Modeling Genome Evolution

Creation, Change and Destruction

MATS PETTERSSON
Dissertation presented at Uppsala University to be publicly examined in Zootissalen, Evolutionsbiologistcentrum, Norbyvägen 18, Uppsala, Tuesday, September 25, 2007 at 10:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Historically, evolution has been studied either by looking at morphological traits in living organisms and the fossil record, or by using bioinformatics and comparative genomics. While highly useful for deducing evolutionary history, these approaches are not particularly well suited for studying the mechanisms of evolution. In order to address such issues, other methods are needed. Mathematical modelling is one of the most powerful options available, and it is the approach used in this thesis. By constructing models of biological systems, the work aims to resolve some of the many unresolved questions regarding evolutionary processes, such as how new genes evolve and how selection acts in fragmented populations. Some answers have been reached, and thus the thesis makes a small contribution to our overall understanding of evolution.

The creation of novel genes was studied both directly and by extension of an analogous system, which revolved around reversion of a frameshift mutant. The results pointed to gene amplification as a likely mechanism for both reversion of the frameshift mutant and creation of new genes.

Selection in fragmented populations is shown to be effective even when sub-populations, rather than individuals, are competing against each other. Modeling of a system of bacterial symbionts living in aphids indicates that, although the bacterial population within a single host is small and subject to rampant genetic drift, the bacterial population as a whole is regulated by selection on the host level. Thus, deleterious mutations do no accumulate and the population maintains its fitness over time.

Keywords: Theoretical biology, Population genetics, Stochastic modeling, Genome evolution

Mats Pettersson. Department of Evolution, Genomics and Systematics, Norbyv. 18C, Uppsala University, SE-75236 Uppsala, Sweden

© Mats Pettersson 2007

ISSN 1651-6214
urn:nbn:se:uu:diva-8163 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8163)
To Olga and Björn.

If not for you.
Main references

This thesis is based on the following papers, which will be referred to in the text by their roman numerals.


Introduction

Fundamentals

The necessity of evolution

All lifeforms evolve. The very nature of life dictates that this must be the case. Life proceeds by copying itself, and since it is impossible to create perfect copies, changes are inevitable. Thus, evolution must occur. Changes create variation, and, given that resources within the biosphere are finite, variation gives rise to competition. Consequently, the strongest competitors will prosper at the expense of less well adapted individuals. Therefore, also selection is inevitable. Due to this fact it should be noted that, even if replication without error was possible, perfectly replicating organisms would be eliminated by adaptable competitors. These inherent properties form the basis of Darwinian biology and, even more importantly, are directly responsible for the world we live in today. Therefore, in my mind, studying evolution and the mechanisms involved in it is the most fundamental aspect of modern biology. Any advances will help explain fundamental issues concerning how the present day world arose, and where it is headed.

Evolution is central to all aspects of modern biology, and is heavily studied from almost any conceivable point of view. Despite this vast effort, the topic is sufficiently deep for much to remain unclear. There are many obstacles to the study of evolution, but one stands above all else: time. Evolutionary processes typically stretch over vast amounts of time, slowly playing out across thousands, or millions, of years. Naturally, this makes them notoriously difficult to study experimentally. There are exceptions – under the right conditions bacteria can evolve noticeably in a matter of days – but generally, evolution must be studied indirectly.

The most obvious source of information are the organisms present today. These are, in a sense, the result of all evolution that has occurred to date, and must therefore contain clues to the process that created them. Essentially, the idea is that by studying the outcome of evolution, one can infer the underlying process.

Studies of present day organisms, in particular genome sequences, have certainly led to significant insight into the evolutionary history of these organisms: One particular highlight was the revolutionizing of systematic biology achieved by the construction of “the Tree of Life” (Woese and Fox,
1977); a grand phylogeny linking all living organisms together and establish-
ing the three domains of life, Eubacteria, Archaea and Eukarya. However, the results are often contradictory to some extent, and there are still disagree-
ments on the interpretation of phylogenetic data. “The Tree of Life” remains controversial today, a full thirty years after its introduction. Despite immense sequencing efforts, creating enormous amounts of data, early evolution is still largely a subject of speculation rather than facts, partly due to lack of understanding of the actual mechanisms of evolution. It turns out that study-
ing the organisms of today, while useful for determining history, does not re-
veal how evolution occurs, on the detailed level.

The evolutionary process

The general evolutionary principle – evolution through natural selection – is well known, and has been so since the time of Darwin. In short, it can be de-
scribed as follows: all populations have some amount of variance between individuals. Some variants will be better able to cope with the challenges presented by the environment, and are thus more likely to survive, prosper and, most importantly, propagate. Over time, these favored variants will dominate the population, until an even better variant appears and the cycle begins anew.

While essentially true, this simple description leaves many questions unanswered. In particular, the generation of variation is a black box, and the speed with which the process occurs is not addressed. Obviously, modern bi-
ology have refined this basic picture significantly, but much still remains to be done. The work in this thesis is aimed at improving our understanding of some details of evolution, and thus enable a more precise description of the evolutionary process. In particular concerning mechanisms for generation of variation and effects of population biology and structure on efficiency of se-
lection and, consequently, the speed of evolution.

Genome evolution

This thesis deals with a variety of evolutionary mechanisms, all unified by the fact that they are relevant to the long term evolution of organisms in gen-
eral, rather than being confined to a specific context. Most of the work adopts a holistic view of the evolutionary process, by focusing on evolution of entire genomes rather than individual genes. While there, naturally, are particular circumstances associated with the evolutionary fate of each and every gene, many factors apply equally to all, or at least large groups of, genes. By examining processes that are in effect all over the genome, it is possible to make statements regarding genome evolution without taking all peculiarities of the individual genes into account. This enables more far-
-reaching conclusion than would otherwise be possible.
On the other hand, it is sometimes beneficial to apply the general principles to a defined example, since that allows for direct comparisons between the model and reality. Therefore, the models used in this thesis are adapted to match an actual biological system when suitable, in order to obtain a stronger link to reality, and consequently increased confidence in the results.

Studying evolution on the genome scale means dealing with the forces that govern the long term changes in the overall capacity of a given genome. Depending on the circumstances, the genome might constantly acquire new genes and increase its functional capability, or it might lose genes which are no longer contributing to the survival of the organism. Thus, understanding of the processes involved is needed for both explaining the genetic make up of present day organisms and for predicting their future development.

Traditionally, genomes have been considered to be stable entities, at least over short to moderate time. This is a natural viewpoint, since changes on the genomic scale have drastic effects, potentially involving any number of cellular functions. However, with increasing genomic data becoming available, it has become obvious that this is not necessarily the case. Despite the difficulties connected with genomic changes, it is apparent that genome content and layout is far from static on the evolutionary scale. Instead we have realized that genomes, in particular microbial ones, are volatile and may quickly shrink or grow in size (Bergthorsson and Ochman, 1998). In fact, even within a single species, with some reservation for what truly constitutes a bacterial species, there can be 30% or more difference between isolates.

Genome size affects the biology of an organism in many ways. This is particularly true for microbes, where genome size is closely linked to genome content. A larger genome means greater diversity, and therefore more adaptability. On the other hand, a large genome usually carries many non-essential genes, which might confer a significant cost in terms of unnecessary protein expression or complicated regulation mechanisms. In this perspective, it becomes obvious that the genome size of an organism will depend on the environment in which it lives, or if it moves between different environments, and the challenges it has to face. This means that the evolutionary possibilities of an organism are connected with the features of its genome and shows the importance of understanding the dynamics of genome evolution.

**Approach**

Genome evolution can be studied by bioinformatic methods, comparative genomics in particular, is well suited to the task. However, these methods are limited by the fact that it is only possible to examine a subset (i.e. those that can be cultured or captured) of the organisms currently in existence. Because
of the limitations in the available data, especially the lack of depth in time, any conclusions will have to be based on assumptions concerning whatever processes have shaped the species up until the present day. Naturally, these assumptions are difficult to assess, given that the proposed mechanisms have been at work for millions, or even billions, of years. Thus, it is generally very difficult to directly address the issues experimentally. Instead, more indirect methods are necessary.

The modeling method

Since the main obstacle in examining evolutionary processes is the huge amount of time involved, it is immensely helpful to construct a model of the system, and analyze the model rather than the real thing. If a satisfactory model can be constructed, the system can be studied by mathematical and computational methods. This allows for, almost, arbitrary speed up, making it possible to study processes on any time scale. In addition, models are not restricted to organism that can be cultured or otherwise subjected to experiments.

In order to construct a model of a biological system, the processes involved must be transformed into a mathematical framework, where all possible events are described by appropriate expressions. Once the framework is in place, the behavior of the model can be studied, either by analytical means or by simulations. If the mechanism has been perfectly converted into a model, the results will be just as real as in any practical experiment. While that ideal situation is not practically achievable, even a model with minor deficiencies will produce useful results as long a one is mindful of the limitations of the model. Obviously, this means that substantial prior knowledge of the system is a prerequisite for successful modeling. Nevertheless, the advantages of utilizing mathematical modeling are obvious; in fact, there is often no realistic alternative.

The power of mathematical modeling lies in its versatility. It makes it possible to use a common method to analyze mechanisms which are fundamentally different, giving a unique opportunity to study a complex problem from many angles at once. By connecting experimental data, which is usually on a detailed level, in a common theoretical framework, one can facilitate understanding of the entire process. At its best, the model will be the glue that keeps the pieces of experimental data in their place, allowing the entire puzzle to be solved.

Combining theory and experiments

While modeling is a powerful tool, it cannot be done without making use of knowledge based on experiments. Without extensive knowledge about the system at hand, it is impossible to construct a relevant model, and, generally
speaking, the only way to generate that knowledge is by performing experiments. Furthermore, experimental data is also highly useful for estimating parameter values, which is important if one wants to determine the implication of the modeling results for real biological systems. The work in this thesis draws heavily upon the efforts of others, and would not have been possible without exploiting a large body of experimental data. The modeling adds further value to the experimental data by creating a new context and by doing so, it provides access to information that was otherwise not readily obtainable.
Areas of study

Creation of genes and functions
Papers I and III revolve around gene amplification by homologous recombination as a means of facilitating improvement of a function. For an evolutionary process, this is extremely quick; the entire process can be completed in a few tens of generations. While they differ in scope and emphasis, both papers deals with the same basic mechanism and its effects.

The amplification model
The idea of amplifications as a mechanism for speeding up evolution arose from observations in the Cairns system (Cairns et al., 1988; Cairns and Foster, 1991), where *Escherichia coli* cells that are unable to metabolize lactose, due to a frame-shift mutation in the lac-operon that disrupts the β-galactosidase gene, revert to wild type at rates which are orders of magnitude faster than what is expected from the regular point mutation rate. Three main models have been suggested to explain these results. The first two, Directed Mutation and Hypermutable States (described below), assume that revertants arise as single-step mutations in this non-growing lawn of parental cells, and that the increased speed was due to elevated mutation rates, either specifically in the affected gene (Cairns and Foster, 1991; Foster and Cairns, 1992) or in the genome in general (Hall, 1990). These models have been shown to contain serious flaws (Sniegowski and Lenski, 1995; Torkelsson et al., 1997; Roth et al., 2003) and are most likely not corresponding to the true course of events.

Amplification Mutagenesis was suggested as a third possibility, and has been supported by significant evidence, both experimental (Andersson D. I. et al., 1998; Hendricksson et al., 2002) and theoretical (Paper I). While there are difficulties in explaining all experimental observations also with this model (Stumpf et al., 2007), it is clear that amplification does occur in the system. The amplification model assumes that reversion does not occur in the non-growing parent population, but rather within growing clones. The clones are initiated by cells with a duplication of the mutant lac region, which retains some weak functionality (Andersson D. I. et al., 1998; Hendricksson et al., 2002; Slechta et al., 2003). In this model, growth improves as unequal recombination adds more lac copies and, concomitantly, the
probability of a reversion (-1 frameshift) event goes up due to the increase in lac copy number within developing clones. It is critical that cells with a duplication of the frameshift mutant grow, albeit very slowly, on lactose. Because the growth rate will be essentially proportional to the number of lac operons present in the cell, selection will strongly favor cells with larger copy number and such cells will accumulate in micro-colonies, explaining the high number of revertants.

Paper I deals specifically with the dynamics within the Cairns system. By adapting the model to a specific system, it becomes possible to use the experimental data available to verify the validity of the model and guide parameter choices, making the conclusions stronger and more precise than what is usually possible when modeling evolutionary processes. The modeling results show that amplifications can account for the observed increase in reversion rate, with biologically relevant parameter values used. While this does not prove that the amplification mechanism is the true explanation of the phenomenon, it significantly increases the credibility of the model.

Paper III examines the implications of amplifications in a wider evolutionary context, by examining the potential of the adaptive amplification mechanism for establishing completely new functions. This makes the conclusions much more widely applicable, but also makes it difficult to relate the modeling to experimental results. The paper contains a detailed description of the process, and compares the amplification model to other proposed models for establishment of new functions. The results indicate that exploiting the amplification mechanism is advantageous under most circumstances. Thus, it seems likely that amplifications are responsible for a large fraction of the new genes created in nature.

Loss of genes and functions

Evolution does not necessarily favor increased complexity. There are situations where trimming of excess genetic material is driven by selection, and there are also situations where the population structure causes loss of genes through genetic drift, to the detriment of the organisms involved. The first case occurs in very stable environments, and is usually associated with parasitism or symbiosis (Andersson and Kurland, 1998). Due to the precise regulation inherent in all living cells, organisms living inside other organisms experience a degree of stability and predictability that cannot be otherwise achieved. Thus, it is commonly seen that intracellular organisms have lost, or are in the process of losing, large amounts of genetic material compared to their free-living relatives. This is particularly true if the symbiont/parasite is specific to a single host (Shigenobu et al., 2000; Alsmark et al., 2004; Andersson S. G. E. et al., 1998). Loss due to drift is, on the other hand, most of-
ten linked to small population sizes, where selection is ineffective. However, it can also occur if the population is oddly structured. Such odd population structures include populations where only a few individuals propagate, as well as patchy or otherwise compartmentalized populations.

Muller's ratchet

Inevitable accumulation of deleterious mutations in asexual organisms, is known as Muller's ratchet (Muller, 1964). Paper II focuses on a special case of this phenomenon. It deals with a two-level population structure with small populations of symbionts living inside a large population of hosts. In particular, paper II examines small populations nested in large ones, such as symbionts within hosts. As a model system, we use aphids and their symbionts, of the genus *Buchnera*. The peculiar population biology of *Buchnera*, and other vertically transmitted symbionts, shares many features with that of the eukaryote organelles, making it an interesting system to study. This type of two-level systems have been studied before (Bergstrom and Pritchard, 1998, Rispe and Moran, 2000), but there is still much to resolve.

The *Buchnera* are transmitted exclusively from mother to offspring, and have no way of moving between unrelated host organisms (Baumann *et al.*, 1997), much like mitochondria. However, the symbionts still reproduce independently within their host, and thus compete with each other for transmission to the next generation. This gives rise to a cyclic growth pattern, with bottlenecks occurring at each host generation and clonal expansion in between. In addition, the symbionts have a direct impact on the fitness of their host, since they synthesize metabolites which the aphid can not supply on its own. Consequently, there will also be host level selection acting on the symbionts. Of particular interest is the behavior at relatively large population sizes, which has hitherto not been addressed. To deal with large populations, it becomes necessary to find simplifications. The paper focuses on the interplay between the two levels of the system, with the specific goal of identifying the circumstances under which it is possible to simplify the system while still retaining a faithful description of its behavior. Furthermore, it examines parameter regions where the ratchet proceeds slowly and ultimately stops, which is crucial to the ultimate fate of the species involved.

The *Buchnera* genomes carry obvious marks of destructive forces at work and are in partial decay. In particular, there are many incomplete metabolic pathways (Shigenobu *et al.*, 2000). Muller's ratchet has been proposed as a major factor in this genome degradation (Moran, 1996; Brynnel *et al.*, 1998; Andersson and Andersson, 1999), making it interesting to determine if such deleterious traits can be fixed by the ratchet under the present circumstances, and at what speed fixation would occur. Ultimately, this comes down to whether or not the symbionts are headed for extinction due to accumulating genetic damage. Our results (Paper II), as well as data from comparative ge-
nomics, indicating that the *Buchnera* genomes are exceptionally static (Tamas et al., 2002), support the notion that extinction due to the ratchet is unlikely. In fact, it is uncertain if the ratchet can even operate in these two-level populations, since the results show that it is disrupted by selection within the host population, provided it is sufficiently large.

Deletion rate

Paper IV also deals with reductive evolution, although in a different context. It addresses a situation where reduction of genome size is adaptive, and decreased genome size is favored by selection. This creates an interesting dynamic, since there is separation between the gene that regulates the deletion rate and the selection pressure – which acts on the actual deletions. The consequence is that the efficiency of selection for increased deletion rate is dependent on the amount of linkage disequilibrium present in the organism in question. Thus, selection on the deletion rate will be ineffective in species with sexual recombination, or rampant conjugation. The results suggest an interesting interaction between the deletion rate and the density of essential material in the genome, which could confer a possible dynamic regulation of genome size via the deletion rate. However, this manuscript is still in preparation, and the simulation results are preliminary.
Summary of the papers

Paper I – The amplification model for adaptive mutation

Background
Paper I is a theoretical examination of amplifications in the Cairns system, in which lac E. coli cells revert to the wild type at rates several orders of magnitude faster than expected from the point mutation rate (Cairns et al, 1988). The phenomenon has been shown to depend on the particular mutation involved, a frame-shift mutation which confers some residual $\beta$-galactosidase activity (roughly 1% of the wild type). This, together with the fact that the fast reversion does not occur in recombination deficient E. coli, led to the proposal that reversion is achieved via a temporary amplification of the lac-operon, driven by selection for the residual activity of the frame-shift mutant. Paper I examines this idea by formalizing the thought model into a mathematical framework, and uses that framework to explain the experimental observations.

Models
The models cover the process of forming a colony from a single cell. Thus, it is possible to analyze any stage of colony development. Two types of genetic events are included: increase or decrease in the number of lac copies by recombination (the fundamental amplification mechanism) and correction of the mutant lac allele (reversion to lac+) by a frameshift mutation. Recombination is assumed to be equally probable between all lac-copies and to occur at a constant intrinsic rate. The probability of a reversion event is, similarly, proportional to the copy number of the mutant lac allele per cell); this rate constant is the probability of mutation per lac copy per cell division. In order to mimic colony formation, the population is growing exponentially from a single cell. Since the amplification process can only take place with two or more copies present, the original cell is assumed to carry at least two copies of the lac-operon. Studies have shown that a few percent of the bacteria carry such a duplication at any given time. Due to the strong selection on copy number, these cells will give rise to the vast majority of all revertant colonies.
The recombination rate per cell division is a function of the copy number, and the intrinsic probability of recombination between two copies. Recombination events will often, although not always, yield daughter cells with copy number different from that of the mother. For simplicity, the model assumes that the copy number of the mother cell remains the same. If recombination was instead treated as reciprocal, every recombination event causing a change will yield one cell with higher copy number than the original and one with lower copy number. Simulations show that this is essentially the same as using a higher recombination rate. Thus, for simplicity, we consider the mother cell to remain constant after recombination. Recombination is closely tied to fitness, since the fitness of a cell is determined by its \textit{lac} copy number.

Throughout Paper I, the model is either implemented as a Monte Carlo simulation or analyzed by numerical integration, depending on the specific context. For the details of the computational approach, see the appropriate section of Paper I.

Parameters

The intrinsic recombination rate is estimated to in the range of a few percent for a long, perfect, tandem repeat. This estimate is supported by experimental data (Andersson and Roth, 1977; Reams and Neidle, 2003). The rate of reversion (back-mutation) of the \textit{lac} frameshift mutation used is $10^{-8}$, based on experiments with unselected growth. The growth rate contribution from each copy is derived from the \textit{\textgreek{z}}-galactosidase activity yielded by one copy of the frame-shift mutant, which has been shown to be around 1 \% of the wild type (Andersson D. I. et al., 1998). The selection coefficient is the product of the activity and the copy number of the F'- plasmid, which is between 2 and 3. The load parameters reflect the cost of maintaining and expressing many copies of \textit{lacZ} and other genes included in the duplication, as well as the increasing risk of deletion events where a part of the array is looped out and lost. These are difficult to estimate from experimental data, but fortunately the exact choice of load parameters is not crucial for the behavior of the system. As long as the expression follows the assumption of the model, which is that fitness increases with increasing copy number up to some sufficiently large (>50) optimum value, the system is robust to changes in these parameters.

Results

The Monte Carlo simulations show that after a period of slow growth of cells with low copy number, the average copy number takes off quickly. The reason for the initial lag phase is the need for time to create variation, before se-
lection can take effect. Once the average copy number starts to increase, the appearance of the first revertant becomes inevitable. Once the revertant appears, the amplification loses its selective advantage and becomes counter selected, due to the load penalty. Thus, the revertant sub population quickly loses the extra copies, and stable (single-copy) cells appear a few generations later and start rising rapidly in the population (Figure 1).

In fact, since the timing of the first reversion event is so crucial to the history of the clone and its probability can be analytically calculated, it is possible to answer many questions without actually performing a complete simulation. Rather, just following the development of the reversion probability is sufficient. It is not certain that all cells initially carrying a duplication will give rise to a revertant colony within the time of a typical experiment, thus it is of interest to examine also the region where reversion is a low probability event. The resolution of the analytic probability calculation makes it possible to examine the dependence on the recombination rate, which allows estimation of the development of the reversion probability in this key region.

There is a threshold value where recombination is too slow to have any real impact on the system. In this case, reversion occurs only at a colony size of a few times $10^7$ cells, which is what one would expect if the average copy number of the colony remains at two. At the other extreme, when recombination is as frequent as one per division (the maximum), reversion occurs at a frequency that is not very different from a recombination rate of a few percent. This is not unexpected as the probability of reversion depends on the total number of lac copies in the colony. Thus even if the optimal copy num-
ber is reached quickly, the population must be large before reversion is likely. Figure 2 illustrates the effect of the recombination rate on the probability of reversion.

Figure 2: Reversion probabilities for different recombination rates

While all colonies are initiated by a single cell, that cell does not necessarily have exactly two lac-copies. Calculations on neutral duplications indicate that, while most cells with multiple gene copies have only two copies, there is a thin but long tail of cells with higher copy numbers. However, initializing the colony with a higher copy number has limited effect on the outcome. As an example, shifting the initial copy number from 2 to 8 moves the average reversion time less than half a generation. However, due to the increased growth rate of the initial cells, this can correspond to a more substantial shortening of the actual time needed to form a colony.

The possible loss of the reverted copy through recombination events between flanking non-revertant copies is not included in the models used. Such losses might slow down the process, but only marginally. Given the assumption that all recombination events are equally likely, the probability of loss during a recombination event is:

\[ p_{\text{loss}} = \frac{z(n-z)}{n} \]
where $n$ is the number of copies, and $z$ is the position of the reverted copy. It is obvious that this probability is maximal when the reversion is in the middle copy ($z = n/2$) and that this maximum is 25%. Thus, even in a worst case scenario, 75% of all deletions are allowed, and the process is unlikely to be very much affected, due to the strong selection in favour of the revertants.

Conclusions
The calculations and simulations described here show that the amplification model can quantitatively explain the behaviour of the Cairns selection system, if the variables are appropriate. Recombination rates cannot be below 0.5% per cell/division, or the gene amplification process will not lead to reversion within the time-scale of the experiment. Equally important is that to the genetic system can accommodate copy numbers on the order of $10^2$, or the process will not be quick enough to explain the observed reversion rate.

Paper II – Muller's ratchet in symbiont populations
Background
Paper II deals with the evolutionary fate of insect symbionts, and other organisms that have similar lifestyle. Since the symbionts are completely isolated within their hosts, they have long been believed to risk extinction by genome degradation, due to the inefficiency of selection in such small populations.

In particular, the effect known as “Muller's ratchet” poses a threat. In a asexual population, each cell carries at least as many deleterious mutations as its mother. In a large population, this is of no consequence, as the selective disadvantage of mutated cells will cause them to be eliminated from the population. However, in a small population it is possible that, by chance, all individuals present in the population at a given time happen to carry at least one deleterious mutation. If this occurs, there is no way to recreate a perfectly healthy cell. Such an event is known as a “click” of the ratchet. In the absence of other factors, this process goes on indefinitely, ultimately causing the population to go extinct.

However, while the aphid symbionts themselves are isolated, they do interact indirectly, via their hosts. Since the host depends on metabolites provided by the symbionts, host fitness will be affected by the health of the symbionts. This provides a possible escape from the ratchet, since the hosts have a large population size, and are thus subject to effective selection. Paper II is focused on the interactions between the two population levels, and how that affects the evolutionary fate of the symbionts.
Models

Paper II uses different models, with different degrees of complexity. The starting point is a detailed model for the population-within-population dynamics. This model is shown possible to simplify by treating the two population levels separately, yielding a massive reduction of the computational load and making it possible to simulate large populations.

In the detailed model every host is a unique individual, carrying a specific set of symbionts, which determines its fitness. The symbionts undergo cyclic growth, growing from the size of the transmission bottleneck to the number of symbionts present in a mature host. During growth, symbionts accumulate deleterious mutations with a constant rate per replication. These mutations lower the fitness of both the symbiont and the host, although not necessarily to the same degree. The host population is assumed to be of constant size. The probability for propagation of a given host is equal to the fraction of the population's total fitness that host possesses. When a new host is born, it receives a random sample of symbionts from its mother. This is the transmission bottleneck. Finally, the symbionts grow and accumulate mutations. This completes the cycle, and a new host is chosen for the next repetition.

Simplification is achieved by separating the two levels, and treating them as two different one-level ratchet systems, where the symbiont dynamics within individual hosts are used to produce a mutation rate of sorts for the host population level. First, a single symbiont population going through growth and bottlenecks is simulated. The ratchet rate for this system is subsequently used as the “mutation rate” for the host population. As long as this model holds, the host dynamics are reduced into a one level ratchet system, which can be described by a preexisting model (Haigh, 1978) and is relatively straightforward to analyze.

Results

If the simplified models are to be of any use, they must accurately predict the behavior of the full model. Thus, the first part of Paper II is devoted to validating the simplification by comparisons with the full model. The results show that reducing the symbiont dynamics to an internal ratchet rate is generally satisfactory, but there are cases where the models diverge (essentially when the variation in fitness between hosts on the same click is large). However, a slight modification of the simplified model makes it an excellent predictor of the full model throughout the entire tested parameter range.
Using the reduced model, it is possible to analyze the behavior of the ratchet at large populations sizes. The ratchet time will start to increase very rapidly with increasing population size at some point. In short, this occurs when the size of the best class becomes large enough to support effective selection. This steep barrier forms an effective limit to which mutations can be fixed in populations with a given size. The results are shown in Figure 3. Concerning the specific case of the *Buchnera*, the effective barrier for progress of the ratchet lies at a host level selection coefficient on the order of \(-10^{-4}\). In other words, only very weakly deleterious mutations have any practical chance of fixation.

Using the internal ratchet rate, the host population dynamics can be described with Haigh's model. There are existing analytical approximations of single level systems (Stephan *et al.*, 1993; Gordo and Charlesworth, 2000), which can be applied. Although these approximations are not accurate in all limits, they will in general produce adequate estimates, and allow the system to be analyzed without performing simulations.

In addition, from the equations used in these approximations, it follows that ratchet times should increase linearly with increasing host population size, when the host level selection and internal ratchet rate are kept inversely proportional to the population size. This makes it possible to perform the calculations with a relatively small population, which is much faster. It is then straightforward to use the results to estimate the behavior of the full scale system.
Conclusions

The qualitative behavior of the system has been shown to change depending on the amount of variation present among the host individuals that have the same best class of symbionts. The magnitude of this variation is mainly determined by the size of the transmission bottleneck. Large bottlenecks cause large inter-class variation and as a consequence progression of the ratchet will, in some cases, be much faster than expected in a one-level population. By taking the intra-class fitness variations into account, it is possible to simplify the model and maintain a good approximation of its behavior. The interference caused by variations in host fitness is most pronounced for small population sizes and strong host level selection. By simplifying the complex two-level system into two separate systems, it becomes possible to study a much wider parameter space. In fact, using the inherent scaling properties one can estimate the behavior of arbitrarily large systems.

Paper III – Evolution of new functions

Background

The evolution of new genes or, more fundamentally, new functions is obviously a central evolutionary problem. All evidence suggests that duplications of existing genes is the source from which new genes arise. However, the details of how a duplication turns into a novel gene have not been easy to resolve. The original idea, which simply was that duplicate genes are relieved from selective pressure and will over time evolve a new function (Ohno, 1970) was proven to be untenable due to mechanistic reasons (Lynch and Conery, 2000; Berg and Kurland, 2002). It was shown that, duplicated genes are overwhelmingly more likely to be destroyed or lost than they are to persist and evolve. Extensions have been made to the original model, by adding the concept of subfunctionalization (Lynch and Force, 2000), but this does not resolve all the difficulties. This, combined with the results from the Cairns system (Andersson D. I. et al., 1998; Hendricksson et al., 2002; Paper I) and comparative genomics (Hooper and Berg, 2002), led to amplification being suggested as a possible mechanism. Amplification circumvents the problem of duplication instability, since it acts on a faster time scale. Paper III extends and modifies the model used in Paper I, in order to evaluate the potential of the amplification mechanism for creation of new genes.

Models

The model consists of a constant size population of cells, each carrying a given number of copies of the gene that will eventually give rise to a new
function. Each gene copy has two properties, representing the activities of the old and new function, respectively. At the start of a simulation, all gene copies have full activity of the original function and no activity of the new function. As the cells grow, their copy number changes through recombination, and the gene copies change their activities through point-mutations. The rates of these events depend on the rate constants for recombination and mutation, as well as the number of gene copies present in the cell.

Each function is assumed to have a maximum contribution to fitness, which is reached when the activity of that function is at least 100%. That is, if 100% activity is present in a given cell, adding more or better copies will not further increase the cell's fitness. Until the 100% limit is reached, activity is assumed to be additive, so that two copies with 2% activity each will combine to provide 4% total activity to the cell. The fitness of a given cell is a function of the contributions of activity from each gene copy present in the cell. In addition, there is a load penalty for carrying many copies, reflecting the cost of expressing excessive amounts of protein/mRNA as well as replicating extra DNA.

Parameters

The mutation rate sets the overall speed of the system by governing the inflow of new variants. It has a strong influence on the absolute results, but often has limited effect on the qualitative behavior of the system. The recombination rate per copy and cell division and the de novo duplication rate affect the system primarily through their ratio, which determines the frequency of duplications present in the population in the absence of any new functions. Since duplications are a prerequisite for the amplification process, the prevalence of duplications directly affects the efficiency of the mechanism. The key fitness parameter is the net benefit gained from one copy carrying a single mutation. This value is directly linked to the selective advantage of the first cells with the new function, and is thus crucial. In fact, for the mechanism to be able to operate, the selective advantage must outweigh the cost of carrying the extra copy. This ensures that amplification has a positive net effect. The fitness gain must also compensate for the rate with which the mutated copy is lost, which is related to the recombination rate.

Results

The amplification mechanism revolves around the formation and loss of arrays of gene copies in response to the selective pressure acting on the population during establishment of a new function. Initially, all cells carry a single gene without any activity for the new function. Over time, some cells will acquire spontaneous duplications. These duplications will be short-lived, as long as they confer no selective advantage. However, as time pass-
es, mutations create activity of the new function in cells carrying a duplication, and the average copy number is likely to increase. Further on, as the new function becomes increasingly optimized, copy numbers will again decrease, since the selective advantage of maintaining a large set of copies is lost. Finally, the population will approach a stable distribution with a maximum at two gene copies. Figure 4 illustrates the establishment process by means of a series of snapshots of the copy number distribution within the population.

![Figure 4](image)

Figure 4: Time evolution curves showing the distribution of copy numbers at a set of time points during establishment of a new function. The legend indicates the number of generations since the start of the simulation.

Under favorable conditions the amplification mechanism is very effective, and facilitates rapid establishment. However, changing the load penalty has a strong detrimental effect on the mechanism. This is due to the low fitness benefit of an extra gene copy with minimal activity. When the fitness gain is small, even a minor, in absolute terms, load penalty is significant.

If the mutated copies are frequently lost due to recombination, and the fitness contribution from a gene copy with one mutation is low, a single mutation might not be enough to stabilize the array. In that case, establishment requires a double mutant to occur, and establishment time will increase faster than linearly with decreasing mutations rates. Under such conditions, the mechanism will not be effective at the mutation rates expected in nature.
Conclusions

The amplification model is efficient only for genes that can be duplicated at costs that are negligible, or at least low, relative to the potential fitness gain and when a single mutant is sufficient for stabilization of a gene array. The amplification mechanism allows tandem, and thus unstable, duplications to contribute new functions to the genome, whereas other models require duplications to be stable in order to lead to establishment of a new function. Since no duplications are truly stable, this provides a strong argument in favor of the amplification mechanism.

Even if establishment is possible without amplifications, the results show that the process proceeds significantly faster via the amplification mechanism, especially since inactivating mutations ensure that no neutral duplication is truly stable. Thus, amplifications yield higher adaptability, which is an important trait in most environments.

Paper IV – Deletion rate and indirect selection

Background

Evolution of genome size is a topic that have proven difficult to grasp, despite many efforts. The organisms present on earth vary over many orders of magnitude in genome size, and even closely related species can have large differences (Bergthorsson and Ochman, 1998; Gregory and Herbert, 1999). The reason for this large variation is remains unknown, but several ideas have been proposed. Some authors suggest that genome size is governed by the unstoppable accumulation of junk DNA (Ohno, 1972; Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Others believe that the amount of non-coding DNA is a selected property, due to effects on the phenotypic traits such as cell size, cell division rates and many others (Gregory and Herbert, 1999). Finally, Petrov (2002) have suggested the “mutational equilibrium model”, which proposes that the balance between the rates of deletions and insertions determines the genome size. The work included in Paper IV suggest a new option, since the results indicate that selection on the deletion rate shifts depending on the fraction of essential DNA present in the genome. This introduces the possibility of dynamic regulation of the genome size, via regulation of the deletion rate. However, the relevancy of this observation is still uncertain, since lowering the deletion rate does not necessarily confer an increase in genome size.

Paper IV examines a model for deletion rate evolution, based on tracking the fate of deletion rate mutants. This trait belongs to a peculiar class known as modifier genes, since it has no direct influence on the fitness of the organism. However, if the majority of deletion events are positively selected, due to pressure towards a smaller genome size, it will be advantageous to in-
crease the deletion rate, and vice versa. Experimental work in E. coli have shown that indirect selection can be effective (Sniegowski et al., 2000; Shaver et al., 2002). Also known is that modifier genes are sensitive to the amount of genetic linkage in a genome, and consequently to the frequency of sexual recombination (Johnson, 1999). Paper IV studies how the indirect selection acts in this system, and examines the impact of sexual recombination.

Models

The model uses a Moran population with overlapping generations, in which the cells grow, die and evolve in a population of constant size. Each cell has a given genome size, which is the sum of the core genome and the non-essential genetic material. The amount of non-essential material changes over time due to deletions and insertions and is different from cell to cell. The population initially contains cells with two different deletion rates, and eventually one will become fixed and the other one lost. This setup allows for estimation of fixation probability and, by inference, the effective selection for or against the mutant allele. Deletions and insertions happen during replication, with frequencies dependent on both the intrinsic rates and the current genome size, and the fitness of a cell is given by the difference between the number of deletions and insertions it carries.

Results

As expected, the effective selection strength decreases with increasing sexual recombination. However, a significant fraction of the selection pressure remains at intermediate recombination frequencies, if the density of essential genetic material is low. Thus, there is indications that indirect selection can act also in sexual organisms, but only under favorable conditions.

Increased frequency of deletions means increased risk of lethal events. If the strength of selection is related to size of the deletion, long, but rare, deletions will be strongly favored over short, but frequent, ones. This is amplified by recombination, since the increased risk of lethal mutations cannot be mitigated by moving the deleterious deletion to another cell.
Figure 5 shows the fate of deletion rate mutants in a population where one cell division out of ten involves sexual recombination. As can be seen, the density of essential material in the genome proves to be crucial for the outcome. There appears to be a threshold density where the deletion rate is essentially unselected. At densities higher than the threshold, selection favours deletion rates lower than the wild type; and at lower densities, the situation is reversed.

**Conclusions**

Firstly, it is important to note that this manuscript is only in its early stages, and all results are to some degree preliminary.

The results show that the efficiency of indirect selection is inversely related to the frequency of sexual recombination events. In organisms where there is little time for selection between intra chromosomal recombination events, the selective advantage of a deletion rate mutant is greatly reduced, or lost entirely. However, in genomes that have very small fractions of essential material, higher deletion rates will be favoured even if the recombination frequency is high.

Selection on the deletion rate proves to be sensitive to the density of essential genetic material in the genome, a property that is directly affected by deletions taking place. Thus, there is a feedback loop caused by the link between deletion rate and density separates the deletion rate from other similar modifier genes. However, this feedback will not necessarily mean that the genome size will gravitate towards the threshold density. Due to the selective
pressure, deletions will still accumulate faster than insertions, unless the insertion rate is very high compared to the selection coefficient of an in/del event.

A further complication introduced by sexual recombination is that the deletions created in mutant cells will, over time, spread to the wild type cells and increase their fitness. This will reduce the selective advantage of the deletion rate mutant, and thus decrease the chance that it takes over the population. It can be expected that the likelihood of fixation has a negative relationship with the time the mutant is present in a mixed population. This, in turn, means that deletion rate mutants will be sensitive to the population structure in sexual organisms. Furthermore, the average fitness of the population will increase if deletion rate mutants are prevalent, even if no such mutant reaches fixation. This could be of importance if there is competition between sub-populations.
Concluding remarks

The work presented in this thesis provides insights into a few unresolved evolutionary problems and thus contributes to the overall effort of understanding evolution. I believe that continuing upon this avenue of research is highly important for biology in the near future. Comprehending the details of evolution is essential, both as a basis for biological research in general and for preserving the credibility of the field. The latter point is probably more important now than it has previously been, given the recent upsurge of the intelligent design movement. Thus, I feel that refining the scientific foundation is one of the crucial issues for evolutionary biology during the upcoming decade or so.

I also feel that mathematical modeling as a tool, could, and should, be applied more widely in biology, as it often provides options that are otherwise not present. It is my hope that this thesis provides some indication of the usefulness of modeling work, and might contribute to a more widespread use of mathematical models.
Modellering av genomevolution

Evolution är en nödvändighet. Alla organismer forplanta sig genom att kopiera sig själva, och all kopiering innefattar fel i någon utsträckning. Detta leder till variation mellan individer. Eftersom biosfärens resurser är begränsade kommer dessa individer att konkurrera med varandra, och de som är mest framgångsrika i konkurrensen förbättrar sina chanser att forplanta sig vidare. Så enkel är evolutionen i princip, men trots det vet vi ganska lite om hur evolutionära förändringar inträffar i detalj.

Till stor del beror vår okunskap på att evolutionen verkar över så långa tidrymder att processen är svår att observera direkt. För att komma tillräckligt med det problemet måste man använda sig av andra alternativ. Ett sätt är att jämföra olika närbesläktade arter, och använda deras skillnader och likheter för att förstå vad som har hänt sedan deras sist gemensamma förfader levde. Den typen av analys har gett många insikter, men det är svårt att få insikt i vilka mekanismerna är när man bara kan studera slutresultatet. En annan möjlighet, vilken är den som använts i denna avhandling, är att konstruera en modell av systemet och sedan simulera förloppet med hjälp av modellen.

Arbetet är inriktat på generella frågeställningar rörande evolutionära förlopp. Metoden är att formalisera modellerna i matematiska termer, och sedan använda dessa för att genomföra simuleringar. På så sätt kan man studera förlopp som i verkligheten utsträcker sig över tusentals eller miljontals år.

De teman som behandlas är: hur funktioner förändras eller skapas, hur en populations struktur kan påverka ansamlande av skadliga mutationer och effekten av variation i deletionshastighet (hur ofta genetiskt material går förlorat). Resultaten visar att amplifiering, d v s skapandet av ett stort antal kopior, av gener är en viktig mekanism för att förändra eller skapa nya funktioner. När antalet kopior av en gen ökar, ökar också sannolikheten för att positiva mutationer inträffar. Dessutom kan den selektiva effekten av en förbättrad gen förstora genom kopiering, vilket minskar risken att den går förlorad genom genetisk drift.

Vidare framgår det att ansamlandet av skadliga mutationer, en process känt som “Müller's ratchet”, är tydligt påverkad av populationens struktur. Mitt arbete har inriktat sig på en struktur som är typisk för bakterier som lever som symbionter i insekter. Populationen inom varje insekt är separat, och har inget utbyte av individer med omvärlden. Var för sig är dessa små
populationer  utsatta för “Müller's ratchet”, men simuleringarna visar att konkurrensen mellan insekterna leder till att bara de friskaste populationerna av symbionter fortlever, vilket stoppar ratchet-effekten sett över hela populationen.

Evolution av deletionshastigheter är fortfarande oklar i mycket, men de preliminära resultaten visar på tydliga skillnader mellan organismer med respektive utan sexuell rekombination. Eftersom effekten av en mutation som ökar deletionshastigheten utgörs av andra händelser (de deletioner som inträffar allt eftersom) kommer effekten att bero på hur länge dessa sekundära mutationer förblir kopplade till den primära mutationen, d v s den som påverkar deletionshastigheten. Resultaten tyder också på att densiteten av essentiellt material är avgörande för om selektionen favoriserar högre eller lägre deletionshastigheter vid ett givet tillfälle.
Acknowledgments

My time at the department has been a great experience, I could not have wished for a better group of colleagues. Since my very first day, I have felt at home. Thus, I extend a heartfelt thanks to all the people, past and present, who have made my PhD-studies a pleasure.

Naturally, extra thanks are directed to Otto, my supervisor, who has been tremendously helpful throughout my stay. If he had not provided me with the opportunity to pursue the, admittedly odd, kind of biology that interests me, I would most likely not be in science today. In addition, his insight and attention to detail has repeatedly saved my research from going astray. Again, thanks.

Next are the rest of the modeling group Jonas and Sean. Sean, being something of a legend, needs no introduction but deserves thanks for adding color (baby-blue) to the work day, for being a generally good man and, finally, for assisting me in courting my, at the time, wife-to-be. Jonas, while less spectacular, is no less a good man. Thanks for many long discussions, sometimes even about science, good games and general friendship. May the force be with you, always.

While not group-mates, the rest of the PhD-students at the department have been wonderful colleagues, and have provided me with a never ending flow of laughs, smiles and interesting stories. Beginning with the old guard, who have now moved on: Haleh, Olof, Björn I, Ivica, Wagied, Carolin and Lisa have my thanks for welcoming me into the department, despite having had to suffer me as a student previously. Almost, but not quite, gone are Hillevi (you owe me one), Magnus and Hans-Henrik, awesome researchers, who has impressed me since the very beginning. They are also the last who know the true meaning of processivity, and I trust them to protect the secret with their lives.

The younger generation, Björns II and III, Eva, Johan, Kasia and, unfortunately absent, Karin E have, reversely, have had to suffer me as a teacher rather than a student. I apologize for any trauma inflicted, and thank you for remaining friendly. With luck I might even have passed on a useful idea or two, in between all the random modeling nonsense.

Thanks are due also to Håkan for, generally, keeping the computers alive, Ola for making pretty pictures and horror movies, Siv for running the place and recruiting my wife, Karin O for truly running the place, Rolf for adding
scientific passion, **Micke** for being the department's connoisseur and **Robert** for providing the old-school touch.

Due to space constraints (read memory deficiencies) not everyone who has worked at the department can be mentioned by name. I simply repeat my thanks to everyone who has graced the department with their presence.

On a separate note I thank my friends **Petter, Kalle, PJ, Erik, Lars, Jonas P, Tobias B** and **Tobias N.** I have many fond memories of the years we spent together. I thank **Petter** for catching crayfish with me as a kid, and, all these years later, being my best man. **Tobias B** gets extra thanks for his persistence on the badminton court, despite years of dismal results. **Tobias N**, the master of the arcane disaster, get thanks for years of role playing games, a sunken kayak and his ambitious, and amusing, schemes. It's a shame we never finished the trebuchet. I thank **Lars** for the coffee breaks. **Kalle, PJ** and **Erik** get additional thanks for not throwing me into the river. Yet.

Most important of all, I thank my family, **Olga** and **Björn IV**, for their, surely undeserved, love and support. Without you, there is no doubt I would never have finished my studies. I am a very lucky man. Again, thanks to **Sean** and **Siv** for match-making efforts. Thanks to my **mother** for constantly worrying over my well being, my **father** for leading by example and getting me interested in science, my **sister** and **brother** for being good people, and therefore not making my responsibilities as the oldest sibling too difficult to deal with.

More formally, thanks to the **National Graduate School in Scientific Computing (NGSSC)** and the **Swedish Research Council** for funding my position.
References


Hall, B. G. 1990. Spontaneous point mutations that occur more often when advantageous than when neutral. Genetics 126:5-16.


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 324

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)