Front. Phys. China, 2009, 4(2): 241–251
DOI 10.1007/s11467-009-0014-x

VIEWPOINT

Law of genome evolution direction: Coding information quantity grows

Liao-fu LUO (罗辽复)

Laboratory of Theoretical Biophysics, Faculty of Science and Technology, Inner Mongolia University, Hohhot 010021, *China*
E-mail: lolfcm@mail.imu.edu.cn

Received December 9, 2008; accepted January 17, 2009

The problem of the directionality of genome evolution is studied. Based on the analysis of C-value paradox and the evolution of genome size, we propose that the function-coding information quantity of a genome always grows in the course of evolution through sequence duplication, expansion of code, and gene transfer from outside. The function-coding information quantity of a genome consists of two parts, p-coding information quantity that encodes functional protein and n-coding information quantity that encodes other functional elements. The evidences on the law of the evolutionary directionality are indicated. The needs of function are the motive force for the expansion of coding information quantity, and the information quantity expansion is the way to make functional innovation and extension for a species. Therefore, the increase of coding information quantity of a genome is a measure of the acquired new function, and it determines the directionality of genome evolution.

Keywords genome evolution, function-coding information quantity growing, p-coding information quantity, n-coding information quantity, C-value paradox

PACS numbers 87.10.+e, 87.23.-B, 87.23.kg, 05.90.+m

The extension of the diversity of species and their evolution toward higher function more adaptive to environment shows that the life evolution obeys a law with definite direction. Darwin expressed the law as "survival of the fittest." It means that the "designed" properties of living things better adapted to survive will leave more offspring and automatically increase in frequency from one generation to the next, while the poorly adapted species will decrease in frequency. Accompanying the development of molecular biology and with ever increasing understanding on genomes, we are able to express the law on life evolution more quantitatively and precisely. Here, the key point is the introduction of information. Life consists of matter and energy, but it is not just matter and energy. The life of an individual comes from the DNA of its parents. DNA, which weighs only 10^{-12} gram for human, is insignificant in terms of matter because, like many other things on earth, it is composed of nitrogen, oxygen, sulfur, etc. In addition, DNA, as a source of energy, is also unimportant, since it is just composed of the similar level of chemical energy as other macromolecules that can be produced by experiment. However, different from matter and energy, information constitutes the third fundamental category in natural sciences. Schrödinger [1] was the first who recognized the importance of information and indicated that the characteristic feature of life that differentiates from an inanimate piece of matter is the large amount of information contained in its chromosomes. He said, "We believe a gene – or perhaps the whole chromosome fiber – to be an aperiodic solid. With the molecular picture of genes, it is no longer inconceivable that the miniature code should precisely correspond with a highly complicated and specified plan of development and should somehow contain the means to put it into operation." Sixty years have passed; but now, as we try to formulate the law on the directionality of life evolution, we should put our discussion based again on the concept of information. Thanks to the discovery of cell totipotent, despite the complexity of multicellular organism the full genetic information of an organism is stored in chromosomes of just one cell. We can express the basic law on species evolution through the genomic information contained in chromosomes of one cell. Many studies were carried out about the evolution of genome size [2]. The size of genomes for sibling species can change

more than several tenfold, for example, 340-fold for flatworms, 70-fold for nematodes, 170-fold for arthropods (insecta), 350-fold for fish, 130-fold for amphibians,196-fold for algae (chlorophyta), 500-fold for pterridophytes, and 1000-fold for angiosperms, etc. Moreover, the evolutionary complexity of a species is irrespective of its genome size. For example, the C-values of lungfishes are higher than human about 8 to 20 times. Some salamanders can also have large genomes with C-values 15 times or more than human. Next, about the genome size evolution, it was speculated that plants might have a "one-way ticket to genomic obesity" through amplification of retrotransposons and polyploidy. The similar assumptions were also made that the animal genome sizes might change in the direction of increase. However, there has been considerable evidence that both increases and decreases may occur in plant and animal lineages. In the meantime, no fossil evidence on the genome size variability in the single direction was reported [3]. These have made the issue on the directionality of genome size more complex and interesting. However, from our point of view the genome size of a species is not a proper measure of evolutionary directionality since it is irrespective of genetic information necessary for encoding the biological function. Instead, when our investigation is based on the function-coding information, we will be able to obtain an unambiguous picture on the evolutionary law of genomes.

1 The law of genome function-coding information quantity growing

Definition. For an n-long sequence (called sequence A) written by symbols A_1, A_2,..., and A_n, where A_i taking $s_i$possible values (i=1,...,n), if the sequence A encodes certain function, then we define the function-coding information quantity of the sequence:

$$I_C = \log_2 \prod_i s_i \tag{1}$$

For prokaryote genome, the symbol A_i takes four values A, G, C, or T. For eukaryotes, the chromatin remoulding and histone modification are other type of variables that has the potential to influence fundamental biological processes and may be epigenetically inherited. More than 30 histone modifications have been found for human and other vertebrates [4]. They form the source of heredity information in addition to four kinds of bases A, G, C, and T. Moreover, the DNA methylation (mainly the cytosine methylation for higher vertebrates) can inhibit gene expression and be epigenetically inherited, and therefore, the methylated bases give additional symbols applicable in genetic language [5].

Following the said definition, a function-coding DNA sequence segment (or its extension that includes the chromatin remoulding variable, methylated cytosine, etc.) consists of two basic types: p-coding sequence that encodes functional protein and n-coding sequence, which encodes all other functional sequence segments,

$$\begin{aligned} I_C =& (\text{p-coding information quantity}) \\ &+ (\text{n-coding information quantity}) \end{aligned} \tag{2}$$

If histone modifications and epigenetic inheritances are neglected, then the four-symbol DNA sequence is the only source of genetic information. In this case, p-coding information quantity equals twice of the length of protein-coding sequence (or CDS in common genome terminology),

$$\text{p-coding information quantity} = 2\times(\text{CDS length}) \tag{3}$$

and the n-coding information quantity includes the contribution from promoters and other transcriptional regulatory elements upstream of the transcriptional start site (TSS) and the contribution from non-protein-coding RNAs (ncRNAs) in genes and intergenic sequences in primary transcripts,

$$\begin{aligned} &\text{n-coding information quantity} \\ &= 2\times(\text{length of promoters and other transcriptional} \\ &\quad \text{regulatory elements upstream of TSS} \\ &\quad + \text{length of ncRNAs in primary transcripts}) \end{aligned} \tag{4}$$

(other transcriptional regulatory elements include enhancers, silencers, insulators, etc.) Evidently, from Eq. (2) to Eq. (4), the function-coding information quantity I_C for most genomes is smaller than twice of the genome size due to the existence of a nonfunctional part (for example, pseudogenes) in the DNA sequence. How the promoters and ncRNAs contribute to the genome function will be discussed later in Section 2.

Law of function-coding information quantity growing (CIQG)

Due to the interaction among DNA, RNA, and protein under the possible influence of chromatin remoulding, chemical modification, and other factors, the function-coding information quantity I_C of a genome sequence always grows in the course of evolution

$$\frac{\mathrm{d}I_C}{\mathrm{d}t} \geqslant 0 \tag{5}$$

through sequence duplication, expansion of code, and horizontal gene transfer.

The evolutionary law was first proposed in references [6, 7] as an assumption. Before its demonstration, we shall give some comments on the expression of the law:

(1) Entropy increase is a universal law of nature. Due to the randomness of stochastic movement, the entropy of any isolated physical system always increases. However, what stated here is not identical with the physical law of entropy growing. The law of CIQG refers to the evolution of function-coding information quantity of genome. Here, the mechanism responsible for the law

is functional selection. As a biological law of evolution, there may exist exception; for example, some genome evolves in strange and peculiar environment. However, the exception should be very rare. On the other hand, the time scale $\mathrm{d}t$ in $\frac{\mathrm{d}I_C}{\mathrm{d}t} \geqslant 0$ is determined by the minimal time interval required for natural selection acting on heredity process. For a given species in an environment which is stable in the time duration longer than $\mathrm{d}t$, $\mathrm{d}t$ = several generations of the species, the law $\frac{\mathrm{d}I_C}{\mathrm{d}t} \geqslant 0$ will manifest itself as the clock with accuracy of several generations is used. The equality occurs only for those old species without any evolution. For most species, $\frac{\mathrm{d}I_C}{\mathrm{d}t} > 0$, so $I_C \frac{\mathrm{d}I_C}{\mathrm{d}t} > 0$. This means I_C can be looked as a Lyapounov function and the evolution of species is generally Lyapounov unstable.

(2) Evolutionary law is closely related to the environment. Recently, the evidences for environmental change as a determinant of mass extinction and its biological selectivity in the fossil record were indicated [8]. For a genome in stable environment, I_C grows with time.For a genome in varying environment, I_C grows adapting to the change of environment. The speed of coding information quantity growing can be viewed as a scale of evolutionary rate of species. However, for a genome in a suddenly changing environment, if the speed of growing I_C cannot adapt to the sudden change of environment (for example, the food deficiency), then the species would be close to extinction; or new species would otherwise emerge, wherein the genes of which can adapt to the functional needs under new environment. From fossil records, we know that evolution has a range of rates, from sudden to smooth, and both punctuated equilibrium and phyletic gradualism have occurred [9]. The punctuated equilibrium may be somewhat commoner than phyletic gradualism, and in new species formation, the evolution always shows a sudden rate. Whatever the change is in sudden or smooth phase, the law of CIQG holds universally.

(3) The loss of function in parasitism of some bacteria resulting in the decrease of coding information quantity of these genomes is a phenomenon of retrogression, which should not be included in the scope of the law. The law holds only for free-living genomes. The environment inside the host cell contains many of the nutrients and defense systems that a bacterial cell needed, since genes that are needed in a free-living bacterium to provide the resources are not needed in an intracellular bacterium. In this case, the gene loss may be advantageous since a cell with less DNA can reproduce faster.

(4) Recently, Taft, Pheasant, and Mattick [10] proposed that the relative amount of non-protein-coding sequences increases with genome complexity. The point is different from ours in two aspects: (i) The non-protein-coding sequences include n-coding sequences as in our proposal, but they also include sequences that do not encode any function at all. The latter part sequences should not contribute to the evolutionary complexity of a genome. (ii) The evolution of species is always hierarchical and tree-like. The quantitative comparison of the complexity between organisms in different clades is generally difficult and even confusing in many cases. However, we abandon the undefined concept of complexity comparison but express the evolutionary law as the change of a well-defined quantity, the function-coding information quantity I_C, with time t along any clade of the tree. The statement is clearer in logic and will be checked easily by future experiments.

2 Supporting facts in demonstration of CIQG law

(1) Is the law of function-coding information quantity growing for a genome consistent with experimental data? Biologists know that the evolutionary complexity of a species is irrespective to its genome size. However, as plotting the range of genome size in different evolutionary phyla, we found that there is an increase in the minimum genome size in each group as the complexity increases [5, 11]. This is due to the quantity of proteins in different species of a group that are basically the same. So the size of the minimum genome reflects the p-coding information quantity, while the p-coding information quantity is correlated well to the total function-coding information quantity from prokaryotes to lower eukaryotes.

(2) The prokaryote genome size changes from 1 Mb to 10 Mb for free-living bacteria. The variation is relatively smaller than for eukaryotes. Because there is so little noncoding DNA in prokaryotes, the mechanism for generating the variation must largely involve changes in gene number. Both gene number increase and reduction over the course of prokaryotes evolution are observed [12]. However, the gene loss has happened primarily among parasites and symbionts. In the meantime, the shift of a genome from free-living to host-dependent lifestyle renders many genes obsolete. They are no longer maintained by selection but by deactivated by mutation, and then, they are subsequently lost. The process may be long, and it leads to pseudogenes remaining intact in some genomes [13]. Since DNA deletion is mainly related to parasites and pseudogenes, the free-living bacterium always shows evolutionary directionality toward the increase of function-coding information quantity. The consistency of CIQG law with data of prokaryote genome evolution can be proved by the following observation.

Suppose the phylogenetic relations among prokaryotes have been known, and a tree of life for free-living bacteria has been reconstructed. The terminal nodes of the tree represent the present species, and the internal nodes represent the past-time genomes. The coding information quantity of a present bacterium can be deduced directly by calculating the functional gene number in the genome, which equals the total number of genes minus the number of pseudogenes. While for ancestor species, it should be obtained based on some assumptions. Since only the free-living bacteria are concerned, as a working hypothesis, we suppose that the genome size of an ancestor species takes the minimal value of sizes of its first descendents in the clade. Thus, we are able to obtain a self-consistent solution for the coding information quantity of all free-living bacteria. We find the solution satisfying the law of coding information quantity increasing in evolution. Note that due to functional diversity, some genomes on the tree, for example, *Bradyrhizobium*, *Streptomyces*, etc., possess sizes much higher than their sibling branches. However, regardless of these large differences in genome size, we have shown the CIQG law consistent with prokaryotic genome data.

(3) For eukaryotes, the DNA loss was also frequently observed in genome evolution. However, the lost DNA is generally related to the nonfunction region. There has been a proposition that at the level of small (<400 bp) insertions and deletions (indels), the deletions in a genome are more frequently occurred than the insertions. Petrov *et al.* [14] studied the indel spectrum in *Laupala* crickets and in *Drosophila*. The former has a genome size 11 times larger than latter. They found that the DNA loss in *Laupala* is more than 40 times slower than in *Drosophila* and indicated the possible inverse correlation between genome size and DNA loss rate. However, the DNA losses discussed by these authors are mainly related to pseudogenes that lack the capacity to encode functional proteins. Therefore, they successfully explained that the high rate of DNA loss in small genomes results in a lower steady state number of pseudogenes. The result does not mean that the loss of functional DNA has happened in some eukaryotic genomes. Generally speaking, gene duplication and loss is a powerful source of functional innovation. Wapinski *et al.* [15] studied the evolutionary principles of gene duplication in fungal genomes through the determination of orthology and paralogy relations across 17 species and demonstrated that gene duplication and loss is highly constrained by the functional properties and interacting partners of genes. They indicated that the duplicated genes typically diverge with respect to regulatory control, and therefore, the gene duplication may drive the modularization of functional networks through specialization. This means that the trend of the function-coding information quantity increases in evolution.

(4) An important approach of new species formation is the whole genome duplication (WGD) that leads to several new species by different ways of gene-loss after the duplication. For example, Scannell and Byrne *et al.* [16] studied reciprocal gene loss in polyploidy yeasts. A whole-genome duplication occurred in a shared ancestor of yeast species *S. cerevisiae*, *S. castellii*, and *C. glabrata*. These authors traced the subsequent losses of duplicated genes and showed that the pattern of loss differs among three species at 20% of all loci. Although three species lose genes, respectively, in their divergence, the total amounts of genes increase due to genome duplication, the increment 14%for *S. cerevisiae*, 16%for *S. castellii*, and 11%for *C. glabrata*. The pattern of reciprocal gene loss demonstrated the mechanism of reproductive isolation, and the further analysis indicated the rapid divergence of three species shortly after the whole-genome duplication. Aury *et al.* [17] studied the whole-genome duplications in ciliate. These authors reported that most of the nearly 40 000 genes of the unicellular eukaryote *P. tetraurelia* arose through at least three successive whole-genome duplications, and the most recent duplication gave rise to the *P. aurelia* complex of 15 sibling species. They observed that the gene loss occurs over a long timescale, and the temporary maintenance of many duplicated genes is simply due to dosage constraints. Then, they estimated the gene number of the species about 20 000 in the initial phase, from 20 000 to 40 000 in the first WGD, from 21 000 to 42 000 in second WGD, from 26 000 to 52 000 in third WGD, and 39 642 after the third genome duplication and speciation. The series of gene number 20 000, 21 000, 26 000, and 39 642 approximately reflect the increase trend of functional gene. Therefore, in spite of a large amount of gene loss in WGD that was observed in the above two examples, the total number of functional genes in a genome still increases through many gene duplications. This is consistent with the law of coding information quantity growing in evolution. In fact, the gene duplication is associated with the acquisition of some new gene function. The new function associated with gene duplication can arise in different ways, for example, neofunctionalization and escape from adaptive conflict etc. [18]. Here, the essential point behind apparently different approaches is the increase of coding information quantity, which can serve as a measure of the acquired new function for a genome.

(5) It was estimated that 70% of all angiosperms had experienced one or more episodes of polyploidy in their ancestry. Intuitively, the polyploids should have larger C-values than diploids with its C-value increasing in direct proportion to ploidal level. However, the loss of DNA following polyploid formation may be a widespread

phenomenon [19]. There are three potential outcomes for the plant genes duplicated by polyploidy: first, both copies remain functional; second, one copy becomes silences or lost, while the other retains the original function; or third, the two copies may diverge in function. The second possibility is that the gene loss can occur rapidly. For example, recent studies have demonstrated the polyploidy-induced rapid and reproducible elimination of DNA sequences in the wheat group. However, the eliminated sequences were not homologous to known genes and likely represented noncoding regions [20].

(6) In the Cambrian explosion of animals (metazoan expansion) and in the fast evolution from reptiles to birds and to mammals, one can find more evidences on CIQG law from these adaptive radiation events. An adaptive radiation means that a small number of ancestral species in one taxon diversifies into a large number of descendant species, occupying a broader range of ecological niches. Darwin was interested in why evolution usually shows a diverging tree-like pattern. He explained the pattern by competition. More similar forms will compete more strongly, which tends to push species apart during evolution. The proliferation of animals with hard skeletons in the Cambrian explosion may be explained by predators evolving escalated skills around this time. The new function needs the increase of information quantity to encode it. The animals with hard parts may have higher coding information quantity than their soft-bodied predecessors [9]. The origin of mammals can be traced back before 85–100 million years ago, through a series of small changes in mammal-like reptiles [21]. The reported reptilian C-values vary from 1.1 pg to 5.4 pg, while the mammalian C-values range from 1.7 pg to 8.4 pg [3]. Considering the C-values measured only for a small fraction of modern species and the extinction of many evolutionary predecessors, the C-value data given above is far from complete. If the coding information quantity can be estimated by the lower bound of the observed C-values in an appropriate group of species, then the evolution from reptiles to mammals seems not to be inconsistent with CIQG law. The origin of bird flight is another important event in vertebrate evolution. The reported avian C-values range from 1.0 pg to 2.2 pg for 2% of the studied species [3]. The minimal C-values observed in birds are close to those of reptiles. Adapted to the flight condition, the birds should have a relatively small genome. Recently, some fossil evidences showed that the birds evolve from nonavian dinosaurs, which is the saurischian dinosaur lineage [22]. The latter has less repetitive DNA than other typical ancestral Dinosauria. These genomic characteristics of fewer repetitive elements and less noncoding DNA should be added to the list of attributes previously considered avian but now thought to have arisen in nonavian dinosaurs, such as feathers, pulmonary innovations, and parental care and nesting [23]. They exhibited a structure of preadaptation that happened to evolve a new function of flight. Although we do not yet know the details of genome variation about the acquisition of avian flight function, the fast evolution from reptiles to birds should not be in conflict with CIQG law also.

(7) Both increase and decrease of genome size occur in plant and animal lineages. The DNA amount of a genome reflects the dynamic balance between the opposing forces of expansion and contraction. However, the steady state changes adiabatically (more slowly than the force acting on genome) toward the higher functional state through selection. We shall investigate the contraction forces in more detail. The mechanism of sequence reduction induced by contraction forces include mainly the sequence recombination by unequal cross-over during meiosis and unequal sister chromatid exchange during mitosis. The process of unequal intrastrand homologous recombination occurs between the long terminal repeats of LTR-retrotransposons and can lead to the deletion of the internal DNA segment and one LTR, leaving behind only one 'solo LTR'. In this case, there is still a net gain of DNA with each insertion of transposable element [24]. On the other hand, the illegitimate recombination, irrespective of homologous sequences, can act on a larger fraction of the genome. It may involve the deletion of all intervening sequences between the two LTR elements. It was indicated that the illegitimate recombination is the main driving force behind genome size decrease in *A. thaliana* [25]. The repair of double stranded breaks in DNA is another important mechanism responsible for sequence deletion observed in some plants. The comparison between *A. thaliana* and tobacco genomes shows that there exist marked differences in their repair pathways after double stranded breaks [26]. In all of these mechanisms of sequence reduction, we speculate that the lost DNA segments may be redundant, not functionally important or, though being functional elements, they have been deleted and replaced by a more efficient functional network through a repair pathway as observed in the case of double stranded breaks. Therefore, the sequence reductions are not likely to contradict the CIQG law.

(8) The p-coding information quantity grows slowly in higher eukaryotes, not proportional to the increase of genome complexity. For example, the sequence length of protein-coding DNA in human genome is nearly the same as in mouse and even in all vertebrates. The extent of protein-coding DNA remains relatively static over a wide range of developmental complexity [10]. On the other hand, we know that the gene density in a genome decreases explicitly with the growing evolutionary complexity of species. The gene density is 1000 genes per Mb for prokaryotes to 500 genes per Mb for yeast and 20 genes per Mb for mammals. This indicates the com-

plexity of regulation mechanism increasing with evolution [27]. Therefore, we have to consider n-coding information quantity in addition to p-coding information quantity. In fact, the complexity of a genome originates from the gene function, which is determined mainly not by the number of genes contained in it but by the interaction among genes and the gene regulation. The comparison between human proteome and other eukaryotes shows that most protein domains appear to be common to the animal kingdom. However, in humans, there are many new protein architectures – new combinations of domains. The greatest increase occurs in transmembrane and extracellular proteins, which may be related to the addition of functions required for the interaction between the cells of a multicellular organism [5, 11]. Recently, the ENCODE project analyzed the functional elements in 1% of the human genome. In the "constrained sequences" (genomic sequences orthologous to the ENCODE regions from 14 mammalian species) that serve some functional roles and correspond to 4.9% of the nucleotides in the studied regions, they found that the protein-coding sequences only amount to 32% [28]. This indicates that in the functional region of human genome a large portion of sequences are not responsible for protein-coding but possibly for gene regulating. Consider the gene expression regulation at transcriptional level. The rate of transcription is modified by binding to enhancer sequences or other regulatory elements (RE) by transcription factors (TF). REs are typically 5 to 10 bp sequences, often assembled into regulatory modules (enhancers, repressors, etc.). A module typically contains binding sites for 4 to 8 TFs. Here, the complexity consists in the following: related TFs may recognize similar binding motifs (so there can be cross reaction), the RE sequences for a given TF can vary (variation in recognition sequences can lead to binding by different TFs), and the location of REs also varies among genes and among species (on either side of the gene near the gene or tens of kilobases away). Therefore, the number and nature of regulatory elements are easily changed in a reasonably short periods of evolutionary time. Therefore, even if the protein-coding part of a gene is not altered, its usage may be. The information contained in the regulatory mechanism is n-coding information quantity. Taking the variability of RE into account, if there are 4^{10}REs of 10 bp, and each RE occurs once in the genome, then it should amount to 1.68×10^8bp, about 1/20 of human genome. The above estimate only accounts for cis-acting transcriptional regulatory elements. The trans-acting regulatory elements and the gene regulation at other levels need more coding information quantity. Following the ENCODE project, more than 93% of bases in human genome were detected in primary transcripts, and many transcripts were discovered in distal regions to protein-coding loci. Of course, a part of primary transcripts is nonfunctional. It was estimated that at least 19% of pseudogenes in the ENCODE regions are transcribed. After removing nonfunctional pseudogenes and protein-coding parts, the remaining large portion of transcripts in human genome belong to non-protein-functional elements. They reside in intergenic, intron, and other untranslated regions and constitute n-coding sequences. These non-protein-coding RNAs (ncRNAs) include two main classes, namely, structural RNAs (transfer RNAs, ribosomal RNAs, and small nuclear RNAs, etc.) and regulatory RNAs (for example, micro-RNAs and small interference RNAs [29, 30]). Some authors [31] gave more detailed functional classification of established and emerging noncoding RNAs, which includes RNA processing and modification, transcription, translation, protein trafficking, regulation of gene expression, and genome stability, etc. Therefore, one may conclude with assurance that for higher organisms, the n-coding information quantity is much larger than p-coding information quantity in a genome. The sum of p- and n-coding information quantity will be able to show a stronger correlation with genome evolution.

Many discussions on junk DNA were carried out in recent years. "Not junk after all," people said [32, 33]. From ultraconserved nongenic DNA sequences in mammalian genomes [34, 35] to non-coding RNAs and to mobile elements [36], all these elements may code for certain functions. They should have contributed to n-coding information quantity. Recent studies on human chromosome 18 provide a new example of functional role of non-protein-coding elements [37]. Despite the low density of protein-coding genes on chromosome 18, it has been found that the proportion of non-protein-coding sequences evolutionarily conserved among mammals is close to the genome-wide average. These sequences might serve a structural role, with a constant density of such elements required to maintain chromosome structure independent of gene density.

(9) The law of coding information quantity growing has found a more direct evidence in the evolution of human genome [38]. Through the comparison of human genome with chimpanzee, baboon, and lemuer, Liu and Eichler *et al*. found a 15% to 20% expansion of human genome size over the last 50 million years of primate evolution. More plainly, these three species – chimpanzee, baboon and lemuer – are estimated to have diverged from human at three different time point, approximately 5.5, 25, and 55 million years ago; and compared to chimpanzee and lemuer, the human genome is estimated to have expanded 30 Mb and 550 Mb, respectively. Under the assumption that the genome size of lemuer, baboon, and chimpanzee increases slowly, respectively, after their divergence and the expansion of human genome is related

to certain events coding for new functions, the following picture can be deduced: the coding information quantity of human genome (including p-coding and n-coding sequence) has grown 30 Mb for the last 5.5 Myr and 550 Mb for the last 55 Myr. In fact, orthologous comparisons have shown that 90% of the human expansion is due to new retroposon insertions, and it is reasonable to assume that the insertion and fixation of retroposons L1 and Alu in the human genome are related to the emergence of new functions. For example, BC200, a brain-specific RNA that is part of a ribonucleoprotein complex preferentially located in the dendrites of all anthropoid primates appears to have been derived about 35–55 million years ago from an Alu transposable elements [39]. Many of the genetic differences between humans and other primates resulted from transposable element activity – through the action of regulatory regions that descended from transposable elements (TE) and/or via TE-mediated exon splicing and deletion mutations. The inactivation of the CMP-N-acetylneuraminic acid hydroxylase enzyme gene in humans is accomplished by an exonic deletion caused by the insertion of human-specific Alu elements. This deletion mutation occurred prior to brain expansion during human evolution [40]. Sometimes, the deletion of an old gene is favorable to gain a new gene or to form some new biochemical pathways for advanced function. The total coding information quantity of the genome is increased in the process. It was estimated that a total of 700–1800 duplications have occurred in the human genome since the split with chimpanzees. These duplications and accompanying some deletions of genes may generate the variation in the gene expression between chimps and humans, especially in the brain [41]. The loss of body hair is another example of the role of gene deletion in human evolution. The ape hair keratin is functional and encoded in clustered gene families, meaning that they were produced by duplication, while in humans, a type I hair keratin pseudogene inactivated by a single point mutation was observed [42].

3 Approaches to information quantity growing in a genome

There are four mechanisms responsible for the expansion of coding information quantity in a genome, namely, the sequence duplication that increases the genome size, the functionalization of transposable elements and other "junk" DNA, the formation and employment of diverse modes for coding, and the gene horizontal transfer from other species.

The sequence duplication includes global duplications (in which the entire genome or a chromosome is duplicated) and regional multiplication. The distribution of bacterial genome size is discontinuous, showing major peaks at around 0.8 Mb, 1.6 Mb, and 4.0 Mb. This distribution has led to the hypothesis that the larger genomes evolved from smaller ones by genome duplication. The polymodal distribution of genome sizes in many groups of eukaryotes (for example, in monocotyledon plants) also suggests that polyploidy is a major mechanism in eukaryotic evolution. There is strong evidence that two rounds of whole-genome duplication occurred in the vertebrate evolution. The regional multiplications are more frequently observed in many genomes. About 38%–45% of genes in an *E. coli* genome are identified as paralogous families that arisen from gene duplication, and about 47% of genes in *B. subtilis* genome constitute paralogous gene families, etc. In the human genome, about 53% of DNA is occupied by repetitive sequences. The localized regional increase of genomic DNA through tandem duplications may be created by replication slippage, slipped-strand mispairing, hairpin formation, etc. Replication slippage provides a powerful mechanism for the rapid proliferation of tandemly repeated sequences within a genome. In the case of long repeated units, the major mechanism for expansion is unequal crossing-over or DNA amplification due to multiple replication of the same replica [43].

The repetitive DNA sequences constitute a large portion of the genomes of eukaryotes. They arise from sequence duplication. The "selfish DNA" hypothesis proposes that they are maintained by their ability to replicate within the genome. They are "junk" produced in neutral mutation. For example, the inverse correlation between development rate and amount of highly repeated DNA in flowering plants and salamanders is compatible with nonselective mechanisms for maintaining repeated arrays since slower developing species could accumulate larger quantities of repetitive sequences generated by array size expansion. The large differences in amounts of satellite sequences between closely related species and the absence of measurable fitness effects of large deletions or duplications of heterochromatin in *Drosophila* also support nonselective mechanisms [44]. Therefore, there is a considerable proportion of neutral elements (repetitive DNA sequences) that do not confer a selective advantage or disadvantage to the organism. However, the neutral nonselective mechanism for the maintenance of repetitive sequences does not mean they are genomic garbage with completely no use. In fact, genomes are dynamic entities: New functional elements appear and old ones become extinct. The neutral pool of sequence elements may turn over during evolutionary time, emerging via certain mutations and disappearing by others. Those sequences having not been eliminated by evolutionary process might be related to their acquisition of some new biological role. The evolutionary picture is that some repetitive sequences acquire new func-

tions and form new genes or regulatory elements, making the increase of coding information quantity; while some lose the activity and change to pseudogene, without contributing to the coding information quantity in the further evolution. The above points were indicated by several authors in the last decade, including that by a Chinese monograph published in 2000 [45]. Now, more and more biologists recognize the importance of functionalization of "junk" DNA and regarded the neutral pool of sequence elements as a genomic treasure [28].

The most important examples of the functionalization of neutral pool elements are retrotransposons. The mobile elements (transposable elements) comprise a group of distinct DNA sequences that have the ability to integrate into the genome at a new site. These elements include DNA transposons, LTR (long terminal repeats) retrotransposons, non-LTR autonomous retrotransposons, and non-LTR nonautonomous retrotransposons. DNA transposons occupy about 3% of human genome. These elements are generally excised from one genomic site and integrated into another by a "cut-and-paste" mechanism. Retrotransposons are transcribed and reintegrated into the genome, thereby duplicating the elements. LTR retrotransposons is similar to retroviruses in structure. Non-LTR retrotransposons are classified into autonomous (with ability to encode endonuclease and reverse transcriptase) and nonautonomous (not encoding endonuclease and reverse transcriptase). The former is typified by LINE-1 (long interspersed nucleotide elements-1) of mammals, and the latter is typified by SINE (short interspersed nucleotide elements) or Alu sequences in human. Both they are highly-repetitive sequences. The amount of retrotransposons in a genome increases with evolution, about 3%–5% for lower eukaryotes but close to 50% for mammals. LINE-1 sequences have accumulated to 17% of the human genome, while Alus have expanded to 1.1 million copies or 11% of it. These elements have driven genome evolution in diverse ways: they are drivers of genome evolution [36]. Mammalian L1 elements affect the genome in many ways, for example, repair of double-strand breaks, expression of genes 5' to full-length L1s via an antisense promoter in L1, use of L1 in coding regions of genes, etc. A large burst of Alu insertion has happened 40 million years ago in hominid lineage but after that the activity declined markedly. Alu elements modulate gene expression at the posttranscriptional level in at least three independent manners: alternative splicing, RNA editing, and translation regulation. Therefore, the fast evolution of vertebrates has put the transposable elements to use. Based on three-element-interaction among DNA, RNA, and protein, the retrotransposons, carrying a large amount of information, should be regarded as a large reservoir of regulatory functions that have been actively participating in mammalian evolution. Especially that the L1 and Alu sequences may have played distinct roles in the genome evolution of hominid lineage.

New code formation is another important origin of information quantity growing in a genome. The employment of diverse modes for coding leads to the expansion of genome coding information quantity. The formation of code in a complex system is caused by the stochastic interaction among its subsystems. It emerges from the structural matching of subunits and the physicochemical interaction among them. Once the new mode for coding is formed and selected for functional use, the mode will survive in evolution. The emergence of 21^{st} amino acid selenocysteine and the 22^{nd} amino acid pyrrolysine are two examples. Both of them are produced from the formation of new coding rules by reinterpreting nonsense codons [46]. We emphasize that the code used in genome is far from only one single mode – the amino acid code. For instance, microRNA (miRNA) and small interfering RNA (siRNA) provide examples of diverse codes other than amino acid code. Both of them are 21–25 nucleotide noncoding RNAs that regulate gene expression in a sequence-specific manner. These molecules are recognized and processed by a common RNase-III processing enzyme – Dicer. They are assembled into the RNA-induced silencing complex (RISC). The effector complex RISC subsequently acts on its target by translational repression or mRNA cleavage depending partly on the level of complementarity between the small RNA and target [30]. Therefore, the gene-silencing function of miRNA and siRNA is essentially related to a code existed in the 21–25 nucleotide RNA sequence. The example also shows the formation of any code relation is a very complex process that needs the participation of many factors. Splice junctions at the exon–intron boundaries have a code GU at the 5' end and AG at the 3' end of the intron. However, to decide that GU is really a 5'splice site, and AG is really a 3' splice site in human genes, a consensus sequence A/CAG<u>GU</u>RAGU around GU and a consensus sequence YNYUR<u>A</u>Y near branching point A, followed by polyprimidine Y10-20, and then by Y<u>AG</u> around 3' splice site are required [47, 48]. These consensus sequences around splice sites form the splice code. The five small nuclear RNA (snurps)–U1, U2, U5, U4, and U6, together with 50–100 additional proteins, form the splicesome responsible for the lariat-like splicing mechanism. The splice code indicates the location in the pre-RNA sequence, where the splicesome can bind, and the splicing mechanism can work. Recently, it is known that 75% of human genes have alternative splicing. There are five alternative splice forms, namely, intron retention, alternative donor introns, alternative acceptor introns, cassette exons, and mutually exclusive exons. All these splice forms can be understood based

on the interaction between splicesome and splice code sequence. The alternative splicing leads to a considerable increase of protein products, so it is one of the most significant components for the functional diversity in human genome evolution.

Generally, the extension of code is dependent on three-element interaction among DNA, RNA, and protein. However, evolution is a mender; sometimes the modes for coding have been beyond the scope of DNA, RNA, and protein interaction. In eukaryotes, the chromatin remoulding (i.e., the change of chromatin structure) is an important event for controlling gene expression. Histone octamer in nucleosome can be modified by methylation, acetylation, and phosphorylation. They alter the local chromatin structure and activate the gene. Recently, it was found in human genome that the transcription starts are largely influenced by the feature of chromatin accessibility and histone modification [28]. The chromatin remoulding and the wrapping DNA molecule form two colinear sequences of information source. Interestingly, once some histone modifications are established, such changes in chromatin may persist through cell division, creating an epigenetic state in which the properties of a gene are determined by the self-perpetuating structure of chromatin. The name epigenetic reflects the fact that a gene may have an inherited condition (it may be active or may be inactive) that does not depend on its DNA sequence. The self-perpetuating structure of prions is an example of epigenetic heredity. The common occurrence of epigenetic state in mammalian heredity reflects the evolutionary extension of the mode for coding.

Gene transfer from the outside is another approach to the growing of coding information quantity of a genome. It includes horizontal gene transfer (lateral gene transfer) and symbiotic mergers between species. The horizontal transfer means a gene is copied from the genome of one species into that of another species. It is probably frequent in bacteria. We know that about 24% *B. subtilis* genes have clear orthologous counterparts in *E. coli*, and 66% *H. influenzae* proteins have homologues to *E. coli* proteins. However, the detection of a horizontal-transfer event should be made based on rigorous phylogenetiic analysis or by either recombination or gene conversion [49]. Genes are even known to have transferred between Archaea and bacteria. Genes probably also occasionally transfer from bacteria into multicellular eukaryotes. For example, there are 223 proteins in the human genome that have significant similarity to proteins from bacteria. At least 113 of these genes appear to be present only in vertebrates. However, whether they are horizontal-transfer events from bacteria to vertebrates remains to be tested [50]. The symbiotic mergers mean two species combine their genomes into one in a particularly intimate symbiosis. It is argued that 2000–2500 million years ago, the symbiosis between two bacteria led to the formation of the eukaryotic cell containing a mitochondrion. The newly merged cell might have had two DNA molecules, and one of them has expanded and evolved into the nuclear DNA and another shrunk and evolved into the mitochondrial DNA [9].

4 Conclusions

Different from energy, information is not conservative during its transmission. The information expansion is a basic law in biology with its meaning like energy conservation in physics. A great deal of molecular biological data show the information quantity gradually expanded in evolution. DNA sequence duplication is the first important factor for the expansion. Sequence duplication includes short fragment repetition, regional multiplication, gene and genome duplication, etc. New DNAs are produced and maintained in duplications just by their ability to replicate, and these duplicated sequences greatly contribute to the genome information expansion. Enlargement of genetic information network is another important factor. Life is an information system of three-element interaction among DNA, RNA, and protein. The formation, storage, expression, and transmission of genetic information are generally realized in the interaction network of DNA, RNA, and protein. For eukaryotes, the chromatin remoulding and histone modifications is another element entering in the information interaction network. A Chinese ancient philosopher Laozi said "One generates two, two generates three, and three generates all things in the Universe." Therefore, *three* in this sense means *infinity* (Three=Infinity). Three-, or more-, element interaction opens up more possibilities and shows more complexities than two-element interaction. The interaction network of multielements makes the genetic language complex enough to be able to represent the life. For example, RNA and protein can have arithmetic function operated on the DNA sequence, modifying, deleting, or inserting some segments on it and making the function-coding information of genome increasing. The regulation of gene expression is realized through proteins reacting on DNA in usual regulatory mechanism; it can also be realized through ncRNA interaction in a three-element-interaction network. In fact, as early as Jacob and Monod proposed the regulatory model, they suggested the possibility of the highly specific interaction between operon RNA and transcripts of regulatory gene. The recent discovery of histone modifications that strongly influence the transcription starts in the human genome provides another example of how the genetic information network is enlarged. Therefore, the duplication of DNA sequence and the enlargement of genetic information network centered at DNA sequence are

two internal predominating factors in coding information quantity expansion of a genome.

However, the change of information quantity of a genetic system depends on environment and many stochastic factors. Both DNA expansion and contraction have been found in genome evolution. Only after taking the functional selection into account, the directionality of DNA evolution can then be established. That is, although there exist a lot of accidental factors changing the genome size in the course of evolution, the global trend of the function extension of the genome that makes the species more suitable to the environment and more competitive to win is decisiveness and single direction. The central points of Darwin's evolutionary theory are adaptation and competition. Following Darwin, as environments change and competing species change, species will evolve new adaptations. The competition between similar individuals will make up for the evolution of new adaptations in each one that reduces the intensity of competition; thus, there will be divergence. This explains why phylogeny is always hierarchical and tree-like (tree-like means directionality) and how the evolutionary directionality emerges. Here, the key point is functional selection. A new species formation is at the cost of the failure of many trials of old species. The advantageous function acquired by a species makes it become a winner. Therefore, the directionality of biological evolution is easily understood this way. What we proposed in the article is that the coding information quantity of a genome that codes for the function of species can serve as the measure of directionality. The needs of function are the motive force for the expansion of coding information quantity, and the information quantity expansion is the way to make functional innovation and extension. Therefore, the increase of coding information quantity of a genome is a measure of the acquired new function, and it determines the directionality of genome evolution. We know that there are several time arrows in physics that express the irreversibility of the physical laws. The proposed directionality here is a new time arrow, but it should be understood only in terms of biology. As is well known, physics used to ask "why," while biology used to ask "what for." Purpose of function improvement is fully a biological concept. Based on the functional selection, we are able to find a measure for evolutionary directions, namely, the function-coding information quantity growing in evolution. Although the CIQG law proposed here still needs further quantitative demonstrations, we have found the law consistent (or at least not in conflict) with all up-to-date experimental data in genomics.

The practical meaning of this study is as follows: It will be helpful to understand the expansion of the coding information quantity in genome evolution, to understand the distribution of "junk" DNA as a kind of "dark information" occurred in eukaryotic genomes, and to find the possible new code relations existed therein. The life information emerges from the stochastic background of inanimate Nature through millions of years' natural selection, condensed from a huge amount of accidents. What is the main line of the present theoretical molecular biology? We proposed that the main line should be focused on the flow of life information, that is, on the fundamental laws in heredity, transmission, regulation, and expression of genetic information. We called the main line as information biology [6, 7]. It is expected that the suggested CIQG law will be a first cornerstone of information biology.

Acknowledgements The work was supported by the National Natural Science Foundation of China (Grant No. 90403010). The preprint of the paper appeared in: arXiv: q-bio/0808.3323.

References

1. E. Schrodinger, What is Life? Cambridge: Cambridge University Press, 1944
2. T. R. Gregory, J. A. Nicol, and H. Tamm, Nucleic Acids Research, 2007, 35 (Database issue): D332
3. T. R. Gregory, Genome Size Evolution in Animals. In: Evolution of the Genome (Edited by T. R. Gregory), Elsevier Inc., 2005
4. T. Kouzarides, Cell, 2007, 128: 693
5. B. Lewin, Gene IX, Jones & Bartlet Publishers, Inc., 2008
6. L. F. Luo, Journal of Inner Mongolia University, 2005, 36: 653
7. L. F. Luo, Science in China Ser. C, 2006, 58: 24 (in Chinese)
8. S. E. Peters, Nature, 2008, 454: 626
9. M. Ridley, Evolution, 3rd Ed., Blackwell Publishing, 2004
10. R. J. Taft, M. Pheasant, and J. S. Mattick, BioEssays, 2007, 29 (3): 288
11. B. Lewin, Gene VIII, Pearson Education Inc., 2004
12. T. R. Gregory and R. DeSalle, Comparative Genomics in Prokaryotes. In: Evolution of the Genome (Edited by T. R. Gregory), Elsevier Inc., 2005
13. A. Mira, H. Ochman, and N. A. Moran, Trends Genet., 2001, 17: 589
14. D. A. Petrov, T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw, Science, 2000, 287: 1060
15. I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev, Nature, 2007, 449: 54
16. D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe, Nature, 2006, 440: 341
17. J. M. Aury, J. Olivier, L. Duret, et al., Nature, 2006, 444:171
18. D. L. Des Marais and M. D. Ransher, Nature, 2008, 454:762
19. I. J. Leitch and M. D. Bennett, Biol. J. Linn. Soc., 2004, 82: 651
20. H. Ozkan, A. A. Levy, and M. Feldman, Plant Cell, 2001, 13: 1735
21. R. P. Bininda-Edmonds, M. Cardillo, K. E. Jones, et al., Nature, 2007, 446:507

22. X. Xu and M. A. Norell, Nature, 2004, 431: 838
23. C. L. Organ, A. M. Shedlock, A. Meade, M. Pagel, and S. V. Edwards, Nature, 2007, 446:180
24. M. D. Bennett and I. J. Leitch, Genome Size Evolution in Plants. In: Evolution of the Genome (Edited by T. R. Gregory), Elsevier Inc., 2005
25. K. M. Devos, J. K. M. Brown, and J. L. Bennetzen, Genome Research, 2002, 12: 1075
26. J. Filkowski, O. Kowalchuk, and I. Kowalchuk, Plant Sci., 2004, 166: 265
27. W. Deng, X. Zhu, G. Skogerbo, et al., Genome Research, 2006, 16: 20
28. The ENCODE Project Consortium, Nature, 2007, 447: 799
29. G. Storz, Science, 2002, 296: 1260
30. L. He and G. J. Hannon, Nature Rev. Genetics, 2004, 5: 522
31. A. G. Matera, R. M. Terns, and M. P. Terns, Nature Reviews, 2007, 8: 209
32. W. Makalowski, Science, 2003, 300: 1246
33. I. Wickelgren, Science, 2003, 300: 1646
34. E. T. Dermitzakis, A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier, and S. E. Antonarakis, Science, 2003, 302: 1033
35. G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler, Science, 2004, 304: 1321
36. H. H. Kazazian, Science, 2004, 303: 1626
37. C. Nusbaum, M. C. Zody, and M. L. Borowsky, Nature, 2005, 437: 551
38. G. Liu, NISC Comparative sequencing Program, and E. Eichler, Genome Research, 2003, 13: 358
39. A. F. A. Smit, Curr. Opin. Genet. Dev., 1999, 9: 657
40. H. H. Chou, T. Hayakawa, S. Diaz, et al., Proc. Natl. Acad. Sci. USA, 2002, 99: 11736
41. W. Enard, P. Khaitovitch, J. Klose, et al., Science, 2002, 296: 340
42. H. Winter, L. Langbein, M. Krawczak, et al., Human Genet., 2001, 108: 37
43. L. Patthy, Protein Evolution, Oxford: Blackwell Science, 1999
44. B. Charlesworth, P. Sniegowshi, and W. Stephan, Nature, 1994, 371: 215
45. L. F. Luo, Physical Aspects on Life Evolution, Shanghai: Shanghai Science &Technology Pub., 2000 (in Chinese)
46. T. C. Stadtman, Ann. Rev. Biochem., 1996, 65: 83
47. F. Clark and T. A. Thanaraj, Human Molecular Genetics, 2002, 11(4): 451
48. L. R. Zhang and L. F. Luo, Nucleic Acids Research, 2003, 31: 6214
49. W. H. Li, Molecular Evolution, Massachusetts: Sinauer Associates, 1997
50. International Human Genome Sequencing Consortium, Nature, 2001, 409: 860