bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Quantum Theory on Genome Evolution

LiaoFu LUO

1, Faculty of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

2, School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou, China

email: lolfcm@mail.imu.edu.cn

Abstract

A model of genome evolution is proposed. Based on several general assumptions the evolutionary theory of a genome is formulated. Both the deterministic classical equation and the stochastic quantum equation are proposed. The classical equation is written in a form of of second-order differential equations on nucleotide frequencies varying in time. It is proved that the evolutionary equation can be put in a form of the least action principle and the latter can be used for obtaining the quantum generalization of the evolutionary law. The wave equation and uncertainty relation for the quantum evolution are deduced logically. Two fundamental constants of time dimension, the quantization constant and the evolutionary inertia, are introduced for characterizing the genome evolution. During speciation the large-scale rapid change of nucleotide frequency makes the evolutionary inertia of the dynamical variables of the genome largely decreasing or losing. That leads to the occurrence of quantum phase of the evolution. The observed smooth/sudden evolution is interpreted by the alternating occurrence of the classical and quantum phases. In this theory the probability of new-species formation is calculable from the first-principle. To deep the discussions we consider avian genome evolution as an example. More concrete forms on the assumed potential in fundamental equations, namely the diversity and the environmental potential, are introduced. Through the numerical calculations we found that the existing experimental data on avian macroevolution are consistent with our theory. Particularly, the law of the rapid post-Cretaceous radiation of neoavian birds can be understood in the quantum theory. Finally, the present work shows the quantum law may be more general than thought, since it plays key roles not only in atomic physics, but also in genome evolution.

Key words

genome evolution, quantum evolution, uncertainty relation, the least action principle, evolutionary rate, evolutionary inertia, speciation probability, avian genome

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Genome is a well-defined system for studying evolution of species. There are many topics on genome evolution. A central problem is genome size evolution. The C-value enigma is still puzzling and perplexing [1][2]. How to understand the time arrow of the evolution ? [3] The directionality means there possibly exists a set of basic equations for genome that evolves with respect to time t. The time t is a fundamental variable in the evolutionary theory. To find the genome evolutionary equation with respect to t , this is the first motivation of the present article. The evolution of species is a very complex process. The phylogenetic tree provides an appropriate quantitative description on evolution. However, although there have been proposed many mathematical methods to deduce evolutionary trees, it seems lacking a dynamical foundation for reconstruction of the tree from the genome sequence data. In my opinion, the reconstruction of evolutionary tree should be reconciled to the finding of the genome evolutionary equation. Then, what is the dynamical variable appropriate to reconstruct the evolutionary tree and deduce the genome evolutionary equation simultaneously? Since the k-mer non-alignment algorithm (KNA) is a good tool to obtain the evolutionary tree [4], we try to use the nucleotide (k-mer) frequencies as a set of dynamical variables in differential equations describing evolution. To establish a mathematical theory on genome evolution that can be reconciled to evolutionary tree reconstruction, this is the second motivation of the present article. Thirdly, from the evolutionary phenomenology, two theoretical points, phyletic gradualism and punctuated equilibrium, were proposed to explain the macroevolution. It seems that both patterns are real facts observed in fossil evolution [5]. A deeper research question is what conditions lead to more gradual evolution and what conditions to punctuated evolution, and how to unify two patterns in a logically consistent theory. To give a unified mathematical theory on genome evolution, this is the third motivation of the present article. We shall suggest a set of second-order differential equations for genomic nucleotide frequencies which varies in time. Then, by putting the differential equation in a form of the least action principle, we shall demonstrate that the classical evolutionary trajectory can generally be replaced by the trajectory-transitions among them. Thus, the concept of quantum evolution is introduced. Both classical phase (gradually and continuously) and quantum phase (abruptly and stochastically) that have been observed in the evolution can be explained in a natural way. Fourthly, what rules are obeyed by the new species production? -- this is a difficult problem in existing evolutionary theory. However, the quantum theory gives us a fully new view on genome evolution. We shall demonstrate that the problem can be treated in a quantitative way, demonstrate that the speciation events obey a statistical law and the speciation probability can be calculated from the point of quantum transition. To give a formulation to study the new species production, this is the fourth motivation of the present article.

The extinction of the dinosaurs and pterosaurs at the end of Cretaceous (66 millions of years ago) left open a wide range of ecological niches, allowing birds and mammals to diversity. Most of the vacated avian niches were thought to have been filled by a rapid post-Cretaceous radiation of neognathous birds which today include over 99% of all bird species.[6]-[9]. Moreover, avian genomes show a remarkably high degree of evolutionary stasis at the level of chromosomal structure, gene syntency and nucleotide sequence. The macrochromosome studies reveals that the ancestral pattern has remained largely unaltered in the majority of avian genomes[10]. How to

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

understand the law of the rapid post-Cretaceous radiation of birds, more than ten thousands of avian genomes producing explosively in a period shorter than 10^7 years? How to explain the rapid accumulation of information in the early stage and soon afterwards the evolutionary stasis for avian genomes? To answer these questions we need a quantitative theory of evolution. On the other hand, the bird macroevolution was explored recently by using full genomes from 48 avian species representing major extant clades [11][12]. Then a more comprehensive phylogeny of 198 avian species was established by using targeted next-generation DNA sequencing [13]. The abundant data for avian species afford a splendid field for investigation of the genome evolution. In the present article we shall use the proposed quantum theory of evolution to study avian genome and explain the evolutionary data.

The k-mer frequency distribution is a characteristic of a genome (genome barcode) [14][15]. Several years ago we pointed out that the evolutionary tree can be deduced by non-alignment method – k-mer frequency statistics. By using 16S rRNA (18S rRNA) as molecular clock, we found that the phylogenetic trees deduced from k-mer frequency agree well with life tree when k≥7 [4]. If the genome growth is a fully stochastic process then DNA should be a random sequence and the k-mer frequency in the sequence takes Poisson distribution with the most probable value $\alpha(k)=\frac{N}{4^k}$ where N is genome length. However, the non-random force does exist in genome evolution. The most important non-random forces include functional selection and sequence duplication, etc. Due to the non-random forces the k-mer frequency deviates from Poisson distribution . Set f_i^k is the frequency of the i-th k-mer. The square deviation $\sigma_\alpha{}^2$ is expressed by $\sigma_\alpha^2(k)=\sum_i \frac{(f_i^k-\alpha)^2}{4^k}$. The ratio of the deviation $\sigma_\alpha{}^2(k,N)$ for a genome sequence of length N to the deviation $\alpha(k,N)$ for a random sequence of same length represents the nucleotide correlation strength or non-randomness level of the genome. It was found that there exists a meaningful relation between correlation strength and genome complexity [14]. Aiming to find the dynamical variable of evolution we used k-mer non-alignment algorithm (KNA) to reconstruct the evolutionary relationship for avian genomes.[16] Denote the probability of base a (a =A,G,C or T) occurring in a sequence by p_a , and the joint probability of base a and b occurring sequentially in the sequence by p_{ab} . In general, let $\sigma = abc\ldots$ being an oligonucleotide k bases long, we denote the joint probabilities of the bases in σ occurring in the sequence by p_σ. For any given k we always have $\sum_\sigma p_\sigma = 1$, where the summation over σ is over the set $\{\sigma\}$ of the 4^k oligonucleotides of length k. Given two sequences Σ and Σ' with sets of joint probabilities $\{p_\sigma\}$ and $\{p_\sigma'\}$, respectively, define a distance, called a k-distance, between the two sequences based on the difference of joint probabilities in the two sets as follows

$$D_k(\Sigma,\Sigma')=\sum_\sigma \left| p_\sigma - p_\sigma' \right|$$

$$(k=1,2,\ldots)$$

For long sequences a k-distance is insensitive to minor misalignments between two sequences. By using sequence data for 47 representative avian genomes we calculated the k-distances and deduced the phylogenetic tree for each k . We found that the quality of tree is gradually improved

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and the tree topology is stabilized with increasing k and reaches a plateau at k=11 or 12. The tree deduced by KNA algorithm (Figure A1 in Appendix) is consistent with experimental evolutionary relationship and basically in accordance with those obtained in literatures [12][13]. Above calculations show that the k-mer frequency is a good dynamical variable for describing genome evolution. This provides a conceptional basis for studying evolution through a set of evolutionary equations for nucleotide frequencies in the genome.

The materials are organized as follows. Firstly, three basic assumptions on the quantum theory of evolution are formulated. The classical evolutionary equation of the genome is written in terms of nucleotide frequencies varying with time. The evolution is described by a second-order differential equation. The path-integral quantization of the equation leads to the quantum transition between classical trajectories. To apply the theory to the practical system the supplementary assumptions on the form of evolutionary potential are proposed. In the subsequent sections the classical and the quantum motions of the genome are respectively discussed. In the classical phase evolution, the environmental potential parameters, evolutionary inertial parameter and dissipation parameter will be estimated by fitting avian genomic data . The rapid accumulation of information and the evolutionary stasis for avian genomes will be explained. In the quantum phase evolution the discreteness of quantum state and the ground-state wave function of avian genome are deduced. New species production are calculated by the quantum transition between discrete quantum states and based on the calculations the rapid post-Cretaceous radiation of neoavian birds can be understood more directly. In the final section, a concluding remark will be indicated that the quantum law may be more general than thought, since it plays key roles not only in atomic physics, but also in genome evolution.

Basic assumptions of the quantum theory on genome evolution

Ansatz 1: For any genome there exists a potential $V(x_1,...,x_m,t)$ to characterize the evolution where x_i means the frequency of the i-th nucleotide (or nucleotide k-mer, $m=4^k$) in DNA.

Since $V(x_1(t),...,x_m(t),t) \neq V(x_1(-t),...,x_m(-t),-t)$ in general, the potential can be used to describe the evolutionary direction.

Ansatz 2: The genome evolution equation reads as

$$\frac{d}{dt}(c^2\frac{dx_i}{dt}) = \frac{\partial V}{\partial x_i} - f\frac{dx_i}{dt} \tag{1}$$

where f >0 is a dissipation coefficient representing the effect of fluctuation force. The parameter c^2 is introduced with the dimension of (time)^2 which represents the evolutionary inertia of the

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dynamical variables $\{x_i\}$ of the genome.

The evolutionary law can be reformulated based on Feynman's action integral. Introduce a functional (called information action)

$$S[x_1(t),...x_m(t)] = \int (\frac{c^2(t)}{2}\sum(\frac{dx_i}{dt})^2 + V(x_1,...x_m,t))dt \tag{2}$$

Then the solution of (1) can be expressed as

$$\delta S = \sum_i (\frac{\partial S}{\partial x_i}\delta x_i) = 0 \tag{3}$$

as the dissipation is weak（f can be neglected）. Therefore, the classical evolutionary trajectory $x_i(t)$ satisfies the principle of the least action. By use of path integral quantization the evolutionary trajectory theory can be generalized to a more general quantum formalism.

Ansatz 3: The genome evolution obeys a general statistical law, described by the information propagator [17]

$$U(\mathrm{x}',t';\mathrm{x}_0,t_0) = A\sum \exp(\frac{iS[\mathrm{x}]}{L}) \tag{4}$$

($\mathrm{x}=(x_1,...x_m)$) . The summation is taken over all ideal paths satisfying $\mathrm{x}=\mathrm{x}_0$ at t_0 and $\mathrm{x}=\mathrm{x}'$ at t' which is termed as functional integral in mathematics. (see Appendix 1)

Here L is a quantization constant of time dimension. The path integral $U(\mathrm{x}',t';\mathrm{x}_0,t_0)$ describes the evolution of the genomic statistical state from t_0 to t'. When the virtual variation $\delta x_i(t)$ makes $\delta S >> L$ all terms in the summation (4) will be canceled each other due to phase interference apart from those in the vicinity of classical trajectory where S takes a stationary value (δS =0). Therefore, the classical trajectory is a limiting case of the general quantum theory and the constant L can be looked as the threshold for classical approximation. Since during speciation the definite classical trajectory has lost its meaning and instead the quantum picture of trajectory transitions holds, we assume the quantization constant L is related to the time for new species formation. The constant L is species-dependent because of the generation time and therefore the speciation time varying largely from species to species.

Supplementary assumptions To make concrete analysis on genome evolution we further assume

$$V(x_1,...,x_m,t) = D(x_1,...,x_m) + W_{env}(x_1,...,x_m) \quad (m = 4^k) \tag{5}$$

W_{env} is a selective potential dependent of environment, and V depends on t through the change of environmental variables. D means the diversity-promoting potential

$$D(x_1,...,x_m) = N\log N - \sum_i^m x_i \log x_i, \qquad N = \sum_i^m x_i \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(hereafter the notation log means $\log_2$) Notice that the potential defined by (5) equals Shannon information quantity multiplied by *N*. Set $f_i = \frac{x_i}{N}$, one has

$$D = -N\sum_i^m f_i \log f_i \tag{7}$$

If the correlation between bases can be neglected then the diversity of *k*-mers is *k*-fold of the diversity of single nucleotides. In literatures *D* is called diversity measure which was firstly introduced by Laxton [18]. In their studies the geographical distribution of species (the absolute frequencies of the species in different locations) was used as a source of diversity. Recently, the method was developed and applied successfully to various bioinformatics problems, for example, the intron splice site recognition[19], the promoter and transcriptional starts recognition[20], the protein structural classification[21], the nucleosome positioning prediction[22], etc. Now we shall use it to study evolutionary problem. Many examples show that the genome always becomes as diverse as possible and expands their own dimensionality continuously in the long term of evolution [3]. These observations are in agreement with the evolution-promoting potential given by Eq (6). In his book "*Investigation*" Kauffman wrote: "Biospheres, as a secular trend, that is, over the long term, become as diverse as possible, literally expanding the diversity of what can happen next. In other words, biospheres expand their own dimensionality as rapidly, on average, as they can." and called it "the fourth law of thermodynamics for self-constructing systems of autonomous agents" [23]. Our proposal on evolution-promoting force is consistent with Kauffman's suggestion about the force for "expanding the diversity" in autonomous agents.

Assume only the leading term retained in environmental potential W_{env} which is assumed as a function of $N=\sum_i^m x_i$. As we know that for the case of small kinetic energy $(\frac{dx_i}{dt})^2$ the classical trajectory of a mechanical system is always near the bottom of the potential energy. Through the mechanical simulation one may further assume the high-order term of *N* in W_{env} is neglected. Thus we have

$$W_{env} = \alpha N - \beta N^2 \qquad (\beta > 0 \qquad N = \sum_i x_i) \tag{8}$$

Note the form of W_{env} is invariant under the transformation of the oligomer length from *k* to *k*-1. The environmental selection is called positive as the selective potential W_{env} increases with *N* and called negative as W_{env} decreases with *N*. As $\beta > 0$ and $\frac{\alpha}{\beta} < 2N$, the environmental selection is always negative. The parameters α and β in environmental potential W_{env} are assumed to be constant corresponding to the stable evolution. However, α and β may change with *t* in the varying environment and the genome will select new evolutionary trajectory (new solution

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of Eq (1)) to adapt the environment. Sometimes, the sudden change of the environment makes $\beta(t)$ increasing too rapidly and the genome may be shrunken and degenerate and the old species will be close to extinction.

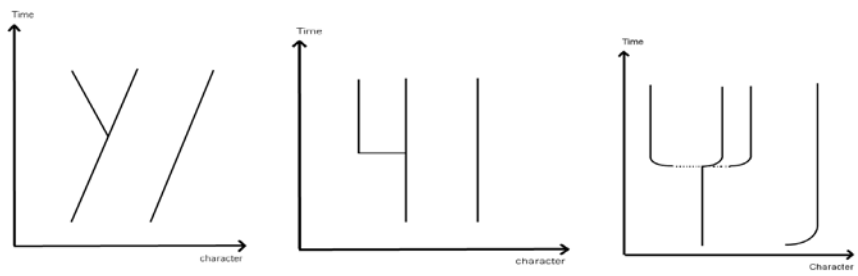
Results and discussions

1 Alternating occurrence of classical and quantum phases

Classical phase means the smooth evolution obeying the classical deterministic law, while the quantum phase means the sudden evolution obeying the quantum stochastic law. The present model of genome evolution predicts the alternating occurrence of both phases.

Many different estimates for the rate of evolution were made from the fossil records. As compiled by Gingerich[24][25] four hundred and nine such estimates were reported and they vary between 0 and 39 darwins in fossil lineage. Paleobiological studies indicated that species usually change more rapidly during, rather than between, speciation events. The smooth evolution always occurs between speciation events and the sudden evolution preferably occurs during speciation. The former can be interpreted as the classical phase and the latter as the quantum phase in our model. Paleobiological studies also indicated that the structurally more complex forms evolve faster than simpler forms and that some taxonomic groups evolve more rapidly than others. All these observations can be interpreted by the alternating occurrence of classical and quantum phases and discussed quantitatively based on evolutionary equations proposed above.

Phyletic gradualism states that evolution has a fairly constant rate and new species arise by the gradual transformation of ancestral species. While punctuated equilibrium argues that the fossil record does not show smooth evolutionary transitions. A common pattern is for a species to appear suddenly, to persist for a period, and then to go extinct. Punctuated equilibrium states that evolution is fast at times of splitting (speciation) and comes to a halt (stasis) between splits. The theory predicts that evolution will not occur except at times of speciation. It seems that phyletic gradualism and punctuated equilibrium are contrasting and contradicting theories.[26] However, from our model of genome evolution both smooth and sudden phase should occur in a unifying theory. By using the evolutionary equation (4) it is easily to deduce that the evolution has a range of rates, from sudden to smooth. Moreover, from the general formalism given by Eq (4) the evolutionary trajectories during speciation should be switched to quantum transitions among them . Thus, from the present theory the punctuated equilibrium and the phyletic gradualism are only the approximate description of two phases of an identical process.



a b c

Figure 1 Comparison of three evolutionary theories

(a) phyletic gradualism, (b) punctuated equilibrium, (c) quantum evolution (dotted line means quantum transition)

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2 Laws in classical phase

From (1) and (5)(6) the classical genome evolution equation reads as

$$\frac{d}{dt}(c^2\frac{dx_i}{dt}) = \log\frac{N}{x_i} - (1-\frac{x_i}{N})\log e + \frac{\partial W_{en}}{\partial x_i} - f\frac{dx_i}{dt} \tag{9}$$

It is easily to prove that

$$\log\frac{N}{x_i} - (1-\frac{x_i}{N})\log e \geq 0$$

(equality occurs at $N=x_i$ only) and the left-hand side increases with $\frac{N}{x_i}$. This means the diversity force always acts in positive direction and the less $\frac{x_i}{N}$ the stronger the force. As all nucleotide oligomers have equal frequency $x_i = \frac{N}{m}$ the same promoting forces are applied on them equally.

Consider single nucleotide evolution (k=1 case) first. To obtain more detail understanding on classical evolutionary trajectory , we shall study avian genome as an example. From the full-genome sequence data of avian species [11]-[13] it is found the size N (sequence length) of genomes takes a value between $(1044 - 1258)\times 10^6$ (Table A1 in Appendix 3) . The informational redundancy R_I is defined by

$$R_I = 2 + \sum_i^4 \frac{x_i}{N}\log_2\frac{x_i}{N} \tag{10}$$

(x_i - the frequency of nucleotide i in a genome). For 48 avian genomes studied in [11][12] we found R_I takes a value between $0.008 - 0.026$ that shows x_i very near to $N/4$.

Inserting Eq (8) into Eq (9) it leads to the equilibrium occurring at

$$(\frac{dx_i}{dt})_{eq} = \frac{1}{f}(\log\frac{N}{x_i} + \frac{x_i}{N}\log e - \log e + \alpha - 2\beta N) \quad (x_i \approx \frac{N}{4}) \tag{11}$$

When N close to $N_0 = \frac{2 - 0.75\log e + \alpha}{2\beta}$ the genome size varies slowly and it tends to stabilization. We call it in stasis phase. Since the majority of avian genomes shows apparently the evolutionary stasis, if the parameter α and β are assumed same for different Aves then one may estimate β from the smallest length N_{small} of the studied avian genomes. On the other hand, the quantum theory gives the genome length distribution for a stationary state having a peak at $N = \frac{2+\alpha}{2\beta}$ (see later, Eq (34)). So, we estimate

$$\beta = \frac{2+\alpha}{2N_{\text{small}}} \approx 1.044\times 10^{-9} \quad (\alpha=0.18) \tag{12}$$

where α is assumed a small number between 0 and 2.

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Through integration of Eq (9) one obtains

$$\frac{d}{dt}\{\frac{1}{2}c^2(\frac{dx_i}{dt})^2 - D - \alpha N + \beta N^2\} < 0 \tag{13}$$

due to the damping effect of friction force. The dissipation of friction force is accumulated in time. In the initial evolution of a genome one may neglect f term in Eq (9) and deduce that x_i increases rapidly with a dimensionless acceleration

$$Ac_i = \log\frac{N(t_0)}{x_i(t_0)} - \log e(1-\frac{x_i(t_0)}{N(t_0)}) + \alpha - 2\beta N(t_0) \tag{14}$$

at $t=t_0$.

Introducing dimensionless $\tau = t/c$ Eq (9) is rewritten as

$$\frac{d^2x_i}{d\tau^2} = \log\frac{N}{x_i} - (1-\frac{x_i}{N})\log e + \alpha - 2\beta N - \frac{f}{c}\frac{dx_i}{d\tau} \tag{15}$$

$$(N = \sum_i^4 x_i)$$

Eq (15) can be solved for each avian genome. The only parameter in equation (15) is f/c in addition to common environmental parameters α and β. The solutions for 12 Aves with branching time longer than 50 million years are shown in Fig 2. For example, the branching time of Common osterich is 51Mya and N= 1.228×10^9 at present. In calculation the parameter f/c =0.00169 is chosen for Common osterich. We found x_i and N of this genome increase rapidly with the dimensionless time τ in a period of Δτ=3600 or the corresponding acceleration time Δt= 4M. Then, after the initial expansion the evlution comes to stasis. From Δt and Δτ we obtain c=1100yr for Common osterich. The inertia of other genomes can be deduced in the same way. We found c obeys

$$c = \frac{\text{const}}{\sqrt{N}} \tag{16}$$

approximately. In fact, for $x_i \approx \frac{N}{4}$ the equation (15) describes a damping oscillation with $\omega = \sqrt{8\beta}$. Since β is oppositely proportional to sequence length, one can obtain Eq (16) immediately. On the other hand, from the input f/c and calculated c for Common osterich we obtain f=1.86 yr for all Aves. Because many birds are in stasis phase from 50 Mya till now one

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

estimates

$$(\frac{dx_i}{dt})_{eq} = \frac{1}{4}(\frac{dN}{dt})_{eq} < \frac{1}{4} \times \frac{(1258-1044)\times 10^6}{50\times 10^6} \approx 1.1/yr$$

The dissipation force $f\frac{dx_i}{dt} \cong 1.86\times 1.1 \cong 2$ matches well with the promotion force $\log\frac{N}{x_i}$, consistent with experimental data.

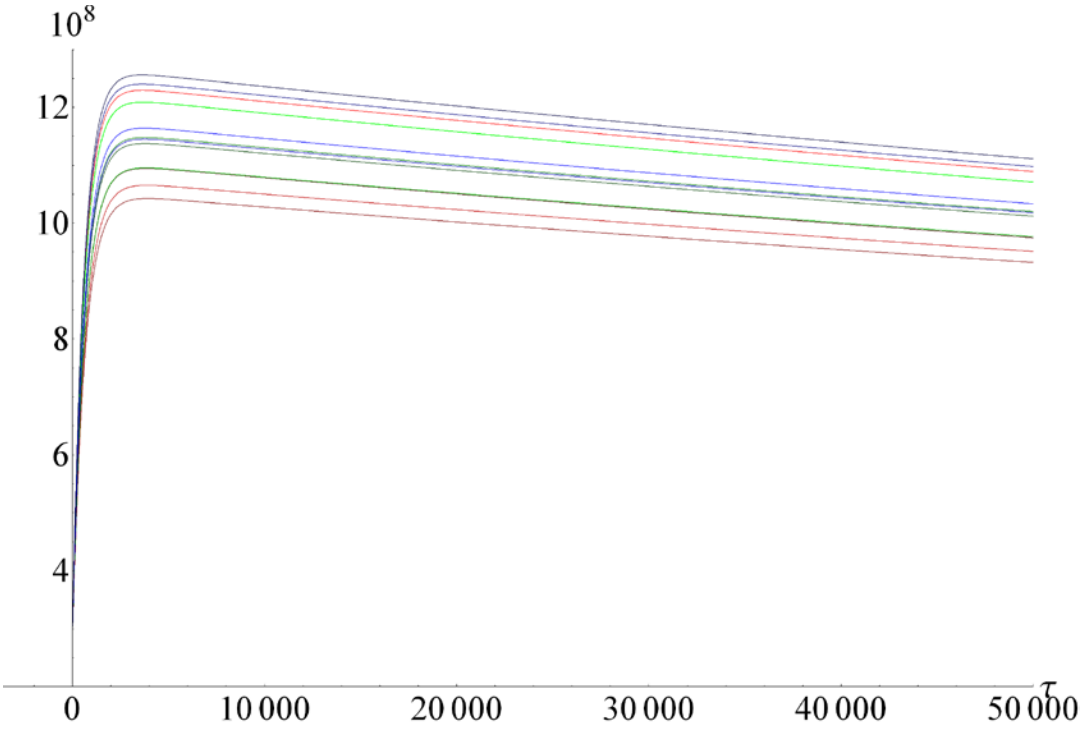


Fig 2 The evolutionary trajectory of 12 avian genomes

The genome length N versus the dimensionless time τ for 12 Ave (Penguin, Crested ibis, Common osterich, Hoatzin, White-tailed tropicbird , Cuckoo-roller , Red-throated loon, Red-crested turaco, Sunbittern, Mac Queen's bustard, Carmine bee-eater, and Rifleman from top to bottom) are plotted. N increases rapidly and attains maximum at τ smaller than 4M. The parameters $\beta = 1.0435\times 10^{-9}$, α=0.18, $f = 1.86$ yr are taken and c can be found in Table A1 in Appendix. The evolutionary inertia c increase for 12 genomes from top to bottom.

Note that in above calculation the environmental parameter is supposed to be a constant. However, if one assumes that β changes with time in the early stage of N expansion, the evolutionary trajectory will retain its basic character as shown in Fig 2 since βN^2 term in equation is small as N not too large.

Now we will discuss the evolution of oligonucleotide frequency of a genome. The k-mer frequency evolution is closely related to the single-nucleotide frequency evolution (k=1 case) of a genome. Generally the functional evolution is described by the former while the latter is more related to the macroevolution of a genome.

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Eq (9) still holds for *k*-mer case (*k*>1). From Eqs (6) and (7) one has $D(m=4^k)=kD(m=4)$ for independent sequence where $D(m=4)$ means diversity of single nucleotide. That is, if the correlation between neighboring bases is switched off then the *k*-mer diversity equals *k*-fold of the single nucleotide's. If the consistency between *k*-mer and single nucleotide evolutionary equations in weak correlation is required then the environmental potential $W_{env}=\chi(N)$ for k-mer should be equal k-fold of $W_{env}=\alpha N-\beta N^2$ (Eq (8)) for single nucleotide evolution. Therefore, one obtains a similar set of equations for *k*-mer frequency as Eq (15)

$$\frac{d^2x_i}{d\tau^2}=\log\frac{N}{x_i}-\frac{1}{\ln 2}(1-\frac{x_i}{N})+\frac{d\chi}{dN}-\frac{f}{c}\frac{dx_i}{d\tau}$$
$$(i=1,\ldots,m=4^k),\qquad (\sum_i x_i=N) \tag{17}$$

but with the replacement

$$\chi(N)=k(\alpha N-\beta N^2) \tag{18}$$

As *N* is much smaller than 4^k only a small fraction of the total set of *k*-mers are involved and only this subset of *k*-mers should be considered in detail in equations (17) . Consider *k*=12 case. For 47 avian genomes the 12-mer frequency distribution peaks at some subet. Nearly for all avian genomes only about 11 to 18 oligomers of the total set of $m=4^{12}$ have frequency equal or larger than 5. The highest frequency of a single 12-mer is 16. These *q* （*q*=11, 12,....., or 18） kinds of 12-mers form a subset for each genome. Set $X=x_1+x_2+\ldots+x_q$ representing the total frequency of the subset. We have

$$\frac{d^2X}{d\tau^2}=\sum_i \log\frac{N}{x_i}+\frac{X}{N\ln 2}-\frac{q}{\ln 2}+q\frac{d\chi}{dN}-\frac{f}{c}\frac{dX}{d\tau} \tag{19}$$

where

$$\sum_i^q \log\frac{N}{x_i}\cong q\log\frac{qN}{X}$$

and

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\frac{d^2N}{d\tau^2} = \sum \log\frac{N}{x_i} - \frac{m-1}{\ln 2} + m\frac{d\chi}{dN} - \frac{f}{c}\frac{dN}{d\tau} \tag{20}$$

where

$$\sum_i^m \log\frac{N}{x_i} \cong q\log\frac{qN}{X} + (m-q)\log\frac{N}{N-X} + (m-q)\log(m-q)$$

$X(t)$ and $N(t)$ can be solved from Equations（19）and（20）.

3 Basic characteristics in quantum phase

The nucleotide frequency moving on classical trajectory and obeying deterministic equation cannot explain the new species production. The sudden change and the stochastic property in speciation event should be understood from a broad view. Suppose the statistical state of the DNA evolution is represented by a wave function $\psi(\mathrm{x},t)$ that describes the probability amplitude ($|\psi|^2$ representing probability) of nucleotide frequency $\mathbf{x}$ ($\mathbf{x}=\{x_i\}$) at time t. The propagation of wave function is determined by

$$\psi(\mathrm{x},t) = \int_{-\infty}^{\infty} U(\mathrm{x},t;\mathrm{x}_0,t_0)\psi(\mathrm{x}_0,t_0)d\mathrm{x}_0 \tag{21}$$

where U is given by Eqs (4) and (2). From Ansatz 3 we can prove $\psi(\mathrm{x},t)$ satisfies Schrodinger equation (see Appendix 1)

$$\begin{aligned} iL\frac{\partial}{\partial t}\psi(\mathrm{x},t) &= H\psi(\mathrm{x},t) \\ H &= -\frac{L^2}{2c^2(t)}\sum\frac{\partial^2}{\partial x_i^2} - V(\mathrm{x},t) \end{aligned} \tag{22}$$

H is called Hamiltonian of the genome. The quantum evolutionary equation (22) is the logic generalization of the classical equation (1) (as f=0). The generalization is valid not only for the evolution in stable environment but also for the evolution in varying environment where the evolutionary potential V and inertia c^2 are time-dependent. As a stochastically generalization of the deterministic equation the quantum evolutionary theory is applicable in studying the new species formation.

In addition to the least action principle and Feynman's path integral approach the quantum generalization of the classic evolutionary equation can be deduced as well from Bohr's correspondence principle. Niels Bohr indicated that there exists a good correspondence between classic-mechanical motion and quantum motion in atomic physics. As the mass of a particle is small enough the classical mechanics should be replaced by the quantum laws. The similar correspondence exists in present case. The inertia parameter c^2 is time-dependent in general. During speciation the inertia of nucleotide frequency in the genome decreases largely and the classical motion will be changed to quantum. Therefore, from both approaches we can deduce the

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

stochastic evolutionary equation Eq(22) , the quantum generalization of the deterministic equation (1). The deterministic equation describes the classical phase of evolution while the Schrodinger equation describes the quantum phase of evolution.

We will discuss the relations between classical and quantum evolutions in more detail. The time evolution of the genome is described by the propagation function $U(\mathrm{x}',t';\mathrm{x}_0,t_0)=A\sum\exp(\frac{iS[\mathrm{x}]}{L})$ (Eq (4)). Assuming L is proportional to the average generation time λ for given species, one estimates L=(2000-3000) λ crudely from the human or bacterial speciation data . From Eq (2) the contribution of trajectory variation $\delta\mathrm{x}(t)$ to $\frac{1}{\boldsymbol{L}}\int(\frac{c^2}{2}\sum(\frac{dx_i}{dt})^2dt$ is proportional to c^2. So the variation of phase $\delta S/L$ in Eq (4) is much influenced by c^2. The large c^2 would make the contribution of different trajectories to the summation in (4), namely to $U(\mathrm{x}',t';\mathrm{x}_0,t_0)$, easily canceled each other, while the small c^2 does not. We assume that the inertia c^2 is a constant for genome moving on classical trajectory under stable environment. However, during speciation the genome sequence undergoes tremendous changes and the evolutionary inertia of nucleotide frequency is lost, c^2 jumping to a much lower value $(c')^2$. For example, one may assume $c'=10^{-3}c$ during speciation. The nucleotide changing rate $\frac{\boldsymbol{dx_i}}{\boldsymbol{dt}}$ is assumed to be $v=1/T$. In the time duration T the variation of phase for different paths in functional integral

$$\frac{\delta S}{L}\approx\frac{c^2}{2LT}. \tag{23}$$

We have estimated c=1100yr for birds in previous section. Taking L=2000 λ , λ=1yr and T=1yr one obtains $\delta S/L$=250. The phase variation is large enough and the strong interference makes only the classic trajectory retained. However, if the inertia is changed to $(c')^2=10^{-6}c^2$ then $\delta S/L$= 0.00025 and the non-classical path cannot be neglected definitely. It means the evolutionary picture should be switched from classical to quantum and the quantum evolution occurs during speciation. To avoid confusion, in the following we shall use c' to denote evolutionary inertia in the quantum theory, instead of c on classic trajectory.

The uncertainty relation is a fundamental characteristic of quantum system. By using wave function $\psi(\mathrm{x},t)$ one obtains the average of nucleotide frequency x_i

$$\langle x_i\rangle=\int\psi^*(x,t)x_i\psi(x,t)d^mx \qquad (d^mx=dx_1...dx_m) \tag{24}$$

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Define the conjugate of x_i as $p_i = -\boldsymbol{iL}\frac{\partial}{\partial \boldsymbol{x_i}}$. By using Schrodinger equation the average conjugate of frequency

$$\int \psi^*(x,t)\, p_i \psi(x,t) d^m x = c'^2 \frac{d}{dt}\int \psi^*(\mathrm{x},t) x_i \psi(\mathrm{x},t) d^m x$$
$$or \qquad \langle p_i \rangle_{av} = c'^2 \langle v_i \rangle_{av} \tag{25}$$

is deduced where v_i means the changing rate of nucleotide frequency. Moreover, by defining square deviations

$$\Delta x_i = \int \psi^*(\mathrm{x},t) x_i^2 \psi(\mathrm{x},t) d^m x - (\int \psi^*(\mathrm{x},t) x_i \psi(\mathrm{x},t) d^m x)^2$$
$$\Delta p_i = \int \psi^*(\mathrm{x},t)(p_i)^2 \psi(\mathrm{x},t) d^m x - (\int \psi^*(\mathrm{x},t) p_i \psi(\mathrm{x},t) d^m x)^2 \tag{26}$$

one easily obtains

$$\Delta x_i \Delta p_i \geq \frac{L^2}{4} \qquad (i = 1,\ldots,m) \tag{27}$$

or

$$\sqrt{\Delta x_i}\sqrt{\Delta v_i} \geq \frac{L}{2c'^2} \qquad (i = 1,\ldots,m) \tag{28}$$

for any given time. Here $\sqrt{\Delta x_i}$ means the uncertainty of nucleotide frequency and $\sqrt{\Delta v_i}$ the uncertainty of the changing rate of nucleotide frequency. Eq (27) (28) means the uncertainty relation of nucleotide frequency and its changing rate in DNA sequence. Generally the RHS of inequality (28) is not negligible. Combining Eq (28) and Eq (23) we find the less the variation of phase the stronger the uncertainty. Since x and v cannot be determined simultaneously to an enough accuracy, the evolutionary trajectory cannot be accurately defined in principle. The small c'^2 requires the simultaneous occurrence of the larger frequency deviation and the larger deviation of the frequency changing rate. This leads to the nucleotide frequency and the frequency changing rate no longer simultaneously measurable to an enough accuracy. Therefore, the picture of classical definite trajectory should be replaced by a large amount of rapid trajectory-transitions. Moreover, the uncertainty of nucleotide frequency $\sqrt{\Delta x_i}$ can be understood as the single nucleotide polymorphism. So Eq (28) gives a lower bound of the uncertainty for the changing rate of nucleotide frequency.

Now we shall discuss the quantum state of genome evolution. As an example, consider the avian genome evolution. Start from Eq (22) and assume $V(\mathrm{x},t)$ given by (5)-(8). The quantum state described by the eigenstate of Hamiltonian is generally not continuously but discrete. To be definite study the single-nucleotide frequency only. The eigenstate of H satisfies

$$H\psi(\mathrm{x}) = E\psi(\mathrm{x}) \tag{29}$$

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Hamiltonian given by Eq (22) can be rewritten as

$$H = H_0 + H'$$

$$H_0 = -\frac{L^2}{2c'^2}\sum_i \frac{\partial^2}{\partial x_i^2} - (2+\alpha)N + \beta N^2 \tag{30}$$

$$= -\frac{2L^2}{c'^2}\sum_i \frac{\partial^2}{\partial N^2} - (2+\alpha)N + \beta N^2$$

$$H' = NR_I = \frac{2}{N\ln 2}\sum_i x_i^2 - \frac{N}{2\ln 2} \tag{31}$$

where Eqs (5)-(8) and (10) have been used. H' is a small quantity for avian genome (see the note after Eq (10)) which can be looked as a perturbation. Following perturbation theory, setting $\psi(\mathrm{x}) = \psi^{(0)}(N) + \psi^{(1)}(\mathrm{x})$, the zero order wave function $\psi^{(0)}$ depends on N only, satisfying

$$\{\frac{-2L^2}{c^2(t_0)}\frac{\partial^2}{\partial N^2} - (2+\alpha(t_0))N + \beta(t_0)N^2\}\psi^{(0)}(N) = E^{(0)}\psi^{(0)}(N) \tag{32}$$

The parameters α, β and c take values at time t_0. The solution of Eq (32) is

$$E^{(0)} = E_n^{(0)} = (n+\frac{1}{2})\sqrt{8\beta}\frac{L}{c'} - \frac{(2+\alpha)^2}{4\beta} \qquad (n=0,1,2,\ldots) \tag{33}$$

$$\psi^{(0)}(N) = \psi_n^{(0)}(N) = C_n \exp(-\frac{\xi^2}{2})H_n(\xi)$$

$$H_0(\xi) = 1 \quad H_1(\xi) = 2\xi \quad H_2(\xi) = 4\xi^2 - 2, \ldots \tag{34}$$

$$\xi = \sqrt{\frac{c'}{L}}(\frac{\beta}{2})^{1/4}(N - \frac{2+\alpha}{2\beta})$$

The first-order correction to $E_n^{(0)}$ is

$$E_n^{(1)} = H'_{nn} = \int d^4x \psi_n^{(0)} H' \psi_n^{(0)} \quad , \quad d^4x = dx_1 dx_2 dx_3 dx_4 \ . \tag{35}$$

The first-order correction to wave function $\psi^{(0)}(N)$ can also be calculated from perturbation theory.

Eq (33) shows the Hamiltonian-level is equally spaced by $\sqrt{8\beta}\frac{L}{c'}$. If c' is replaced by the classical value c, estimated by 1100 yr for avian genome in previous section, then the spacing $\sqrt{8\beta}\frac{L}{c}$ is a small quantity. So the eigenstates are basically continuous in classical phase of the evolution. However, during speciation the evolutionary inertia of nucleotide frequencies drops to a lower value as if in this time all evolutionary events happened more rapidly. If c' as a parameter of time dimension decreases to 10^{-3} of the classical value, $c' = 10^{-3}c$, then the picture of definite trajectory, namely $x_i(t)$ as a function of t, ceases to be correct and the state will be switched to

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

quantum. A series of discrete eigenstates occur in quantum phase. In the quantum state the nucleotide frequency always takes some statistical distribution but not a definite value. Eq (34) shows the statistical distribution of N in ground state (n=0) peaks at $N=\frac{2+\alpha}{2\beta}\approx\frac{1}{\beta}$ with width equal $\sqrt{\frac{2}{\beta}}\frac{L}{c'}$.

Therefore, the quantum theory regards that the speciation event is essentially a quantum transition between initial "old" species and final "new" species. There always exists a Hamiltonian-level gap between low-lying ground state (n=0) and excited state (n=1,2,…), that can be seen from the large spacing $\sqrt{8\beta}\frac{L}{c'}$ in a quantum genome. In fact, the gap of ground state should be deeper than $\sqrt{8\beta}\frac{L}{c'}$ if the higher- order terms of N^3 etc have been taken into account in the environmental potential equation (8) of W_{env}. Due to the deep gap occurring in the Hamiltonian ground level one should consider only the transition between ground states in studying speciation event.

4 Speciation studied from theory of quantum transition

In the following we shall discuss the speciation event from the view of quantum transition. Based on Schrodinger equation the speciation rate can be calculated. Suppose the initial wave function of the "old" species denoted by $\psi_I(\mathrm{x})$ satisfying Eqs (32)-(34) and the final wave function of the "new" species by $\psi_F(\mathrm{x})$ satisfying the same equation but $c'(t_0), \beta(t_0)$, $\alpha(t_0)$ replaced by $c'(t_F), \beta(t_F)$, $\alpha(t_F)$ respectively. The transition from "old" to "new" is caused by a time-dependent interaction $H_{\mathrm{int}}(t)$ in the framework of quantum mechanics. One may assume H_{int} comes from the time variation of evolutionary inertia and environmental potential , namely

$$H_{\mathrm{int}}(t)=H(c'=c'(t),\beta=\beta(t),\alpha=\alpha(t))-H(c'=c'(t_0),\beta=\beta(t_0),\alpha=\alpha(t_0))$$

$$=\frac{\partial H}{\partial c'}\Delta c'+\frac{\partial H}{\partial \beta}\Delta\beta+\frac{\partial H}{\partial \alpha}\Delta\alpha\,. \tag{36}$$

Thus the transitional probability amplitude is expressed by [27]

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$
\begin{aligned}
T_{FI} &= \frac{-i}{L_I}\int \psi_F^*(\mathrm{x})\int_{t_0}^{t_F} dtH_{\mathrm{int}}(t)\exp\{i(\frac{E_F}{L_F}-\frac{E_I}{L_I})(t-t_0)\}\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4 \\
&= \frac{-i}{L_I}\int \psi_F^*(\mathrm{x})\int_{t_0}^{t_F} dtH_{\mathrm{int}}(t)\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4 \\
&= \frac{-i}{L_I}\int \psi_F^*(\mathrm{x})\int_{t_0}^{t_F} dt\{(\frac{\partial H}{\partial c'}\frac{dc'}{dt}+\frac{\partial H}{\partial \beta}\frac{d\beta}{dt}+\frac{\partial H}{\partial \alpha}\frac{d\alpha}{dt})(t-t_0\}\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4 \\
&= \frac{-i}{L_I}\int \psi_F^*(\mathrm{x})\{H(t_F)(t_F\text{-}t_0)\text{-}\int_{t_0}^{t_F} dtH(t)\}\,\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4
\end{aligned}
\tag{37}
$$

where

$$
\frac{t_F-t_0}{L_F}<<1, \qquad \frac{t_F-t_0}{L_I}<<1
$$

(L_I and L_F are the quantization constant of initial and final genomes respectively) has been taken into account and the partial integration has been used. On the other hand,

$$
\begin{aligned}
T_{FI} &= \frac{-i}{L_I}\int \psi_F^*(\mathrm{x})\int_{t_0}^{t_F} dtH_{\mathrm{int}}(t)\exp\{i(\frac{E_F}{L_F}-\frac{E_I}{L_I})(t-t_0)\}\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4 \\
&= \frac{-i}{L_I}\int \psi_F^*(\mathrm{x})\int_{t_0}^{t_F} dtH_{\mathrm{int}}(t)\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4 \\
&= \frac{-i}{L_I}\int \psi_F^*(\mathrm{x})\int_{t_0}^{t_F} dt\{H(t)\text{-}H(t_0)\}\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4 \\
&= \frac{-i}{L_I}\int \psi_F^*(\mathrm{x})\{\int_{t_0}^{t_F} dtH(t)\text{-}E_I(t_F\text{-}t_0)\}\,\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4
\end{aligned}
\tag{38}
$$

So,

$$
T_{FI} = \frac{-i}{2L_I}(E_F-E_I)(t_F-t_0)\int \psi_F^*(\mathrm{x})\psi_I(\mathrm{x})dx_1dx_2dx_3dx_4 \tag{39}
$$

E_I and E_F - the eigenvalue of H, given by Eq (33) with n=0 where parameters c', β, α take values at t_0 and t_F respectively . The transitional probability

$$
|T_{FI}|^2 = \frac{(t_F-t_0)^2}{4L_I^2}\{(E_F-E_I)_1+(E_F-E_I)_2\}^2(\int \psi_F^*(N)\psi_I(N)dN)^2 \tag{40}
$$

is determined by three factors, namely

$$
(E_F-E_I)_1 = \sqrt{2\beta_F}\,\frac{L_F}{c'_F}-\sqrt{2\beta_I}\,\frac{L_I}{c'_I} \tag{41}
$$

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$(E_F - E_I)_2 = \frac{(2+\alpha_I)^2}{4\beta_I} - \frac{(2+\alpha_F)^2}{4\beta_F} \tag{42}$$

and the wave function integral $\int \psi_F^*(N)\psi_I(N)dN$. By using

$$\psi_I(N) = \pi^{-1/4} a_I^{-1/4} \exp(-\frac{(N-N_I)^2}{2a_I})$$

$$\psi_F(N) = \pi^{-1/4} a_F^{-1/4} \exp(-\frac{(N-N_F)^2}{2a_F}) \tag{43}$$

$$a_{I,F} = \sqrt{\frac{2}{\beta_{I,F}}} \frac{L_{I,F}}{c'_{I,F}} \qquad \boldsymbol{N_{I,F}} = \frac{2+\alpha_{\boldsymbol{I,F}}}{2\beta_{\boldsymbol{I,F}}} \tag{44}$$

($a_{I,F}$ - the frequency distribution width, and $N_{I,F}$ the frequency distribution centers for two genomes respectively, that can be found in Eq (34), the overlap integral can be calculated and it gives

$$\int \psi_F^*(N)\psi_I(N)dN = \sqrt{2}(a_I a_F)^{1/4}(a_I + a_F)^{-1/2} \exp\{-\frac{1}{2}(\frac{N_F^2}{a_F} + \frac{N_I^2}{a_I} - \frac{a_F a_I}{a_F + a_I}(\frac{N_F}{a_F} + \frac{N_I}{a_I})^2)\} \tag{45}$$

The transition time t_F-t_0 can be obtained by the normalization of the total transition probability from initial (I) to all possible final (F) states, namely

$$\sum_F |T_{FI}|^2 = 1 \tag{46}$$

Under stable environment , $\beta_I = \beta_F$, and $\alpha_I = \alpha_F$, one has $(E_F - E_I)_2 = 0$. The transition is determined only by $(E_F - E_I)_1$ and the overlap integral. However, under varying environment, considering $\beta \cong \frac{1}{N}$ is a small quantity one has $(E_F - E_I)_1 << (E_F - E_I)_2$ generally. This explains why new species formation occurs most probably in the condition of varying environment. For example, a long-term experiment with *Escherichia coli* was made where the population evolved with glucose as a limiting nutrient. There was only a hypermutable phenotype appeared by about generation 26500.[28] However, in our experiments with *Escherichia coli* by repetitious ion irradiation, the colony polymorphisms of mutants were observed in the very early stage. Therefore, ion irradiation as a varying environment can effectively accelerate the genome evolution of *E. coli*.[29] . No loss of generality, in the following calculation we assume

$$\alpha_I = \alpha_F = \alpha, \text{ but } \beta_I \neq \beta_F$$

$$\beta_I = \frac{2+\alpha}{2N_I} \qquad \beta_F = \frac{2+\alpha}{2N_F}, \quad (N_I \neq N_F) \tag{47}$$

On the other hand the exponent factor in Eq (45)

$$\exp\{-\frac{1}{2}(\frac{N_F^2}{a_F} + \frac{N_I^2}{a_I} - \frac{a_F a_I}{a_F + a_I}(\frac{N_F}{a_F} + \frac{N_I}{a_I})^2)\} = \exp\{-\frac{(N_I - N_F)^2}{4\bar{a}}\} \tag{45.1}$$

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

($\bar{a}$ is some average of a_F and a_I) that shows the wave function integral or the transition is not negligible only when N_F near N_I which means β_F near β_I , the variation of environmental parameter being small during quantum transition. The point is understandable because $t_F - t_0$ is a small quantity and the changing environment only causes a small variation of environmental parameter β. Moreover, Eq. (45.1) shows the transition probability is irrespective of the symbol of $N_I - N_F$. The sequence length of posterity genome can be longer , or shorter as well, than ancestry in quantum transition.

If the quantum transition occurs in stable environment, assuming $c'_I = c'_F$, from (41) and (42) one has

$$E_F - E_I = \frac{\sqrt{2\beta}}{c'}(L_F - L_I)$$

Set L proportional to the generation time λ. One obtains

$$E_F - E_I \propto \lambda_F - \lambda_I \tag{48}$$

For several posterity genomes produced from one ancestry the branching ratio of transitional probability between different channels is proportional to $(\lambda_F - \lambda_I)^2$ if the difference between the overlap integrals can be neglected. However, the speciation event often happens in varying environment and in this case the initial and final environmental parameters are different and one should use the original equation (40) to calculate the transitional rate.

The rapid post-Cretaceous radiation of neoavian birds provides a vast amount of experimental data to test the above quantum theory on new species production. Although different evolutionary trees -- for example, trees published in Science [12], in Nature [13] and IMU tree[16] -- show the basically same sister relationships among clades the earliest branches in Neoaves during the rapid radiation after the Cretaceous-Paleogene mass extinction event have quite different topology. We study 12 kinds of birds whose branching time was between 66 Mya to 50 Mya, namely Common ostrich, Carmine bee-eater, Rifleman, Sunbittern, Hoatzin, Mac Queen's bustard, Cuckoo-roller, Barn owl, White-tailed tropicbird, Red-throated loon, Crested ibis and Penguin, (denoted as bird 1 to bird 12 respectively) and assume that they were produced from a quantum speciation event from one ancestry bird (denoted as Anc) at time t between t_0 and t_F. Then, these genomes of posterity birds follow their classical trajectories respectively after the quantum transition (Fig 2). The ancestry bird Anc is assumed to have the same characteristics, inertia c' and generation time λ , as the posterity bird 1, Common ostrich . The peak of the Anc genome sequence length is supposed at pN_1 (p is a fraction 1/4,1/3 or 1/2 etc , N_1 is the present genome length of Aves 1). The peak of the i-th posterity genome at time of quantum transition is supposed at (referred to Fig 2)

$$N_i' = pN_1 + 10^5 \times \frac{N_i - N_1}{N_1} \qquad (i = 1,\ldots,12) \tag{49}$$

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where N_i means the present sequence length of the *i*-th genome. By using Eqs (40)-(47) we calculate the transitional probabilities of new avian production and the results are summarized in Table 1.

Table 1 New avian production probability from ancestry Anc

p	Tm(yr)	Aves1	Aves 2	Aves 3	Aves 4	Aves 5	Aves 6	Aves 7	Aves 8	Aves 9	Ave10	Ave11	Ave12
1/4	0.88	0.1069	0.0459	0.0375	0.0731	0.1038	0.0718	0.0863	0.0745	0.1003	0.0903	0.1044	0.1051
1/2	0.85	0.1015	0.0533	0.0462	0.0770	0.0989	0.0761	0.0857	0.0752	0.0971	0.0895	0.0993	0.1003

Tm means transition time t_F-t_0. Transition probabilities for 12 posterity Aves are listed in the table. In calculation c'=c/1000 and L=3000 λ are assumed. The environmental parameter α =0.18 is taken, β_I and β_F are calculated from eq (47) where N_I=pN_1 and N_F equal the average sequence length $N_i^{'}$ of the posterity Aves. (p=1/4 corresponds to classical trajectory given in Fig 2; for reference the case of p=1/2 is also calculated and listed).

Although the rapid post-Cretaceous radiation of neoavian birds arouses many difficulties in reconstructing avian evolutionary tree we found the quantum evolution gives new insight to the understanding of the complexity in the problem. Moreover, several general laws on new species production can be deduced from the example of Aves genomes:

1) The speciation is a stochastic event, obeying a statistical law. Generally an ancestry produces several posterities with different probabilities.

2) The transition time $t_F - t_0$ is very short as compared with parameter *L*, $\frac{t_F - t_0}{L}$ being about 1/1000 as seen from Table 1. The result is consistent with the concept of quantum transition.

3) The quantum transition preferably occurs in varying environment but due to the transition completed in a short time, the change of environmental parameter during speciation is small. In above example of Aves the difference between the initial and final environmental parameters, β_I and β_F, is lower than 10^{-5} .The calculation holds irrespective of different choices of parameters *L* and *c'* (for example, *L*=2000 λ to 3000 λ and *c'* =*c*/1000 to *c*/2000).

4) Both ancestry and posterity genomic sequence lengths obey statistical distributions. The peaks of two distributions in initial and final states of quantum transition are near to each other; otherwise, the transition rate is too small and cannot be observed.

Conclusions

It is interesting to make comparison between the present quantum model for genome and the conventional quantum theory for electrons, atoms and molecules (even for biological macromolecules [30]). Both systems have wave function satisfying Schrodinger equation. Both theories have their classical limit. The classical trajectory of an atom or an electron is given by the position coordinate of the particle as a function of *t* , while the classical trajectory of a genome is given by the nucleotide frequency of DNA as a function of *t* . The former is constrained by particle's energy while the latter is constrained by genome's information. Energy is conserved with time but information always grows in evolution. The atomic quantum theory has an elementary constant, the Planck's constant, while the genomic quantum theory has a

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

corresponding quantization constant *L*. The Planck's constant is universal and has dimension of energy but the genome constant *L* is species-dependent and has dimension of time. The classical limit of both theories can be deduced from Planck constant *h* or genome constant *L* approaching to zero respectively. However, the classical limit of atomic quantum theory is a new mechanics for macroscopic particle of large mass. While the genome inertia *c* is changeable in the course of evolution. For the same genome it changes thousand-fold in different phases of evolution. Two constants, quantization constant *L* and evolutionary inertia *c*, both in the dimension of time, play important roles in the present quantum evolutionary theory. The former is related to the realization of the quantum picture of the evolution, while the variation of the latter gives rise to the switch from classical phase to quantum phase or *vice versa*.

Acknowledgement The author is indebted to Dr Qi Wu for numerous helpful discussions. He also thanks Drs Liron Zhang, Yongfen Zhang and Xun Zhou for their contributions in avian genome analyses and Dr Yulai Bao for his kind help in literature searching.

Appendix 1 Deduction of wave equation in quantum phase

Suppose the statistical state of $\mathrm{x}=(x_1,\ldots x_m)$ represented by a wave function $\psi(\mathrm{x},t)$.

$$\psi(\mathrm{x},\varepsilon)=\int U(\mathrm{x},\varepsilon;\mathrm{x}_0,0)\psi(\mathrm{x}_0,0)d\mathrm{x}_0 \qquad (\varepsilon\text{- a small quantity}) \qquad \text{(A1)}$$

where *U* is a functional integral, expressed as

$$\begin{aligned} U(\mathrm{x}',t';\mathrm{x}_0,t_0) &= A\int \exp(\frac{iS[\mathrm{x}]}{L})d(\mathrm{x}) \\ &= \lim_{n\to\infty} A\int d\mathrm{x}_1\ldots\int d\mathrm{x}_{n-1}\exp(\frac{iS(\mathrm{x}_1,\ldots\mathrm{x}_n)}{L}) \end{aligned} \qquad \text{(A2)}$$

and $t'-t_0=\varepsilon$ has assumed. Inserting the information action functional $S[\mathrm{x}]$ (Eq (5))

$$S[\mathrm{x}(t)]=\int(\frac{c^2(t)}{2}(\frac{d\mathrm{x}}{dt})^2+V(\mathrm{x},t))dt \qquad \text{(A3)}$$

into (A2), after integrating over $(\mathrm{x}_1,\ldots\mathrm{x}_{n-1})$ and taking limit $n\to\infty$ we obtain [10]

$$U(\mathrm{x},\varepsilon;\mathrm{x}_0,0)=(\frac{c^2(0)}{2\pi Li\varepsilon})^{m/2}\exp\frac{i}{L}\{\frac{c^2(0)(\mathrm{x}-\mathrm{x}_0)^2}{2\varepsilon}+\varepsilon V(\frac{\mathrm{x}+\mathrm{x}_0}{2},0)\} \qquad \text{(A4)}$$

Eq (A4) has been normalized to $U\to\delta(\mathrm{x}-\mathrm{x}_0)$ as *V*=0. In the deduction of (A4) the integral formula

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\int_{-\infty}^{\infty} dy_1 ... \int_{-\infty}^{\infty} dy_{n-1} \exp\{-\sum_{j=0}^{n-1} \frac{1}{i}(y_{j+1} - y_j)^2\} = \frac{1}{\sqrt{n}}(i\pi)^{\frac{n-1}{2}} \exp\{\frac{-1}{ni}(y_n - y_0)^2\} \tag{A5}$$

has been used. By using (A4), Eq (A1) can be written as

$$\psi(\mathrm{x},\varepsilon) = (\frac{c^2(0)}{2\pi Li\varepsilon})^{m/2})\int \exp\frac{ic^2(0)(\mathrm{x}\text{-}\mathrm{x}_0)^2}{2\varepsilon L}\exp\{\frac{i\varepsilon}{L}V(\frac{\mathrm{x}+\mathrm{x}_0}{2},0)\}\psi(\mathrm{x}_0,0)d\mathrm{x}_0 \tag{A6}$$

Here $\exp\frac{ic^2(\mathrm{x}\text{-}\mathrm{x}_0)^2}{2\varepsilon L}$ is a rapidly oscillating factor, only important near $\mathrm{x}\text{-}\mathrm{x}_0=0$. By setting $\mathrm{x}_0\text{-}\mathrm{x} = \eta$, Eq (A6) is rewritten as

$$\psi(\mathrm{x},\varepsilon) = (\frac{c^2(0)}{2\pi Li\varepsilon})^{m/2})\int \exp\frac{ic^2(0)\eta^2}{2\varepsilon L}\exp\{\frac{i\varepsilon}{L}V(\mathrm{x}+\frac{\eta}{2},0)\}\psi(\mathrm{x}+\eta,0)d\eta \tag{A7}$$

After expanding $\psi(\mathrm{x}+\eta,0)$ and $\exp\frac{i\varepsilon}{L}V(\mathrm{x}+\frac{\eta}{2},0)$ with respect to η and ε, finally we obtain

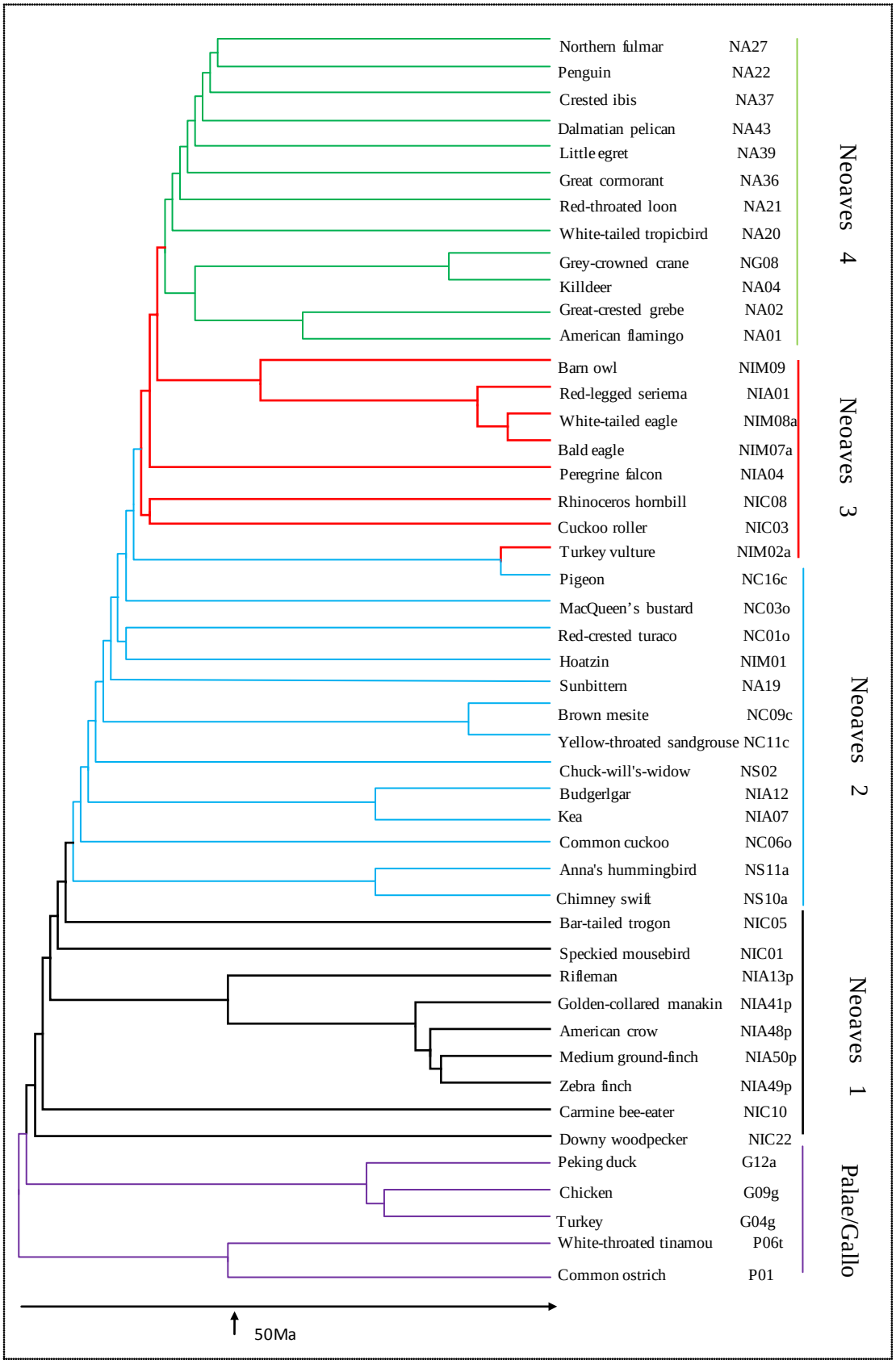
$$\psi(\mathrm{x},\varepsilon) = \psi(\mathrm{x},0) - \frac{i\varepsilon}{L}\{-\frac{L^2}{2c^2(0)}\sum\frac{\partial^2}{\partial x_i^2} - V(\mathrm{x},0)\}\psi(\mathrm{x},0) \tag{A8}$$

Therefore, $\psi(\mathrm{x},t)$ satisfies Schrodinger equation,

$$iL\frac{\partial}{\partial t}\psi(\mathrm{x},t) = \{-\frac{L^2}{2c^2(t)}\sum\frac{\partial^2}{\partial x_i^2} - V(\mathrm{x},t)\}\psi(\mathrm{x},t) \tag{A9}$$

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Appendix 2 Phylogenetic tree of 47 representative birds



bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure A1 Phylogenetic tree deduced by KNA algorithm

The evolutionary tree of 47 birds was deduced in [16] that can be compared with trees published in [12] [13].

Appendix 3 Notations and parameters of 47 Aves

Table A1 Parameters of 47 Aves used in theory of genome evolution

	Notation	Common name	N	λ(yr)	c(yr)
1	P01	Common ostrich	1228437082	3	1100
2	P06t	White-throated tinamou	1059687971	2	1184
3	G04g	Turkey	1061982190	1.5	1183
4	G09g	Chicken	1108466630	1	1158
5	G12a	Peking duck	1105049323	1	1160
6	NIC22	Downy woodpecker	1174542329	1	1125
7	NIC10	Carmine bee-eater	1066568727	1	1180
8	NIA49p	Zebra finch	1235794146	1	1097
9	NIA50p	Medium-ground finch	1072489490	1	1177
10	NIA48p	American crow	1095175346	1	1165
11	NIA41p	Golden-collared manakin	1160253066	1	1132
12	NIA13p	Rifleman	1044655775	1	1193
13	NIC01	Speckied mousebird	1086303309	1	1169
14	NIC05	Bar-tailed trogon	1085711737	1	1170
15	NS10a	Chimney swift	1125117173	3.5	1149
16	NS11a	Anna's hummingbird	1114626264	1	1155
17	NC06o	Common cuckoo	1159208195	1.5	1132
18	NIA07	Kea	1062026465	3	1183
19	NIA12	Budgerlgar	1117358947	3	1153
20	NS02	Chuck-will's-widow	1145149681	1	1139
21	NC11c	Yellow-throated sandgrouse	1079547375	1	1173
22	NC09c	Brown mesite	1101046125	1	1161
23	NA19	Sunbittern	1098400457	2	1163
24	NIM01	Hoatzin	1209082675	2	1109
25	NC01o	Red-crested turaco	1175440150	2	1125
26	NC03o	Mac Queen's bustard	1095295809	2	1165
27	NC16c	Pigeon	1111581692	1	1156
28	NIM02a	Turkey vulture	1179513848	2	1123
29	NIC03	Cuckoo-roller	1148076042	1.5	1139
30	NIC08	Rhinoceros hornbill	1105159377	3.5	1160
31	NIA04	Peregrine falcon	1174046505	2	1125
32	NIM07a	Bald eagle	1258793725	4	1087

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

33	NIM08a	White-tailed eagle	1144733688	4	1140
34	NIA01	Red-legged seriema	1145705463	1	1139
35	NIM09	Barn owl	1138815065	1	1142
36	NA01	American flamingo	1144756512	3	1139
37	NA02	Great-crested grebe	1152104262	2	1136
38	NA04	Killdeer	1225129931	1	1101
39	NG08	Grey-crowned crane	1138861064	4	1142
40	NA20	White-tailed tropicbird	1166684827	3.5	1129
41	NA21	Red-throated loon	1145076370	2	1139
42	NA36	Great cormorant	1154672768	1	1135
43	NA39	Little egret	1211513874	2	1108
44	NA43	Dalmatian pelican	1168624213	3	1128
45	NA37	Crested ibis	1240309383	2	1095
46	NA22	Penguin	1257483768	3.5	1087
47	NA27	Northern fulmar	1153271365	1.5	1135

The notations of 47 birds are given by the following rules. Referred to 198 birds in ref [13], several letters are introduced to indicate the taxon of each bird. P01-P09 referred to Palaeognathae, G01-G16 referred to Galloanserae, NS01-NS13 referred to Neoaves- Strisores, NC01-NC16 referred to Neoaves-Columbaves, NG01-NG09 referred to Neoaves-Gruiforms, NA01-NA43 referred to Neoaves-Aequorlitornithes, NIM01-NIM10 referred to Neoaves-Inopinaves-Accipitriformes and related, NIC01-NIC26 referred to Neoaves-Inopinaves-Coraciimorphae, NIA01-NIA56 referred to Neoaves-Inopinaves-Ausralaves. The last lower-case letter indicates a subclass, for example, the letter p in NIA13p-NIA56p means the subclass passeriformes. These notations are also used in Fig A1. N means the genome sequence length. λ means the generation time. **C** means the evolutionary inertia, calculated from the analysis of the classical trajectory of the genome evolution, see text.

References

[1] Gregory TR. 2005. Genome size evolution in animals. In: *Evolution of the Genome* (Edi. Gregory TR), Elsevier Inc.

[2] Gregory TR et al. 2007. Eukaryotic genome size databases. Nucleic Acids Research **35,** Database issue D332-D338.

[3] Luo LF. 2009. Law of genome evolution direction: Coding information quantity grows. *Front. Phys.* **4**(2): 241–251.

[4] Luo LF. 2004. *Theoretic-Physical Approach to Molecular Biology*. Shanghai Scientific & Technical Publishers, Shanghai.

[5] Ridley M. 2004. *Evolution* (3rd edition). Blackwell Publishing .

[6] Feduccia A. 1995. Explosive evolution in tertiary birds and mammals. Science. **267**: 637-638.

[7] Hedges SB, Parker PH, Sibley CG, Kumer S. 1996. Continental breakup and the ordinal diversification of birds and mammals. Nature. **381**:226-229.

[8] Cooper A, Penny D. 1997. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. Science.**275**:1109-1113.

[9] Simon YW H, Matthew J P. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Sym. Biol. **58**(3):367-380.

bioRxiv preprint doi: https://doi.org/10.1101/034710; this version posted December 28, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[10] Romanov MN et al. 2014. Reconstruction of gross avian genome structure, organization and evolution suggests that the chicken lineage most closely resembles the dinosaur avian ancestor. BMC Genomics **15**:1060.

[11] Zhang GJ et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaption. Science. **346**:1311-1319.

[12] Jarvis ED et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. **346**:1320-1331.

[13] Prum RO et al. 2015. A comprehensive phylogeny of birds (aves) using targeted next-generation DNA sequencing. Nature. **526**:569-573.

[14] Luo LF Gao Y, Lu J. 2010. Information-theoretic view of sequence organization in a genome. arXiv 1004.3843 [q-bio.GN]

[15] Zhou FF, Olman V, Xu Y. 2008. Barcodes for genomes and applications. BMC Bioinformatics. **9**:546.

[16] Zhang YF, Zhou X, Luo LF, Zhang LR. 2016. Reconstruction of evolutionary tree for avian genomes. International Symposium on Big Data of Science, Baotou.

[17] Feynman RP, Hibbs A. 1965. *Quantum Mechanics and Path Integrals*. McGraw Hill, NY.

[18] Laxton RR. 1978. The measure of diversity. J Theor Biol. **70**(1):51-67.

[19] Zhang LR, Luo LF. 2003. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Res. **31**(21):6214-6220.

[20] Lu J, Luo LF. 2007. Predicting human transcription starts by use of diversity measure with quadratic discriminant. *Computation in Modern Science and Engineering: Proceedings of the International Conference on Computational Methods in Science and Engineering* 2007 *(ICCMSE 2007), AIP Conf Proc*. **963**:1273-1277.

[21] Feng YE, Luo LF. 2008. Use of tetrapeptide signals for protein secondary-structure prediction. Amino Acids. **35**(3): 607–614.

[22] Chen W, Luo LF, Zhang LR . 2010. The organization of nucleosomes around splice sites. Nucleic Acids Research **38**:2788-2798.

[23] Kauffman S. 2000 *Investigations*. Oxford University Press, Inc,

[24] Gingerich PD. 1983. Rates of evolution:effects of time and temporal scaling. Science **222**: 159-161

[25] Hendry AP, Kinnison MT. 2001. *Microevolution: rates, pattern, process*. Kluwer Academic, Dordecht, Netherland

[26] Eldredge N , Gould SJ. 1972. Punctuated equilibria: an alternative to phyletic gradualism. In Schopf TJM (ed) *Models in Paleobiology*. Freeman Cooper & Co, San Francisco .

[27] Landau LD, Lifshitz EM. 1958 *Quantum Mechanics*. Oxford: Pergamon Press

[28] Barrick JE, *et al* . 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature **461**: 1243-U1274

[29] Song ZQ，Luo LF. 2012. Escherichia coli mutants induced by multi-ion irradiation. Journal of Radiation Research **53**:854–859, doi: 10.1093/jrr/rrs061

[30] Luo LF. 2014. Quantum theory on protein folding. Sci China – Phys Mech Astron, **57**:458-468, doi: 10.1007/s11433-014-5390-8 ; Luo LF., Lv J. 2016. Quantum conformational transition in biological macromolecule. Quantitative Biology. doi: 10.1007/s40484-016-0087-9.