel, B. R., Satapathy, S. S., Kumar, A., Jha, P. K., Buragohain, A. K., Borah, M. & Ray, S. K. (2009). A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's second parity rule). DNA research, 16(6), 325-343.

• Prehn, R.T. (2005) The role of mutation in the new cancer paradigm. Cancer Cell Int. 26;5(1):9.

• Revell, L. J., Mahler, D. L., Peres-Neto, P. R. & Redelings, B. D. (2012). A new phylogenetic method for identifying exceptional phenotypic diversification. Evolution, 66(1), 135-146.

• Rocha, E. P., Touchon, M. & Feil, E. J. (2006). Similar compositional biases are caused by very different mutational effects. Genome research, 16(12), 1537-1547.

• Rodrigue, N., Philippe, H., & Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proceedings of the National Academy of Sciences, 107(10), 4629-4634.

• Sarkar, S. (1992)"Haldane as Biochemist: The Cambridge Decade, 1923–1932", The Founders of Evolutionary Genetics, Boston Studies in the Philosophy of Science, vol. 142, Dordrecht: Springer Netherlands, pp. 53–81.

• Schmidt, S., Gerasimova, A., Kondrashov, F. A., Adzuhbei, I. A., Kondrashov, A. S., & Sunyaev, S. (2008). Hypermutable non-synonymous sites are under stronger negative selection. PLoS genetics, 4(11), e1000281.

• Sen, P., Kurmi, A., Ray, S. K., & Satapathy, S. S. (2022). Machine learning approach identifies prominent codons from different degenerate groups influencing gene expression in bacteria. Genes to Cells, 27(10), 591-601.

• Sen, P., Aziz, R., Deka, R. C., Feil, E. J., Ray, S. K. & Satapathy, S. S. (2022). Stem region of tRNA genes favors transition substitution towards keto bases in bacteria. Journal of Molecular Evolution, 90(1), 114-123.

• Seplyarskiy, V. B., Kharchenko, P., Kondrashov, A. S. & Bazykin, G. A. (2012). Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. Molecular biology and evolution, 29(8), 1943-1955.

• Seward, E. A., & Kelly, S. (2016). Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. Genome biology, 17, 1-15.

• Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A., & Kondrashov, A. S. (2001). Selective constraint in intergenic regions of human and mouse genomes. Trends in Genetics, 17(7), 373-376.

• Shabalina, S. A., Spiridonov, N. A., & Kashina, A. (2013). Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic acids research, 41(4), 2073-2094.

• Shendure, J. & Aiden, E. L. (2012) The expanding scope of DNA sequencing Nat. Biotechnol. 30(11), 1084-1094.

• Simpson, G. G. (1953). The major features of evolution. Columbia University Press.

• Singer, C. E. & Ames, B. N. (1970). Sunlight Ultraviolet and Bacterial DNA Base Ratios: Natural exposure to ultraviolet may be an evolutionary pressure toward high guanine plus cytosine in DNA. Science, 170(3960), 822-826.

• Stanley, S. M. (1975). A theory of evolution above the species level. Proceedings of the National Academy of Sciences, 72(2), 646-650.

• Stern, C. (1943). The hardy-weinberg law. Science, 97(2510), 137-138.

• Subramanian, S. (2013). Significance of population size on the fixation of nonsynonymous mutations in genes under varying levels of selection pressure. Genetics, 193(3), 995-1002.

• Sueoka, N. (1964). On the evolution of informational macromolecules. In Evolving genes and proteins (pp. 479-496). Academic Press.

• Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. Journal of molecular evolution, 40, 318-325.

• Sukhodolets, V. V. (1986) The role of natural selection in evolution. Genetika, 22(2):181-93.

• Sung, W., Ackerman, M. S., Gout, J. F., Miller, S. F., Williams, E., Foster, P. L., & Lynch, M. (2015). Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments. Molecular biology and evolution, 32(7), 1672-1683.

• Tamura K., Nei M. (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees". *Molecular Biology and Evolution*. **10** (3): 512–26.

• Teng, S., Michonova-Alexova, E., & Alexov, E. (2008). Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. Current pharmaceutical biotechnology, 9(2), 123-133.

• Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequence. *Lecture of Mathematics for Life Science*, *17*, 57.

• Wahl, M. C., & Sundaralingam, M. (1997). Crystal structures of A-DNA duplexes. Biopolymers: Original Research on Biomolecules, 44(1), 45-63.

• Watson, J. D., & Crick, F. H. (1953, January). The structure of DNA. In Cold Spring Harbor symposia on quantitative biology (Vol. 18, pp. 123-131). Cold Spring Harbor Laboratory Press.

• Watson, J. D., & Crick, F. H. (1958). On protein synthesis. In The Symposia of the Society for Experimental Biology (Vol. 12, pp. 138-163).

• Woese, C. R. & Fox, G. E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences, 74(11), 5088-5090.

• Wright, S. (1942). Statistical genetics and evolution.

• Wright, S. (1984) The first Meckel oration: on the causes of morphological differences in a population of guinea pigs. Am J Med Genet. 18(4):591–616.

• Wu, H., Zhang, Z., Hu, S. & Yu, J. (2012). On the molecular mechanism of GC content variation among eubacterial genomes. Biology direct, 7, 1-16.

• Yang, Z., & Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Molecular biology and evolution, 25(3), 568-579.

• Zangerl, R. (1948). The methods of comparative anatomy and its contribution to the study of evolution. Evolution, 351-374.

• Zhao, X., Zhang, Z., Yan, J. & Yu, J. (2007). GC content variability of eubacteria is governed by the pol III α subunit. Biochemical and biophysical research communications, 356(1), 20-25.

• Zhu, Y., Neeman, T., Yap, V. B., & Huttley, G. A. (2017). Statistical methods for identifying sequence motifs affecting point mutations. Genetics, 205(2), 843-856.

• Zuckerkandl, E., Jones, R. T., & Pauling, L. (1960). A comparison of animal hemoglobins by tryptic peptide pattern analysis. Proceedings of the National Academy of Sciences, 46(10), 1349-1360.

• Zuckerkandl, E. & Pauling, L.B. (1962) Molecular disease, evolution, and genetic heterogeneity. In: Kasha, M. and Pullman, B., Eds., Horizons in Biochemistry, Academic Press, New York, 189-225.