Speciation as a Phase Transition

Georgios Tsekenis

Physics Department, University of Illinois at Urbana-Champaign,
Urbana, Illinois, 61801, USA

(Dated, December 13, 2005)

Abstract

In the most recent theory of cellular evolution, the mechanisms of incorporating foreign pieces in an organisms DNA sequence play an extremely important role. The emergence of species has to be perceived as a phase transition. Exact simulations on theoretical models justify that notion and strengthen the argument of the evolution being communal. Extended versions of those models will probably reveal more details clarifying more intuitions.

I. Introduction

Importance of Horizontal Gene Transfer

The most recent theory of evolution of the living organisms, as strongly presented by Woese (ref.1), decisively incorporates the transfer of genetic material between organisms that do not have the parental-offspring relationship. Horizontal Gene Transfer (HGT) is to be considered as important as the familiar Vertical Gene Transfer (VGT) that happens from a parent organism to an offspring one. HGT is evidenced to be able to affect the entire genome and with a considerable impact. This fresh look on the century-and-a-half old subject was triggered by detailed results of the cellular processes based on the decoded genome sequences.

Communality of Evolution

From the specifics of HGT it can be inferred that some level of compatibility (but not complete identity) between donor and recipient is needed for it to be successful. That factor renders HGT specific to the members of a community. Indeed, from experimental evidence there is little to no doubt that the evolution at least for primitive, simple organisms is communal. The size though of the pool that one organism can potentially receive genetic material from depends on the complexity of its structure. And this is because the less trivially a cell operates the more individually specific it becomes. In the same direction the level of incompatibility with other organisms that are outside its class, increases and the probability of successful HGT decreases.

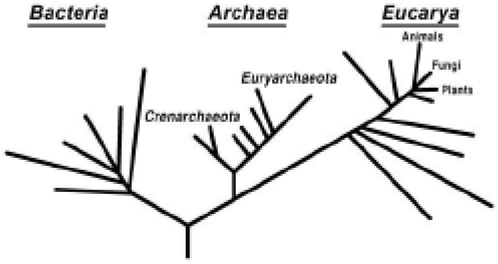

Fig. 1. The universal phylogenetic tree (ref.1). Explicitly is depicted the root and the common evolution line until the various points where the Darwinian threshold is achieved and through a phase transition distinct species emerge proceeding with their own communal evolution to other Darwinian points.

Universal Phylogenetic Tree

Along this very line of thinking, Woese (ref.1) proposed the universal phylogenetic tree originating not from Darwin's still elusive common ancestor but from a common root (fig. 1). At those early stages of simple organismal structure, the different organisms would be roughly the same, just because they were simple. Genetic material could be transferred freely enough. Organisms acquiring the most beneficial changes would easily distribute the beneficial genetic material to others since it was well-adapted and could pass through the defense mechanisms, which would have to be based on the less evolved DNA. Allowing sufficient evolutionary time, changes should have accumulated to a subset of organisms making them evolve apart from the rest, with geographical barriers and natural selection not being trivial factors. At a later time another level of complexity must have been reached and another separation should have occurred. Current analysis of genome data and of other cellular processes, reveal three major forms of life, assuming that any extinction could have happened to a minor or a major branch of the universal phylogenetic tree (ref. 1 & fig. 1).

The Phase Transition

The point where a group of organisms has evolved sufficiently to start evolving apart from the original community was named Darwinian Threshold incisively enough by Woese (ref. 1). At the same time though the notion of the evolving community being separated to smaller individually evolving sub-communities renders the common ancestor idea obsolete and the meaning of a common root more feasible. There is a phase transition that spans the entire commonly evolving community at that point and the emergence of many smaller sub-communities to feed themselves to the same process of collective evolution and unavoidable dispersion. In a similar way modern cellular life must have derived itself from the first ever occurrence of a Darwinian Threshold and the first ever phase transition taking place (fig. 1).

II. Mechanisms of Genetic Alteration

Acquisition Processes

A cell can acquire foreign genetic material from the environment (transformation), via a virus (transduction) or from another cell by direct contact (conjugation). The foreign fragment will be incorporated to the recipient's genome through a process known as recombination (ref. 2).

Homologous Recombination

The homologous recombination is performed by the organism itself and helps regulate the process so that the received DNA fragment will replace the analogous part of the own DNA by demanding sufficient identity. The probability for a homologous recombination to happen is,

$$P \propto e^{-\alpha d} \qquad (1)$$

where d is the divergence[1] between the fragment to be replaced and the fragment to replace and α the minimal length of uninterrupted sequence the new piece has to have with the old so that it is allowed to replace it (ref. 2,3).

Illegitimate Recombination or HGT

On the other hand there is illegitimate recombination or HGT. In general this term is used for events that alter the DNA sequence of an organism but are not controlled by it. It can range from genetic material delivered by bactiriophage integrases to random breakage and repair which can be as small as affecting only one base in which case it is called a point mutation. The fact that the received DNA piece did not go through a cellular acceptation process can allow it be very different from the piece it replaced (ref.2).

[1] The divergence between two DNA sequences is directly related to number of different base pairs relatively to the total number of available pairs for recombination.

III. Speciation as a Phase Transition

Competing Mechanisms

Among the different mechanisms of transferring genetic material between organisms there are two that are clearly competing as was noticed by Vestigian and Goldenfeld (ref. 2). Recombination tends to keep the DNA sequences the same while mutations can cause any king of change at any point of the sequence. What is more, point mutations can inhibit recombination from some level of divergence and above imposing strong genetic boundaries between the cells (fig 2).

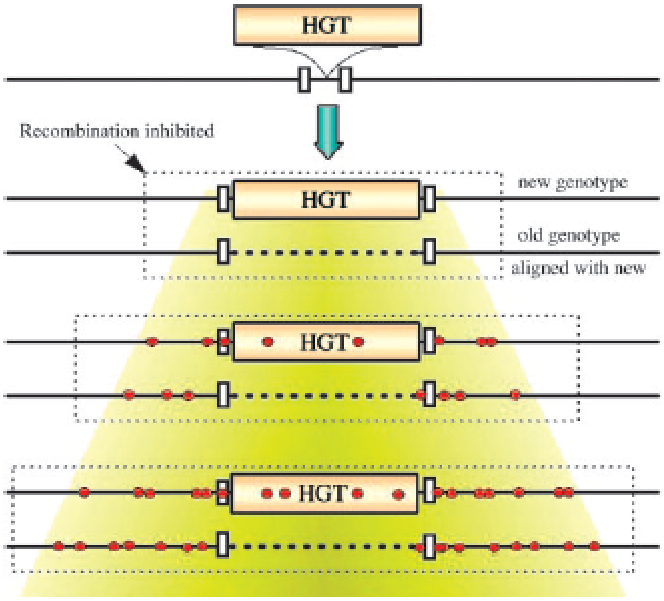

Fig. 2. As a HGT is trying to recombine at a specific part of the sequence, point mutations can accumulate and destroy the end identity required for the process to be successful (ref. 2).

Modeling the Competition

Assume N closed circular sequences where each position can be occupied by one of a total of n letters. That is for a set of N DNA sequences corresponding to N microbes, each sequence is made of n different bases. In each position each letter can be changed to any other letter of the alphabet with a rate m. This is the mutation rate m which is taken to be the same throughout the entire sequence. Additionally each genome receives fragments of size F at a rate r. The fragments can be from any donor in the set of N and attempt to recombine at any position of the target sequence. The initiation of the recombination process requires that a length M of the fragment should be identical with the recipient's end region of attempted incorporation at one end (model I) or both ends (model II) or at no end at all (model III). The success of the attempted recombination was given in (eq. 1).

An appropriate measure of whether the sequences in a population remain almost similar or the population becomes diverge is the order parameter,

$$\psi\left(x\right)=\frac{n}{n-1}\frac{1}{N\left(N-1\right)}\sum_{i=1,j=1}^{N,N}\left(1-\delta_{A_{xi},A_{xj}}\right) \tag{2}$$

It was carefully chosen (ref 2) to compare the letters occupying each position x between all the $N\times(N-1)$ pairs[2] of the N aligned strings. Down to one pair, the position on one sequence can be occupied by n letters with only $n-1$ left to occupy the other if they are to be different. With that normalization chosen[3], it gives $\psi=0$ for highly correlated sequences and $\psi=1$ for uncorrelated ones.

The Phase Transition

In the context of the three models specified above, a series of simulations were performed starting from different initial conditions. The genome sequences could be all the same or very diverse. By varying the strength of the mismatch repair system, quantified in α , it is interesting to look at the system for different rates of mutations relatively to recombination rates, quantified in $\mu=m/r$.

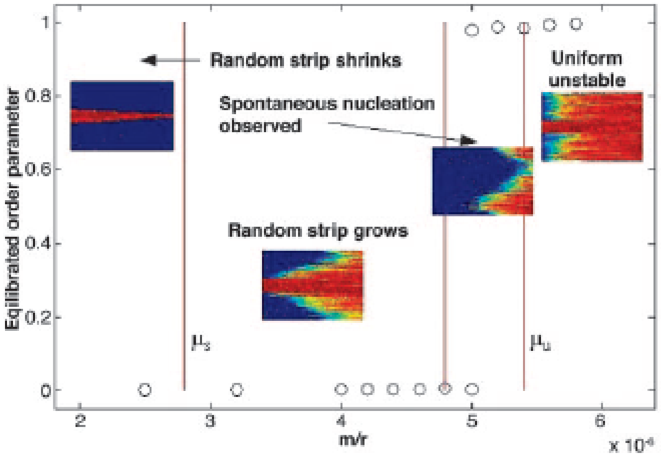

Fig. 3. Starting from the uniform state, the order parameter will relax to one of the two extreme values, 0 for the uniform state and 1 for the diverged. The insets show the genome evolving with time for the different values of the parameters. The position along the genome is on the vertical axis and time is on the horizontal. The color scale relates to the order parameter (blue for the uniform and red for the diverged phase). For intermediate values of $\mu_s<\mu<\mu_u$ a diversification front can be triggered while for $\mu\sim\mu_u$ nucleation can happen (ref. 2).

Starting from a uniform state with $\psi=0$ and above a threshold value of $\alpha>0$, the order parameter will relax with time very close to one of the two limiting values (fig .3). For

[2] The number of pairs here includes double counting but so does the sum.

[3] The equivalent position x in a pair of aligned sequences can be written in a total of $n\times n$ configurations with $n\times(n-1)$ of them being different and the rest $n\times1$ same. In what follows, the order parameter goes to zero for same configurations while for different goes to the ratio of the different over total number of pairs, $n(n-1)/n^2=(n-1)/n$ and consequently has to be normalized with that factor as well.

$\mu < \mu_s$ is equilibrates to $\psi = 0$ in a genetically uniform phase. On the other extreme, for $\mu > \mu_u$, it goes to a diverse phase with $\psi = 1$ where all sequences diverge from each other. For values in between, $\mu_s < \mu < \mu_u$, diversification front propagation occurs with the diverse phase growing and nucleating inside the uniform (fig. 3, ref. 2,6).

The previous observations are very well depicted in the qualitative phase diagram (fig 4). For models I and III, it is apparent that there is a threshold value for the parameter α, above of which the front propagation region is clearly observed. Below that threshold though, there are no distinct boundaries between the diverged phase and the uniform (fig 4a). For model II, the front propagation separates the two phases for all values of α eliminating the threshold. Additionally, the observed width of the front propagation region, $w = \mu_u / \mu_s$, was $w \leq 2$ for models I and III whereas for model III it was $w > 100$, exceeding the simulation limits quite often.

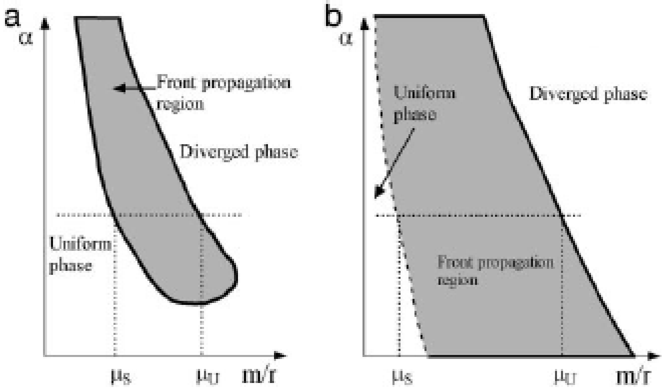

Fig. 4. The quantitative phase diagram depicting the competition between point mutations and recombination for different strengths of the mismatch repair system (ref. 2).

Microbe Classification

The specifics of the transition from a uniform to a diverse phase can be used to classify microbes. Class I can be defined to contain the models I and III that follow the first phase diagram (fig 4a) and class II to contain model II following from the second phase diagram (fig. 4b). For class I diversification fronts can propagate only for a small range of the ratio of the two rates thus rendering them unlikely. However, for class II the diversification fronts can be sustained and therefore allowed to propagate for a wide range of values of the competing mechanisms (ref .2).

IV. Comparison with Genome Data

Analysis of Actual Genome Data

Class II microbes require fragment identity at both ends in order for it to be considered for incorporation in the recipient's DNA sequence. And this is the case in the genus Bacillus (ref. 2,4).

To study the divergence between pairs of DNA sequences of different members of the genus Bacillus, Vestigian and Goldenfeld (ref. 2) devised an appropriate coarse-graining procedure. First the two sequences were aligned and the well-aligned regions were mapped on one of the two along with any small-scale differences. Then a window of width W was slid along the new string of commonly sequenced pieces where the number of identical positions K and different ones D were counted. The divergence d was then determined by,

$$d = \frac{D}{K} \tag{3}$$

In this type of averaging process the completely different fragments that have resulted from insertions or deletions were excluded altogether. Also excluded were any well-aligned regions where the number of the total positions K came out to be less than a satisfactory fraction,

$$f = \frac{K}{W} \tag{4}$$

of the width of the window (fig. 5).

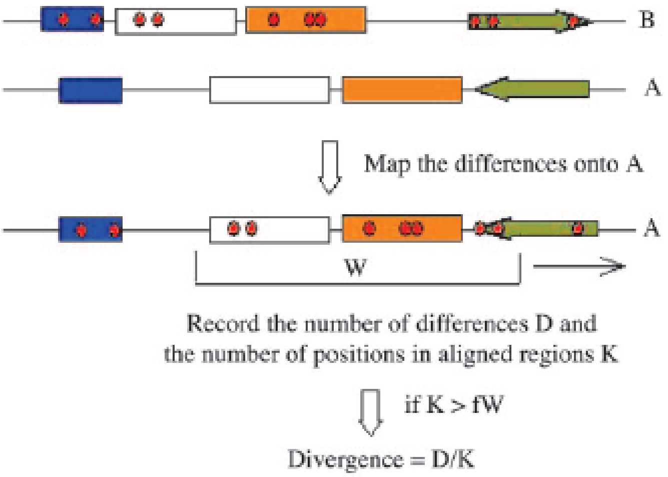
B
A
Map the differences onto A
A
W
Record the number of differences D and
the number of positions in aligned regions K
if K > fW
Divergence = D/K

Fig. 5. The averaging procedure was performed by sliding a window along the well-aligned regions (colored bars and arrows) of the genome sequences and mapping the differences (red circles) (ref. 2).

This coarse-graining procedure results in a step-like pattern (fig. 6) for all of the different components of the genome examined. Apart from the point-wise differences in the sequence of the nucleotides, larger sets composed by more than one nucleotide and responsible for protein production were fed in that procedure and their differences were calculated (ref. 2). This pattern follows to a considerable extent the pattern of the protein coding regions. The number of differences was less in the regions where more information was coded. This implies that the majority of the mutations recorded were silent, thus not affecting the highly organism-specific protein production, which in turn can be a simple consequence of different mutation-susceptible regions in the sequence. And since the components of the organism that lend to it its individuality were less affected by mutations, the main outcome of the previous section that mutations result in the divergence of an initial set of very similar sequences falls into despair. However, the fact that the step-like pattern is observed in all different components suggests that the different density of protein coding regions cannot account for the entirety of the pattern.

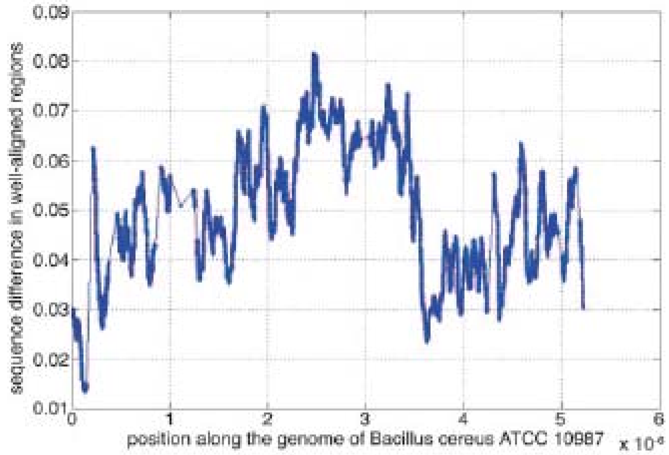

Fig. 6. Step-like sequence difference between two sequences of members of the genus bacillus (ref. 2).

Testing for Recombination Distribution

To solve the riddle Vestigian and Goldenfeld (ref. 2) implemented a test for the statistical analysis of the differences of well-aligned genome sequences not very different from the tests applied by Sawyer (ref. 5). For each pair of strings and for the well-aligned parts of them (as in fig 5) they studied the Distribution of the Lengths of the Maximal Exact Matches (DLMEM). That is, they collected measurements for the lengths x_i of the identical regions, defined as the lengths in between the different positions, and calculated the mean,

$$mean = \langle x \rangle = \frac{1}{Q}\sum_{i=1}^{Q} x_i \tag{4}$$

and the standard deviation,

$$\begin{aligned} std &= \sqrt{\langle x^2 \rangle - \langle x \rangle^2} \\ &= \sqrt{\frac{1}{Q}\sum_{i=1}^{Q}\left(x_i - \langle x \rangle\right)^2} \end{aligned} \tag{5}$$

If the genome had evolved due to point mutations mainly, then the Lengths of the Maximal Exact Matches (LMAM) were expected to follow the Poisson statistics with a ratio of,

$$\left.\frac{std}{mean}\right|_{Poisson} \approx 1 \tag{6}$$

This was not the case. They definitely found a ratio

$$\frac{std}{mean} > 1 \tag{7}$$

indicating that a broader distribution was governing their values which implied recombination processes. There was also a positive correlation between the ratio $std/mean$

and the values of the lengths x_i (fig. 7a), which suggests that longer pieces of well-aligned regions are to be drawn from a wider distribution consequently appearing more rarely in the sequence strengthening the argument that non-aligned regions tent to prevent HGTs from being incorporated to neighboring well-aligned regions (ref. 2).

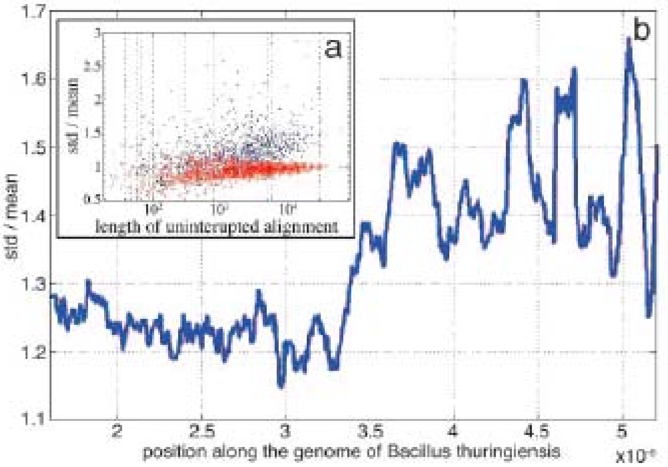

Fig. 7. The ratio $std/mean$ for DLMEM between two members of the family of the genus bacillus. (a) The ratio has positive correlation with the lengths of the uninterrupted region. (b) The ratio along the sequence has values well above unity (ref. 2).

Also, the ratio $std/mean$ obtains different values, as can be determined by again sliding a window along the genome strain, with a step-like pattern (fig.7), indicating that different regions experienced different rates of recombination relatively to mutations. Finally the comparison with a solely mutated strain gives confidence to the ratio $std/mean$ departing from unity to higher values implying recombination as the common mechanism of the clustering of the differences on the genome sequence.

V. Discussion

Analysis of Results

It is apparent from the outcome of the previous study that a phase transition occurs for appropriate mutation versus recombination rates. The single initial community can be driven to a genetically diverse phase where organisms acquire clear genetic boundaries. On the same notion, the front propagation that facilitates the emergence of species and evolution appears to be incompatible with HGTs. And that is because the HGTs rather tend to make the sequences more similar to each other. On top of that beneficial mutations are expected to make sequences more distinct, something which will inhibit HGTs, keeping them distinct (ref. 2).

Ideas to be Further Explored

The model studied here could be extended to include natural selection. This can be done by adding spatial parameters and the ability for organisms to congregate or become isolated depending on the morphology of the environment (ref. 2). However, the largest evolutionary steps were not triggered by the complexity of the environment but from competing with other organisms over the same resources. The inclusion of the parameter of competition among different species will probably uncover significant characteristics of the diversification and in general the evolutionary process.

VI. Acknowledgements

I would like to thank greatly Professor N. D. Goldenfeld for taking the time and discussing with me his recent work which is the major subject this term paper. I am also grateful to him for his inspiring teaching on the subject of phase transitions and the renormalization group and fascinating research projects like the one of (ref. 2). I would also like to thank the Physics Department of the University of Illinois at Urbana-Champaign for its support and hospitality. Finally special thanks to graduate student at UIUC C. Chialvo for helping me in the preparation of the manuscript in appropriate electronic form for internet posting.

VI. References

1. Woese, C. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 8742-8747.
2. Vestigian, K. & Goldenfeld, N. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 7332-7337.
3. Vulic, M., Dionisio, F., Taddei, F. & Radman, M. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 9763-9767.
4. Majewski, J. & Cohan, F. (1998) *Genetics* **148**, 13-18.
5. Sawyer, S. (1989) *Mol. Biol. Evol.* **6**, 526-538.
6. Goldenfeld, N. (1992) Lectures on Phase transitions and the Renormalization Group, *Perseus Books Group*.