Chapter 21 Comparative Genomics: Insight into Human Health and Disease

Toshiaki Nakajima, Akinori Kimura

INTRODUCTION

Comparative genomics is an approach of extraordinary promise for studying the biological significance of the genome from the evolutionary points of view, because genome sequence is a record of the evolutionary history of organisms. It also has a crucial role in furthering our understanding of the pathophysiological mechanisms operated by human disease genes. The functional analyses of a homologue of disease gene in model organisms can provide important information in understanding the biochemical function of the human disease gene.

The comparative genomic analysis consists of 4 main processes, including the following steps. Step1: Generation of sequence data; Step 2: Identification of homologue sequences among related genomes; Step 3: Multiple-sequence alignment; and step 4: Analyses to determine the biological significance of the homologue sequences (Fig. 21.1).[1] The number of species available for genomic comparison has been rapidly increasing. The human genome project, which was started in the late 1980s and ostensibly "completed" in 2003, has actively stimulated the development of genome projects for model organisms, such as *C. elegans*, fruit fly, and mouse. Furthermore, the rapid progress in sequence technology, such as massive parallel sequencing technology,[2] has further accelerated the increase in the number of complete or semi-complete genome sequences of other organisms. These sequences serve as the material for large-scale analysis of comparative genomics. Moreover, rapid progress in the field of bioinformatics, which provides useful tools to reconstruct homologous sequences among related species and to generate base-pair alignment, has advanced studies of comparative genomics. There are several data-bases that are available for large-scale comparative sequence analyses (Table 21.1).[1]

Table 21.1: Main features of the data bases available for large-scale comparative genomics

Data base	*Web site*	*Main feature*
Ensembl Genome Browser[63]	http://www.ensembl.org/index.html	Joint project between EMBL-EBI and the Sanger Center to develop a software system which produces and maintains automatic annotation on eukaryotic genomes.
Galaxy[64]	http://g2.bx.psu.edu.	Interactive system that combines the power of existing genome annotation databases with a simple Web portal to enable users to search remote resources, combine data from independent queries, and visualize the results.
UCSC Genome Browser[65]	http://genome.ucsc.edu/	The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz. The UCSC Genome Browser website contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides a portal to the ENCODE project.
VISTA[49]	http://genome.lbl.gov/vista/index.shtml	The VISTA site includes computational tools for comparative genomics and provides whole-genome alignments of large vertebrate genomes

In this chapter we will focus on the methods used to evaluate the biological significance of homologous sequences (step 4 in Fig. 21.1). In particular, we focus on the methods based on the theory of natural selection, but we will not touch on other steps of comparative genomics (Fig. 21.1). The underlying principle behind the identification of the biological significance is that the genomic regions, which have been evolving more rapidly (positive selection) or more slowly (negative selection) than the local rate of neutral evolution, are tightly linked to biological significance. Human diseases are occasionally linked to human evolutionary process and population history, so that comparative genomics will also provide insight into human health and disease.

HUMAN DISEASE AND EVOLUTION

The molecular evidence that the human evolutionary process is tightly linked to human disease has been steadily accumulating during the last decades. For example, the evolutionary processes, especially natural selection, affect the frequency of disease-related mutations in the general population. Disease-related mutations, which had been introduced through errors in DNA replication, are believed to be rapidly removed by natural negative selection triggered by the deleterious consequences induced by

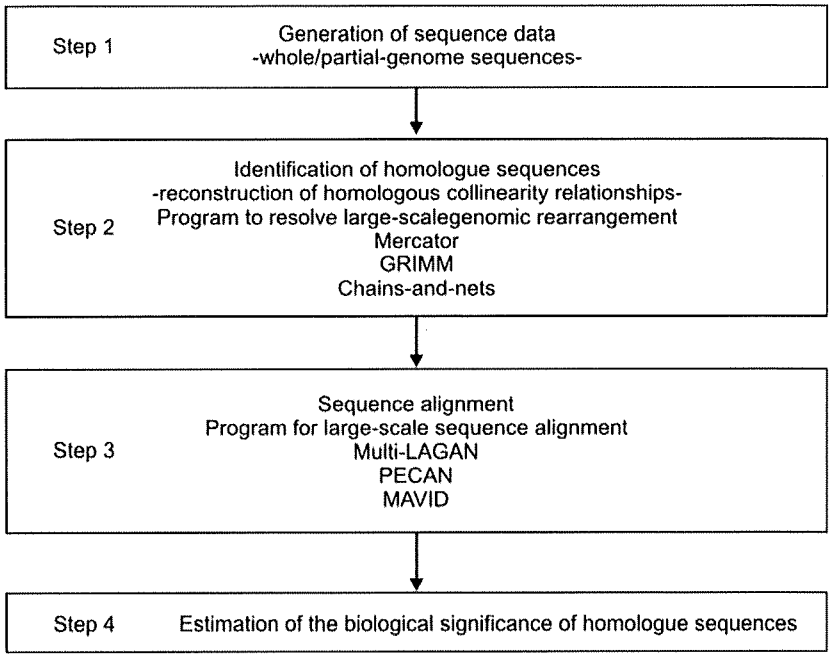


Fig. 21.1: Four main steps of the comparative genomic analyses
Step1: Generation of sequence data. **Step 2:** Identification of homologue sequences among related genomes. Mercator,[66] GRIMM,[67] and Chains-and-nets[68] are programs to resolve large-scale genomic rearrangement in genome-wide scale comparative genomics. **Step 3:** Multiple-sequence alignment. Multi-LAGAN,[69] PECAN,[70,71] and MAVID[72] are programs for large-scale sequence alignment. **Step 4:** Analyses to determine the biological significance of homologue sequences.[1]

the mutations. However, certain environments may result in a special selective advantage for a disease mutation. For example, sickle-cell anemia provides a conspicuous example.[3] Sickle-cell mutations are relatively prevalent in Africa, where *Plasmodium falciparum* malaria is common, because the individuals heterozygous for the sickle-cell mutation, who are resistant to malaria infection conferred by the mutation, have a distinct survival advantage. This example illustrates how disease mutations have arisen and maintained under the pressure of natural selection operating differently in different environments.

It has been reported that among primates humans and chimpanzees are sensitive to HIV infection, but chimpanzees are naturally resistant to the development of AIDS. Interestingly, the nucleotide diversity in the MHC class I genes in chimpanzees is much lower than that in humans, although nucleotide diversity in the chimpanzee genome in general is much higher than that in the human genome.[4, 5] Furthermore, common Patr-B alleles observed in chimpanzees have certain functional relationships to the HLA-B alleles, HLA-B27 and HLA-B57, which are known to be associated with the resistance and/or protection to HIV infection or development of AIDS.[5]

The amino acid sequences of HIV-1 epitopes recognized by CTL in the context of these Patr-B alleles are more or less similar to those of the HIV-1 resistant human alleles, HLA-B27 and HLA-B57. Such evidence suggests that HIV-induced selective sweeps had occurred, and that only the descendants of chimpanzees that were resistant to HIV/AIDS have survived in the course of chimpanzee evolution. This is an example of disease susceptibility genes under the control of natural selection in the course of primate evolution.

The "thrifty genotype" hypothesis advanced by James V. Neel[6] is also an intriguing evolutionary scenario to account for the emergence of common human diseases. The genotypes predisposing to human common diseases (diabetes in Neel's original hypothesis) had previously been advantageous in the course of human evolution, but became deleterious in the changed context of the modern world. A change in the environment or life style factor, especially a dietary factor, may be responsible for the rapid increase in the prevalence of certain human diseases, such as diabetes mellitus and essential hypertension. Certain "thrifty" genotypes had thrived in scarce food environment. In our previous study,[7] genetic diversity in the human angiotensinogen gene, which is a susceptibility gene for essential hypertension,[8] was shown to be in support of this hypothesis.

Alternatively, it has been reported that the genes under the pressure of natural selection are occasionally associated with human diseases. HLA genes, which have the highest nucleotide diversity and the highest nonsynonymous substitution rate in the human genome, are a conspicuous example. The patterns of sequence variations in the HLA genes have been shaped by positive natural selection.[9] It is widely accepted that several human diseases are tightly linked to sequence variations in the HLA genes.[10]

These lines of evidence support a tight association between the evolution of human genome and certain human diseases, so that an improved understanding of the human genome from the evolutionary point of view might provide us with the critically important insight into human health and disease. In particular, understanding the effect of natural selection on the human genes might be helpful to understand the evolutionary aspects of human health and disease. Thus, in this chapter we will describe how to identify the genomic segments under the control of natural selection by comparative genomics.

COMPARATIVE GENOMICS AND NATURAL SELECTION

Assuming that all comparable organisms in a given lineage originate from a common ancestor and that the evolutionary process can be traced by means of genomic sequences, comparative genomics is a useful approach to study the evolution.[1, 11-13] The combined actions of mutation and selection are believed to be powerful driving forces of evolution. Mutations randomly occur in the genome, and their traits are then exposed to intense selective

pressure. The fates of genetic mutations are tightly linked to their functional impacts on the phenotypic traits of organisms in the population.

Mutations are categorized into three types, advantageous, deleterious, or neutral mutations.[14] When a mutation confers a functional advantage on its carrier, it would increase the fitness of its carrier and eventually increase the chance of survival in the population (positive natural selection), resulting in a fixed difference between species. More frequently, a new mutation, which reduces the fitness of its carrier due to its deleterious impact on gene function, will be removed from the population (negative or purifying selection). Most new mutants arising in a population have no impact on gene function. Such mutations are selectively neutral and the fate in the population is determined by chance effect (neutral selection). Comparative genomics make it possible to trace all of these processes of natural selection in comparable organisms.

Evaluation of Selective Pressure on the Coding Regions: Synonymous and Nonsynonymous Nucleotide Substitution Rates

When comparing orthologous coding sequences from related species, the nucleotide differences among the orthologues are classified into two types, synonymous and nonsynonymous substitutions. Nucleotide substitutions which occur without affecting the amino acid sequence are called synonymous (silent) nucleotide substitutions, while those with amino acid replacements are called nonsynonymous nucleotide substitutions. As described in the following sections, estimates of synonymous (Ks) and nonsynonymous (Ka) substitution rates have provided important information for evaluating the selective pressure on the target genomic region in the evolutionary process. Ka and Ks are defined as the number of nonsynonymous substitutions (N_d) per nonsynonymous site (N), and the number of synonymous substitutions (S_d) per synonymous site (S) for any pair of sequences, respectively.[14]

Synonymous and nonsynonymous sites in coding sequence are the sites in a codon, where nucleotide changes would result in synonymous and nonsynonymous substitutions, respectively. One way to calculate them is based on the patterns of nucleotide substitutions in the genetic code and amino acid substitutions (degeneracy in the genetic code).[14] As shown in Table 21.2, nucleotide sites are classified into nondegenerate, twofold-degenerate, and fourfold-degenerate sites. A site is nondegenerate (denoted by G, A, T or C) when all possible changes at this site are nonsynonymous, twofold-degenerate (denoted by Y or R) when one of the three possible changes is synonymous, and fourfold-degenerate (denoted by N) when all possible changes at the site are synonymous. For simplicity, the third position in each of the three codons for Ile (denoted by H) is treated as a twofold-degenerate site, although the degeneracy at this position is actually threefold. When each of the average number of non-degenerate, twofold-

Table 21.2: Degeneracy in the mammalian genetic code

1st base	*2nd base* T	C	A	G
T	TTY (Phe)	TCN (Ser)	TAY (Tyr)	TGY (Cys)
	TTR (Leu)		TAR (Stop)	TGA (Stop)
				TGG (Trp)
C	CTN (Leu)	CCN (Pro)	CAY (His)	CGN (Arg)
			CAR (Gln)	
A	ATH (Ile)	ACN (Thr)	AAY (Asn)	AGY (Ser)
	ATG (Met)		AAR (Lys)	AGR (Arg)
G	GTN (Val)	GCN (Ala)	GAY (Asp)	GGN (Gly)
			GAR (Glu)	

Non-degenerate sites are denoted by G, A, T, or C.
Twofold-degenerate sites are denoted by Y or R.
Threefold-degenerate sites are denoted by H.
Fourfold-degenerate sites are denoted by N.

degenerate, and fourfold-degenerate sites in the two compared sequences is denoted by L_0 (non-degenerate), L_2 (twofold-degenerate), and L_4 (fourfold-degenerate), respectively, the number of synonymous and nonsynonymous are given by the formula,

$$S = L_2 / 3 + L_4 \text{ and } N = 2L_2 / 3 + L_0$$

For example, here are two coding sequences for three codons to estimate the value of Ks and Ka,

Sequence 1: GTC-TTT-GAT (Val-Phe-Asp)

Sequence 2: GTT-GTT-GCG (Val-Val-Ala)

The numbers of the three degenerate types of sites in each of the two sequences are as follows:

	Non-degenerate	*Twofold-degenerate*	*Fourfold-degenerate*
Sequence 1	6	2	1
Sequence 2	6	0	3
Average	6	1	2

Hence, $S = 1 \times 1/3 + 2 \cong 2.33$ and $N = 1 \times 2/3 + 6 \cong 6.67$

Next, to estimate the Ks and Ka, we need to compare codon by codon the two sequences and infer the nucleotide substitutions in each pair of codons. When comparing the two first codons, GTC (Val) and GTT (Val), a

C to A synonymous nucleotide substitution at the four-degenerate site is easily inferred. Similarly, a T to G nonsynonymous nucleotide substitutions at the nondegenerate site is inferred in the comparison of the two second codons, TTT (Phe) and GTT (Val). For two codons with one nucleotide difference, the number of synonymous or nonsynonymous substitutions is easily inferred. However, the case of the two third codons is more complicated. For the two third codons, GAT (Asp) and GCG (Ala), there are two possible minimum evolutionary pathways;

Pathway 1: GAT (Asp) ↔ GCT (Ala) ↔ GCG (Ala)
Pathway 2: GAT (Asp) ↔ GAG (Glu) ↔ GCG (Ala)

In pathway 1, one nonsynonymous nucleotide substitution and one synonymous nucleotide substitutions are estimated, whereas the pathway 2 involves two nonsynonymous substitutions. For the estimation of the number of synonymous and nonsynonymous substitutions, we may apply the weight for the probability of each possible pathway, because synonymous substitutions have occurred more often than nonsynonymous substitutions in the course of evolution. There are several methods to estimate the weight for all possible codon pairs.[15-17] When estimated by Nei-Gojobori method,[16] which considers all pathways with equal probability, but which excludes those pathways that go through stop codons, the numbers of synonymous and nonsynonymous substitutions for sequence 1 and 2 are 1.50 and 2.50, leading the Ks and Ka for them to be 1.46 (1.50/2.33) and 0.52 (2.50/6.67), respectively.

As shown above, inferring the number of synonymous and nonsynonymous substitution is not simple or straightforward. Furthermore, the maximum-likelihood approach taking advantage of the probability theory has been developed.[18] It is necessary to pay attention to the differences in the estimated values of Ka and Ks by several different methods to obtain an accurate picture.

Natural Selection and Synonymous and Nonsynonymous Nucleotide Substitution Rates

In the comparisons of coding sequences from several species, estimates of synonymous (Ks) and nonsynonymous (Ka) substitution rates between two given sequences are extremely useful in evaluating the selective pressure on the target genomic region in the evolutionary process. Because the value of Ks is proposed to be almost equal to the nucleotide substitution rate under the neutral selection, the Ka/Ks ratio is a good parameter for identifying genomic segments under the pressure of natural selection. In the absence of selective pressure, the values of Ks and Ka should be more or less the same ($Ka/Ks \cong 1$). When the genomic segment has been under the pressure of negative (purifying) selection, the value of Ka is expected to be much lower than Ks ($Ka/Ks<1$), because selection would occur at the

amino acid level. Amino acid sequences under the pressure of negative selection have been strictly conserved in the course of evolution. The functional significance of a target for the negative selection is considered as such that certain amino acid sequences have not been altered in the course of primate evolution. In clear contrast, if there has been a pressure of positive Darwinian selection, the value of Ka would be expected to be higher than Ks (Ka/Ks>1). In this scenario, rapid changes in amino acid sequences, which confer novel or gain in function, would have been advantageous in the course of evolution.

Genes under the Pressure of Natural Selection Derived from Whole-genome Comparative Analysis in Primates

Recently, large-scale genome sequences of chimpanzees[19] and rhesus macaque monkeys[20] have been available, and the whole genome sequencings of gorilla, orangutan, and marmoset are under way. Comparative genomic analyses among primates are definitely useful to address the issue of which genetic changes have made us uniquely human. Furthermore, they are also useful to identify the susceptibility genes for human diseases, and to understand the pathophysiological mechanisms of human disease genes, because biological differences among primates, such as differences in the susceptibility to several different diseases, have been reported.[21]

To resolve these issues, identification of the genes that have come under the pressure of natural selection in the course of primate evolution is of critical importance. Actually, comparisons between human and chimpanzee and between human and rhesus genome have already suggested that dozens of genes have been under the pressure of natural selection in the course of primate evolution, in particular those involved in host-pathogen interactions, reproduction, sensory system, and more.[19, 20, 22] However, it is difficult to establish the functional significance of the genes under the natural selection, because of the limitations on the functional experiments required by the ethical reasons. We cannot practically apply, for example, the knockout experiment in primates to reveal the functional differences induced by the genetic difference in the genes under the control of natural selection. Thus, further studies for comparative analyses of the genomes and phenotypes among primates should be undertaken.

Comparative Genomics in Coding Regions: Lesson from TLR Sequence Comparisons among Primates

In this section, we demonstrate an example of comparative genomic analysis. That is the analyses for Toll-like receptor (TLR)-related genes, which might be helpful in understanding how to study the biological significance of orthologous genes based on comparative genomics. TLRs recognize molecules derived from pathogens and play crucial roles in the

innate immune system.[23, 24] To investigate whether the TLR-related genes have come under the natural selection pressure in the course of primate evolution, we compared the nucleotide sequences of sixteen TLR-related genes, including ten *TLRs* (*TLR1-10*), four genes linked to signal transduction (*MYD88, TILAP, TICAM1, and TICAM2*), and 2 genes linked to *TLR4* (*MD2* and *CD14*), among seven primates, including human, chimpanzee, bonobo, gorilla, orangutan, crab-eating macaque, and rhesus macaque.[25] MD2 and CD14 are key molecules in the LPS signaling through TLR4.[26-28]

To evaluate the nonsynonymous/synonymous substitution ratio, we applied the Bn-Bs program.[29] This program uses a modified Nei-Gojobori method[16] to estimate pairwise synonymous and nonsynonymous distances among the sequences, and then estimates the branch lengths in terms of synonymous (bs) and nonsynonymous substitutions (bn) per site by using the ordinary least-squares method, while the tree topology is given. Σbn and Σbs indicate the value summing up the bn and bs scores, respectively, in the lineages. The values of bs and bn are almost identical to those of Ka and Ks, respectively, which were described in the previous section. When the value of Σbn and Σbs and the ratio of Σbn/Σbs were evaluated for the entire coding sequences from each gene, all values of the Σbn/Σbs ratio from the analyzed genes were much lower than 1.0, suggesting that these genes have been under the pressure of negative selection (Table 21.3).

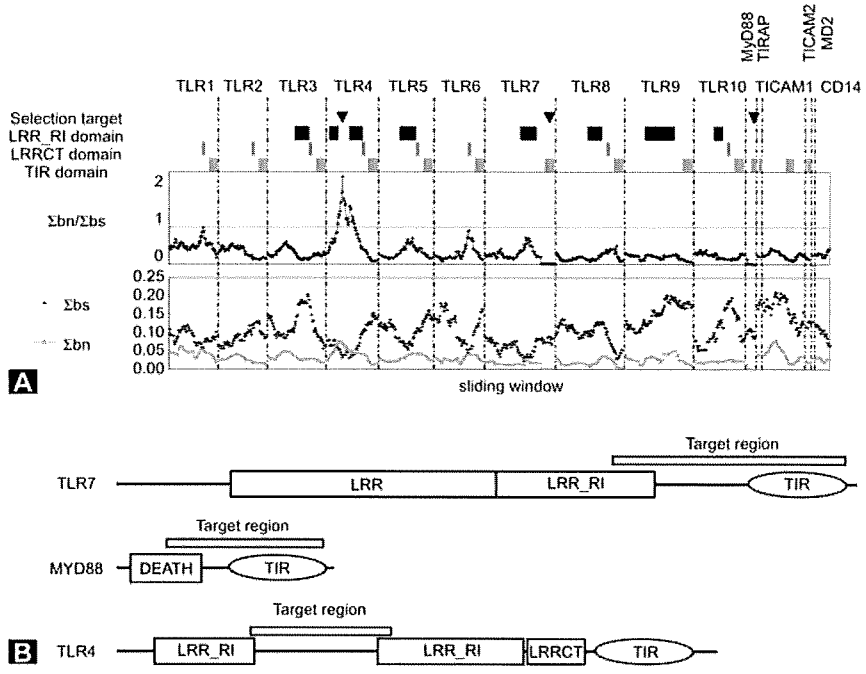
Table 21.3: The nonsynonymous and synonymous substitution ratio, GC content and CBI for 16 TLR-related genes among 7 primates

Gene	*Chromosome (Human)*	*Σbn*	*Σbs*	*Σbn/Σbs*	*G + C 2nd*	*G + C 3rd*	*G + C all*	*CBI*
TLR1	4p14	0.041	0.095	0.429	0.319	0.433	0.394	0.188
TLR2	4q32	0.025	0.086	0.290	0.344	0.452	0.417	0.213
TLR3	4q35	0.032	0.121	0.267	0.314	0.419	0.399	0.193
TLR4	9q32-33	0.038	0.085	0.447	0.323	0.472	0.428	0.214
TLR5	1q41-42	0.030	0.108	0.282	0.33	0.487	0.438	0.189
TLR6	4p14	0.030	0.120	0.240	0.307	0.408	0.384	0.195
TLR7	Xp22.3-p22.2	0.014	0.069	0.202	0.317	0.442	0.409	0.193
TLR8	Xp22.3-p22.2	0.020	0.095	0.209	0.321	0.41	0.394	0.144
TLR9	3p21.3	0.029	0.153	0.187	0.418	0.804	0.619	0.631
TLR10	4p14	0.024	0.106	0.228	0.311	0.368	0.378	0.262
MYD88	3p22-p21.3	0.009	0.096	0.094	0.435	0.712	0.588	0.484
TIRAP	11q23-q24	0.035	0.164	0.216	0.531	0.691	0.608	0.497
TICAM1*	19p13.3	0.039	0.171	0.227	0.523	0.732	0.64	0.486
TICAM2	5q23.1	0.020	0.119	0.167	0.383	0.383	0.427	0.318
MD2	8q21.11	0.015	0.054	0.269	0.353	0.361	0.361	0.408
CD14	5q31.1	0.013	0.040	0.332	0.488	0.704	0.63	0.422

* TICAM1 has a CCT (Pro)-repeat variation.

Next, we performed a sliding window plot analysis (600 bp-window with 30 bp-steps) throughout these genes to identify the genomic segments that might have undergone the natural selection. Analysis of the Σbn/Σbs ratio revealed the presence of both the strictly conserved and rapidly evolving regions among the TLR-related genes. Three candidate segments, where the pressure of the negative or positive natural selection might have operated, were identified in *TLR7*, *MYD88*, and *TLR4* (Fig. 21.2A).

Two target segments showed little nonsynonymous nucleotide differences (Σbn/Σb<<1) among seven primates (Fig. 21.2B). One segment was located at the coding segment for the C-terminal of TLR7 and the other segment was that for the C-terminal of MYD88, both of which encode the intracellular Toll/interleukin 1 receptor (TIR) domain (Fig. 21.2B), implying that the genomic regions corresponding to the TIR domains were the targets of negative natural selection. We then evaluated the Σbn/Σbs ratios for the TIR domains for fourteen genes carrying the TIR domains. The sizes of the genomic segments encoding TIR domains were between 249 and 426 base



Figs 21.2A and B: Sliding window plot analysis for the TLR-related gene to identify the genomic segments which have undergone natural selection. (A) The values of Σbn/Σbs, Óbn, and Óbs based on the sliding window plot analysis for the TLR-related gene (600 bp-window with 30 bp-steps). The arrow heads indicate the candidate segments for the pressure of positive or negative natural selection. Three conserved domain structures, LRR_RI (Leucine-rich repeats, ribonuclease inhibitor-like subfamily), LRRCT (Leucine rich repeat C-terminal domain), and TIR (Toll/interleukin-1 receptor homology domain), were referred from CD-search.[73] (B) Natural selection target in *TLR7* (negative selection), *MYD88* (negative selection), and *TLR4* (positive selection)

pairs (bp) (average 393 bp) that were smaller than the window size (600 bp) used in our analysis, implicating that the window analysis would underestimate the Σbn/Σbs ratios for the TIR domains. The values of Σbn and Σbn/Σbs ratio for TIR domains displayed lower values when compared with those of the non-TIR coding sequences, except for the Σbn/Σbs ratio from *TLR10* (Fig. 21.3). In particular, *TLR7*, *TLR8*, *TLR9*, *MYD88*, and *TICAM2* have much lower values for Σbn in the TIR domains. Taken together, these observations suggest that the TIR domains were under the control of negative/purifying selection. The TIR domains are key components in the TLR signal transduction. In particular, the adaptor protein MYD88 is tightly linked to all the TLR signaling pathways except for TLR3.[23, 24] This may be the reason for that the amino acid sequences have not been altered in the course of primate evolution. Most of the mutations arising at the TIR domains would in all likelihood were deleterious, and thus reduced the fitness of primates which had harbored such changes.

On the other hand, sequence comparisons among the primates support positive Darwinian selection at the extracellular domain of TLR4, for which the Σbn/Σbs ratios were much higher than 1.0 and the highest value in the 600 bp-window was 2.37 that was significant on Z-test[29, 30] with Z-score 2.16 (p-value < 0.01) (Fig. 21.2). The TLR4 target region encoding the extracellular domain next to the domain with Leucine-rich repeats (Fig. 21.2B) has been reported to be hypervariable and contribute to the species-specific recognition of several molecules, such as taxol, lipid IVa,

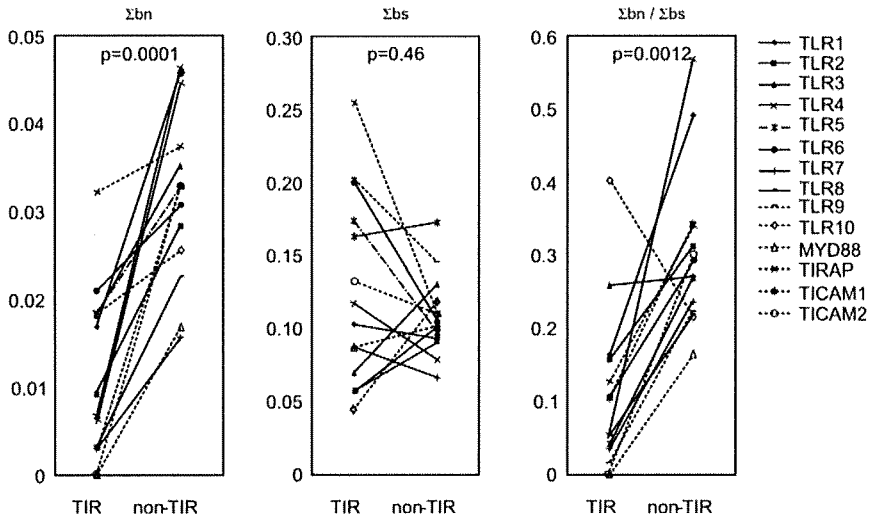


Fig. 21.3: The values of Σbn, Σbs, and Σbn/Σbs for the TIR domains and non-TIR coding regions of TLR family and related genes. Data for fourteen genes, including *TLRs (TLR1-10)*, *MYD88*, *TILAP*, *TICAM1*, and *TICAM2*, are presented

and LPS.[31-33] This target region also was reported to be linked to LPS susceptibility in humans.[34] The missense mutation D299G, where an aspartic acid was replaced by a glycine at the 299 amino acid position of human TLR4, was associated with a blunted response to LPS and increased susceptibility to Gram-negative bacterial infections. An aspartic acid corresponding to the 299 amino acid position of human TLR4 has been highly conserved among great apes and gibbons, whereas it was replaced by a glycine in the lineage of Old World Monkeys. These results indicate that the sensitivity to a certain type of LPS might differ between the lineage of great apes and that of the Old World Monkeys.

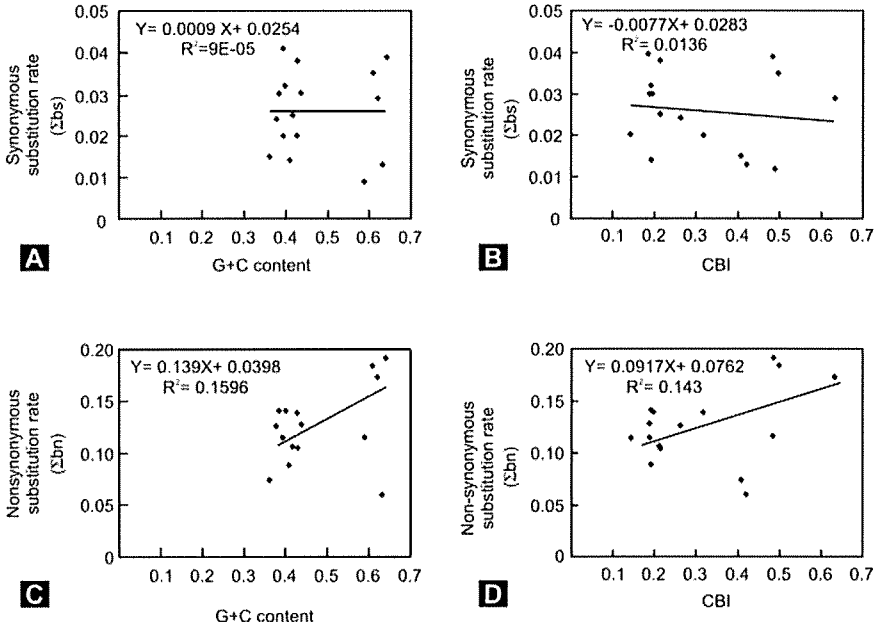
Given that TLR4 recognizes a wide variety of ligands such as LPS and viral envelope proteins, the differences in the species-specific susceptibility to infectious disease might be linked to the natural selection pressure. As shown in the case of TLR-related genes, comparative sequence analyses provide us with the useful information to evaluate the functional significance of genes from an evolutionary point of view.

OTHER FACTORS AFFECT SYNONYMOUS NUCLEOTIDE SUBSTITUTION RATE IN CODING REGIONS

In previous sections, the ratio of Ka/Ks was used as a parameter to estimate the selective pressure on the genomic segment. Since the synonymous substitutions do not cause any changes in amino acid sequences, they were proposed to be equal to the nucleotide substitution rate under the neutral selection. However, several lines of evidence suggest that the synonymous changes have not, in fact, been selectively neutral in the course evolution.[35] It has been reported that the following several factors would affect the synonymous nucleotide substitution rates. Although many synonymous mutations are no doubt free from the selection, the assumption that they are all neutral no longer appears tenable. It is important to keep in mind these effects when evaluating the selective pressure.

Base Composition and the Nucleotide Substitution Rate

The correlation between the synonymous substitution rate (Ks) and the GC content of the genome has been the subject of debate.[36, 37] It is likely that the GC content would affect the nucleotide substitution rates, because the GC content might affect the DNA structure. Therefore, we evaluated the correlation of the GC content with the synonymous and nonsynonymous nucleotide substitution rates by using sixteen TLR-related genes. As shown in Figure 21.4, the synonymous substitution rate, but not the nonsynonymous substitution rate, weakly correlates with the GC content. The genes with higher levels of GC content appeared to be linked to the higher levels of synonymous substitution rates.



Figs 21.4A to D: Correlation analyses of comparative genomics parameters. Correlations between Σbs and the GC comtent (A), Σbs and CBI (B), Σbn and the GC comtent (C), and Σbn and CBI (D) in sixteen TLR-related genes are shown. All values of Σbs, Σbn, the GC content, and CBI were evaluated for the entire coding sequences for the TLR-related gene

Nonrandom Codon Usage and the Nucleotide Substitution Rate

Table 21.4 shows the frequencies of codon usage in human genes, based on 40, 662, 582 codons from 93, 487 coding sequences (Codon usage data base. See http://www.kazusa.or.jp/codon/). If the synonymous changes were completely neutral in the course of evolution, the synonymous codons for an amino acid should have equal frequencies. However, the pattern of codon usage for synonymous codon was apparently nonrandom in the human genome. For example, in the comparisons of the frequencies of six codon usages for leucine, a large bias was observed in their usage (The amino acid, leucine, was most frequently coded by CUG and less often by CUA), suggesting that the pattern of synonymous codons was distinctly nonrandom.

There are several methods to evaluate the nonrandom usage of synonymous codons. We applied the codon bias index (CBI) as a measure of the deviation from the equal use of synonymous codons to investigate the TLR-related genes. Bennetzen and Hall[38] arbitrarily chose twenty-two codons encoding for seventeen amino acids as preferred codons for species of interest. Based on these preferred codons, the value of CBI was defined as:

$$CBI = (N_{pfr} - N_{ran}) / (N_{tot} - N_{ran})$$

where N_{pfr} is the total number of occurrences of the preferred codons, N_{ran} is the expected number of the preferred codons if all synonymous codons were used equally, and N_{tot} is the total number of the seventeen amino acids encoded by the preferred codons. The CBI values range from 0 (uniform use of synonymous codons) to 1 (maximum codon bias).

Figure 21.4 shows the correlation of nucleotide substitution rates using the CBI for sixteen TLR-related genes, that is, a correlation of CBI with the synonymous substitution rates, but not the nonsynonymous substitution rates. The departure of codon usage from random should be associated with the higher levels of synonymous substitution rates. We do not know the exact reason why the CBI was linked to the synonymous substitution rate in TLR-related genes. Since the degree of nonrandom codon usage has been reported to be linked to the levels of gene expression,[39-41] expression levels of the TLR-related genes might be linked to the synonymous substitution rates. An alternative possible explanation is that the GC content might affect the synonymous substitution rate, because the CBI has a positive correlation with the GC content in the TLR-related genes (data not shown).

Table 21.4: The frequencies of codon usage in human genes based on 40,662,582 codons from 93,487 coding sequences

1st base	*2nd base* U		C		A		G	
U	UUU (Phe)	17.6	UCU (Ser)	15.2	UAU (Tyr)	12.2	UGU (Cys)	10.6
	UUC (Phe)	20.3	UCC (Ser)	17.7	UAC (Tyr)	15.3	UGC (Cys)	12.6
	UUA (Leu)	7.7	UCA (Ser)	12.2	UAA (Stop)	1	UGA (Stop)	1.6
	UUG (Leu)	12.9	UCG (Ser)	4.4	UAG (Stop)	0.8	UGG (Trp)	13.2
C	CUU (Leu)	13.2	CCU (Pro)	17.5	CAU (His)	10.9	CGU (Arg)	4.5
	CUC (Leu)	19.6	CCC (Pro)	19.8	CAC (His)	15.1	CGC (Arg)	10.4
	CUA (Leu)	7.2	CCA (Pro)	16.9	CAA (Gln)	12.3	CGA (Arg)	6.2
	CUG (Leu)	39.6	CCG (Pro)	6.9	CAG (Gln)	34.2	CGG (Arg)	11.4
A	AUU (Ile)	16	ACU (Thr)	13.1	AAU (Asn)	17	AGU (Ser)	12.1
	AUC (Ile)	20.8	ACC (Thr)	18.9	AAC (Asn)	19.1	AGC (Ser)	19.5
	AUA (Ile)	7.5	ACA (Thr)	15.1	AAA (Lys)	24.4	AGA (Arg)	12.2
	AUG (Met)	22	ACG (Thr)	6.1	AAG (Lys)	31.9	AGG (Arg)	12
G	GUU (Val)	11	GCU (Ala)	18.4	GAU (Asp)	21.8	GGU (Gly)	10.8
	GUC (Val)	14.5	GCC (Ala)	27.7	GAC (Asp)	25.1	GGC (Gly)	22.2
	GUA (Val)	7.1	GCA (Ala)	15.8	GAA (Glu)	29	GGA (Gly)	16.5
	GUG (Val)	28.1	GCG (Ala)	7.4	GAG (Glu)	39.6	GGG (Gly)	16.5

The frequencies are shown per thousand.

Synonymous Nucleotide Substitutions Having Functional Impact

It is widely accepted that a part of synonymous substitutions would have a functional impact on the genome through their effects on the mRNA stability, translational efficiency, and/or splicing control. It is easily supposed that the synonymous substitutions with these functional affects have not taken place under the control of neutral evolution.

It was reported that some synonymous substitutions could affect the mRNA secondary structure and mRNA stability,[42] which were occasionally associated with the disease susceptibility. Human dopamine receptor D2 gene (*DRD2*) is a well known example of how a synonymous mutation can affect the mRNA stability.[43] Among six synonymous variants in *DRD2*, only the mutation that decreased the mRNA half-life and induced a conspicuous change in the predicted secondary structure was linked to the disease. For another example, synonymous variants in the gene for Catechol-O-methyltransferase (*COMT*), an enzyme responsible for degrading catecholamines, are associated with the enzymatic activity.[44] *COMT* is a key regulator of pain perception, cognitive function, and affective mood. Several haplotypes divergent in synonymous changes exhibited the largest difference in COMT enzymatic activity, due to a reduced amount of translated protein. These synonymous variants affect the mRNA stability due to the resulting difference in the local stem-loop structures of mRNA. The most stable structure was associated with the lowest protein level and enzymatic activity.

Synonymous substitutions could also affect the translational efficiency through the use of codons with rare tRNA.[39-41] Alternative codons specifying particular amino acids can differ in translational efficiency, in part due to the relative abundance of tRNA and the use of codons that are specified by rare tRNA might be linked to slow translation. For example, *MDR1* encodes an ATP-binding cassette transporter, P-gp, which contributes the pharmacokinetics of drugs with altered transport. A synonymous variant in exon 26 was found to be associated with the P-gp activity.[45] This synonymous variant in *MDR1* affected the timing of co-translational folding due to a slow translation resulting form rare codon usage, which might result in the altered function of P-gp.

The evidence for the synonymous mutations leading to human disease by disrupting the splicing process is abundant.[35] These synonymous mutations might create new cryptic splice sites or change splicing-control elements, such as exonic splicing enhancers (ESEs) and silencers (ESSs), which are oligomeric motifs that recruit splicesomal proteins to facilitate splice-site recognition. Alternatively, splicing-control elements were reported to have come under the control of negative selection, because the disruption of splicing-control elements induced by nucleotide substitutions would be deleterious for the gene function.

COMPARATIVE GENOMICS IN NON-CODING REGIONS

The comparison of non-coding genomic sequences is also useful for identifying the genomic regions for regulatory elements, such as components of promoters and enhancers.[1, 11-13] We briefly touch on such comparisons in this section. This approach is based on the idea that functionally important regions should be conserved in the course of evolution. In fact, it has been reported that putative-transcription factor binding sites are enriched in the evolutionary conserved non-coding genomic sequences.[46-48] Furthermore, experimentally determined regulatory elements are indeed enriched in the evolutionary conserved genomic segments. For the prediction of regulatory-regions, comparisons among relatively distant species appears to be more effective than those among closely related species, because the higher similarity in the non-coding sequences can be found among the closely related species, but presumably hard to be observed in the distantly related ones if there were no evolutionary pressures to maintain the sequences.

Two main approaches have been broadly available in the discovery of regulatory regions in the non-coding regions. The first approach is based on the global alignment of orthologous sequences, followed by the identification of conserved regions. Figure 21.5 shows the comparisons of 20-kb genomic sequences encoding human TLR4 gene with its orthologues from six vertebrate species, including mouse, rat, chicken, dog, rhesus, and horse, which were aligned with the computational tool, VISTA.[49] Highly conserved non-coding genomic segments observed between two distant species, such as the human and mouse, are potential candidate regions for regulatory elements.

The second approach endeavors to find multiple orthologous sequences from related species, followed by the evaluation of their functional significance.

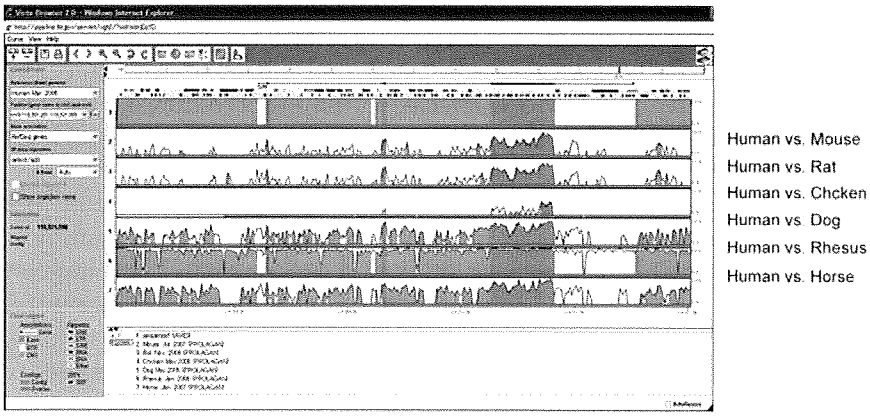


Fig. 21.5: Comparative genomic analysis using a computational tool, VISTA.[49] The 20-kb genomic sequences encoding TLR4 gene among seven vertebrate species including human, mouse, rat, chicken, dog, rhesus, and horse were analyzed

For instance, we tried to identify and compare orthologous sequences of the human microRNA (miRNA), miR759, from other vertebrate species. miRNA is a class of short, non-coding RNAs that post-transcriptionally regulate gene expression by interaction with partially complementary target sites on mRNAs. It has been reported that mammalian miRNAs may regulate the expression of ~30 percent of the protein-coding genes.[50] We identified homologous sequences of miR759 from eight vertebral species. Figure 21.6 shows the sequence alignment among human, rhesus, mouse, rabbit, rat, cattle, and dog. miR759 sequences was found to be strictly conserved in the course of vertebrate evolution, suggesting the biological significance of miR759 in the control of gene expression. The human miRNA miR759 regulates the expression level of human fibrinogen-alpha gene through its binding to the 5'-untransrated region of the human fibrinogen-alpha gene (Chen et al unpublished data).

APPLICATION OF COMPARATIVE GENOMICS TO AN IDENTIFICATION OF HUMAN DISEASE GENES

As discussed in the previous sections, comparative genomics is a promising approach for identifying the genes that may control the susceptibility to human diseases, especially in combination with comparative analyses of phenotype, such as differences in disease susceptibility, among primates. It has been reported that the susceptibility to various diseases differs among primates.[21] For example, Asian Old World monkeys are highly susceptible to infection with *M. tuberculosis bacilli*, while New World monkeys are reported to be more resistant.[51] Furthermore, as described previously, the species-specific restrictions operating on HIV-1 infection are well-known (4; 52; 5). Humans, as well as chimpanzees, but not New and Old World monkeys, are susceptible to HIV-1 infection. Such differences in the species-

```
                    ***.....**************************************************
Mature miR-759      1:                                    GCAGAGUGCAAACAAUUUUGAC 22
human               1: TTATGGAAAGGATTATAATAAATTAAATGCCTAAACTGGCAGAGTGCAAACAATTTTGAC 60
rhesus              1: TTATAGAAAGGATTATAATAAATTAAATGCCTAAACTGGCAGAGTGCAAACAATTTTGAC 60
mouse               1: TTATGGGCAGGATTATAATAAATTAAATGCCTAAACTGGCAGAGTGCAAACAATTTTGAC 60
rat                 1: TTATGGGAAGGATTATAATAAATTAAATGCCTAAACTGGCAGAGTGCAAACAATTTTGAC 60
rabbit              1: TTATGTAAAGGATTATAATAAATTAAATGCCTAAACTGGCAGAGTGCAAACAATTTTGAC 60
cattle              1: TTACGGGAAGGATTATAATAAATTAAATGCCTAAACTGGCAGAGTGCAAACAATTTTGAC 60
dog                 1: TTATGGGAAGGATTATAATAAATTAAATGCCTAAACTGGCAGAGTGCAAACAATTTTGAC 60

                    **********************************************..******.**....
human              61: TCAGATCTAAATGTTTGCACTGGCTGTTTAAACATTTAATTTGTTAGAATGGAAGTAGCG 120
rhesus             61: TCAGATCTAAATGTTTGCACTGGCTGTTTAAACATTTAATTTGTTAGAATGGAAGTAGCG 120
mouse              61: TCAGATCTAAATGTTTGCACTGGCTGTTTAAACATTTAATTTGTTCCAATGGAGGTAGCA 120
rat                61: TCAGATCTAAATGTTTGCACTGGCTGTTTAAACATTTAATTTGTTCGAATGGAGGTAGAG 120
rabbit             61: TCAGATCTAAATGTTTGCACTGGCTGTTTAAACATTTAATTTGTTAGAATGGAAGTAGCA 120
cattle             61: TCAGATCTAAATGTTTGCACTGGCTGTTTAAACATTTAATTTGTTAGAATGGAGGTTCCA 120
dog                61: TCAGATCTAAATGTTTGCACTGGCTGTTTAAACATTTAATTTGTTAGAATGGAGGTAGCA 120
```

Fig. 21.6: Sequence alignment of miR-759-like sequences in human, rhesus, mouse, rabbit, rat, cattle, and dog. The pri-miR-759 sequences are underlined. The mature miR-759 sequence is 22 bp long, as indicated on the top

specific susceptibility to infectious disease might be associated with the differences in gene functions linked to the defense against infectious diseases, which would lead the differences in defense mechanisms mounted against invading pathogens. Natural selection pressure has advanced the species-specific evolution in the susceptibility genes for infectious diseases. In fact, dozens of genes under the control of natural selection in the course of primate evolution have been identified to date. Taking all things into consideration, these genes are candidates for determining the susceptibility to human diseases. In the following sections, we describe the examples.

Comparative Genomics Might Uncover Susceptibility Genes for Human Diseases

TRIM5α plays crucial roles in the intracellular defense against HIV-1,[52] and sequence differences in the SPRY domain of TRIM5α contribute to the differences in anti-HIV-1 activity among primate species.[52] Comparative genomics for TRIM5α shows that this gene has rapidly evolved in the course of primate evolution, and that natural selection has shaped the sequence difference in the SPRY domain. Sequence variations in TRIM5 have been reported to be associated with the susceptibility to HIV-1/AIDS in humans. For another example, APOBEC3G, that also plays an important role in the defense against HIV-1, has been under positive Darwinian selection.[53] A comparison of *APOBEC3G* sequences among primates suggests a rapid evolution of APOBEC3G in the course of primate evolution.

Given that *TRIM5α* and *APOBEC3G* are the genes associated with the susceptibility to HIV/AIDS in humans, comparative genomics can be a useful tool for both identifying the candidate genes for controlling the HIV/AIDS susceptibility and evaluating the pathophysiological roles of the genes from an evolutionary point of view. These examples suggest that comparative genomics is a promising approach to identify the susceptibility genes for infectious diseases. Since susceptibilities differ among primates to not only infectious, but also other common diseases, such as Alzheimer's disease and cardiovascular diseases, comparative genomics may be a crucial tool in providing candidate genes to determine the susceptibility for the other diseases as well.

Bioinformatics Tool to Evaluate Functional Impact of Nonsynonymous Variations based on Comparative Genomics

It is estimated that there are 67,000-200,000 common nonsynonymous single nucleotide polymorphisms (SNPs) in the human genome, and that each individual is heterozygous for 24,000-40,000 nonsynonymous SNPs.[54] Because nonsynonymous SNPs would affect the protein function, some of them might be associated with human health and disease. Recently, various computational approaches to assessing the functional significance of

nonsynonymous SNPs have been developed.[55-61] They predict whether an amino acid substitution induced by nonsynonymous SNPs affects protein function based on the comparative genomics, physical properties of amino acids, and/or three-dimensional (3D) structures of proteins. The main features of the representative methods are summarized in Table 21.5. These programs are also useful to estimate the significance of functional impact induced by nonsynonymous mutations in the genes for single-gene disorders.

Some of the programs are sequence-based amino acid substitution prediction method based on the comparative genomics, which are founded on the concept that the amino acid substitutions affecting protein function tend to occur at conserved evolutionary sites. Such conserved sites, as described in the previous sections, have come under the control of negative (purifying) selection, and are considered to be important for protein function. A multiple sequence alignment among the homologous sequences indicates the conserved sites throughout the course of evolution. The sequence-based amino acid substitution prediction method scores the levels of amino acid substitution based on the amino acids appearing in the multiple alignments, and the severity of the amino acid change based on the physical properties. It has been reported that 25-35 percent of nonsynonymous SNPs are predicted to affect the protein function by the

Table 21.5: The main features of representative prediction methods used to evaluate the impact of nonsynonymous SNPs on protein function

Method	*Web site*	
SIFT [55]	http://sift.jcvi.org/	Sequence-based prediction method using position-specific scoring matrices with Dirichlet priors
PolyPhen[56]	http://coot.embl.de/PolyPhen/	Structure/sequence-based prediction method
SNPs3D[57]	http://www.snps3d.org/	Structure/sequence-based prediction method
PANTHER PSEC[58]	http://www.pantherdb.org/tools/csnpScoreForm.jsp	Sequence-based prediction method using PANTHER Hidden Markov Model families
PMUT [59]	http://mmb2.pcb.ub.es:8080/PMut/	Structure/sequence-based prediction method
TopoSNP[60]	http://gila.bioengr.uic.edu/snp/toposnp/	Structure/sequence-based prediction method A database of topographic mapping of SNPs
MAPP[61]	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Sequence-based prediction method

most widely used prediction methods. Currently, automated prediction methods are being applied on a genome-wide scale, which might accelerate the findings of human disease susceptibility genes in the near future. However, it has been pointed out that there are limitations to the current prediction methods. For example, Thomas et al[62] have reported that the current method may not be useful for identifying certain nonsynonymous SNPs involved in human common diseases. In any events, however, it is reasonable to expect further progress in the field of bioinformatics using these and related methods.

CONCLUSION

In this chapter, we introduced concepts and methods for evaluating the biological significance of homologues sequences, especially focusing on the methods that are based on the theory of natural selection. Although the comparative genomics are not definitive in determining the biological significance of conserved sequences, this approach is nonetheless highly useful as the first step for identifying and characterizing functional regions in the genome. We have introduced here only a small part of comparative genomics. We recommend referring to a number of reviews to cover the wide variety of the features of comparative genomics.[1, 11-13]

Recent rapid progress in the field of bioinformatics and sequencing technology has brought about a breakthrough in the comparative genomic analysis. It is therefore expected that further progress in the comparative genomics will provide a stream of novel insight into health and disease in humans.

ACKNOWLEDGMENTS

This work was supported in part research grants from the Ministry of Health, Labor and Welfare, Japan, the Japan Health Science Foundation, the program of Founding Research Centers for Emerging and Reemerging Infection Disease, the Japan Health Science Foundation, the program of Research on Publicly Essential Drugs and Medical Devices, Grant-in-Aids for Scientific research from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan, grants for Indo-Japan collaboration research from DST and JSPS, and a grant from Heiwa Nakajima Foundation.

REFERENCES

1. Margulies EH, Birney E. Approaches to comparative sequence analysis: Towards a functional view of vertebrate genomes. Nat Rev Genet 2008;9:303-13.
2. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: Methods and goals. Nat Rev Genet 2004;5:335-44.
3. Vogel F, Motulsky AG. Human Genetics problems and approaches 3rd ed. Springer 1996.