UPTEC X 02 019 ISSN 1401-2138
APR 2002

JOHAN WIMAN

An evolutionary simulation of linkage between duplicated genes under the theory of subfunctionalization

Master’s degree project

Molecular Biotechnology Programme
Uppsala University School of Engineering

UPTEC X 02 019

Date of issue 2002-04

Author

Johan Wiman

Title (English)

An evolutionary simulation of linkage between duplicated genes under the theory of subfunctionalization

Title (Swedish)

Abstract

Duplicated genes are of central interest to evolutionary genetic research and are believed to be one of the driving forces behind genetic diversification. Yet, the maintenance of duplicated genes in a genome has not been fully explained. Obviously selectional pressure should erase the copy of a gene if it did not have a function of its own. The theory of subfunctionalization explains this by differential promotor mutations that make the two loci dependent on each other and hence obliged to remain. This explanation implies no fitness reduction, even though it uses only deleterious mutation. To understand more about subfunctionalization and maybe to be able to identify it in physical experiments, a model was compiled that simulates how linkage disequilibrium is affected by this theory. When the alleles in two loci are dependent on each other and go together this should leave an imprint in linkage, a disequilibrium. The model starts from a situation with two identical, perfectly duplicated genes and can be used for various experiments under different conditions in an evolutionary context.

Keywords
Duplicated genes, subfunctionalization, linkage disequilibrium, evolution

Supervisors

Martin Lascoux and Per Sjödin
Department of conservation biology and genetics, Uppsala University

Examiner

David Ardell
Department of cell and molecular biology, microbiology, Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN number 1401-2138	Classification
Supplementary bibliographical information	Pages **21**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

An evolutionary simulation of linkage between duplicated genes under the theory of subfunctionalization

Johan Wiman

Sammanfattning

Duplicerade gener är ett mycket viktigt forskningsområde inom den evolutionära genetiken eftersom de tros ha utgjort en drivande kraft bakom evolutionen och den genetiska diversifieringen. Icke desto mindre är mekanismen som förmår bevara duplicerade gener i arvsmassan ännu inte helt känd. Det naturliga urvalet borde ju rimligen selektera bort en kopia av en gen om den inte uppfyller någon egen och unik funktion. Subfunktionaliseringsteorin förklarar detta med att olika promotormutationer kan uppstå som gör de två lokusen beroende av varandra så att båda måste bevaras, genom att olika uttryckssätt kallade *subfunktioner* slås ut i vardera lokuset. Trots att denna förklaring bara använder sig av deletära mutationer i sin förklaring orsakar den ingen reduktion i fitness.

För att kunna förstå mer om hur subfunktionalisering fungerar och för att eventuellt kunna identifiera processen i rella experiment har en modell framställts som simulerar hur kopplingsjämvikten mellan duplicerade gener påverkas av subfunktionaliserings-teorin. När två alleler är beroende av varandra och nedärvs tillsammans borde detta påverka deras kopplingsförhållande och bringa det ur jämvikt. Modellen startar i en situation där man har två indentiska och perfekta duplicerade gener och kan användas för olika experiment under olika förhållanden i ett evolutionärt sammanhang.

Examensarbete 20 p i molekylär bioteknikprogrammet

Uppsala Universitet april 2002

Contents

1. Introduction

1.1 Gene duplication and evolution

Gene duplication has been of central interest to evolutionary genetic research ever since Susumu Ohno published his theories in the book *Evolution of gene duplication* in 1970 (1). In part this is because one believes that gene duplication, or the duplication of whole genetic segments, chromosomes, or whole genomes, has been one of the driving forces behind genetic diversification and the occurrence of new genetic properties.

The process of gene duplication has been assumed to work in the following way: A part of a genome, big or small, is duplicated and since its genes in this region are then present in two copies, one of them is free to evolve by selection, mutation etc into something new, assuming its function of origin is preserved in the other copy. The factor most often thought to lie behind beneficial evolution in a gene is mutation. As will later be given prominence there are however many difficulties with this model of evolution. But before this I would like to mention something about the purposes in this research field.

Molecular evolution is a research area bordering both to population biology and theoretical biology, descending from the groundbreaking works of Darwin on the origin of species. It struggles for two aims: understanding how evolution works at the molecular level and knowing practically what this implies for the biology of the species.

Today our modern science and technique has given us the possibility to access the genetic heritage, which lies waiting like a giant but unsorted historical protocol in front of molecular evolution. Therefore there now seems to be new possibilities to get to know our origin and how we have evolved.

1.2 Mathematically modelling the evolution

However to understand what the genome has to tell us, we need to know the processes that it has underwent. Therefore mathematical modelling has played an important part in evolutionary biology. To make mathematical models that are applicable to genetic data is beneficial, due to the inherently probabilistic pattern of evolutionary processes. An exact solution can however not be expected other than in a probabilistic sense, since random processes always plays an important role in evolutionary processes. An even very small possibility has to be considered, since it in an evolutionary perspective, according to the law of unrelated events, can grow to be very big with time.

1.3 Example: flowering time in Brassicaceae

An example of a relevant application of the theories about molecular evolution is the attempt to understand the mechanisms behind the variations in length of flowering time in the rape relative *Brassica nigra*, in different parts of the world. The genome of *B. nigra* has gone through multiple duplication processes during evolution. An increased understanding in the area could make flowering time changes possible and/or realise tailor-made climate adaptations for the crop (2).

1.4 Theories about gene duplication

However, when trying to decipher in what way the plant's genetic heritage was created a number of problems are brought to head. To start with one can notice that the phenomenon of gene duplication hardly can be discussed without also discussing the phenomenon of gene loss (1).

How can for instance a duplicated genomic segment be preserved instead of succumbing to mutations, when there shouldn't be any evolutionary gain to make in keeping it, since its copy already fulfils its mission? Empirical data show that a much bigger part of the gene duplicates are preserved than what is expected under the classical model (which predicts that all duplications will be deleted in due time)(3).

Some practical explanations could be that the gene might be part of a genetic network, or a hierarchy (a gene cascade) that's been duplicated as a whole (1,4). Another idea is that the gene is needed in multiple copies to obtain faster transcription. Also the background can be important. If a gene is linked to other genes that leave a lot of copies in the next generation, this gene will also leave a lot of descendants. Conversely, if the gene is linked to others that have a low fitness the gene will leave fewer descendants (5). However most models build on the concept that the gene has to evolve in some way, in such way that some kind of symmetry pattern is broken whereon fixation follows.

An additional question is under which circumstances and how often gene duplication is supposed to have arisen, and for how long a duplication can be preserved before getting lost or being fixed in a population (4). Researchers have studied the half-life of duplicated genes in various genomes to find out more about regularities in the chromosomal clockwork (6). However interesting the results, the half-life only partly works as a measure since there are some gene complexes that are so important that they have endured much longer than what is expected from the half-life (7). Thus tandem duplications have given rise to gene clusters, e.g. the beta globin and the hox clusters, which should have been dissolved a long time ago. These clusters are expected to have existed intact for 200 and 600 million years (4).

The term duplication encompasses a large variation in possible extent. It can concern everything from the copying of a whole genome to the duplication of a certain area of a gene. Accordingly all duplications aren’t perfect. This can then lead to gene functions that are only partly overlapping, or to pseudogenes. Aside from this gene functions can be overlapped by totally different genes that didn’t arise from duplications (paralogous genes).

There are multiple theories to explain the present appearance of genomes. Perhaps the effectiveness of transcription regulation puts a higher limit to the amount of genes a genome can contain. More genes could then be allowed if transcription was made more efficient, which would then play an important regulating role for the rate of duplication (4).

More relevant to the project, with which I am concerned, is the question how duplicated genes can be maintained at all in a genome. One can imagine that duplicate gene copies could function as a type of insurance against harmful mutations. We then get a very unusual kind of selection where the genetic redundancy is supported, not because selection acts on the differential fitness between individuals with different amounts of redundancy, but because the risk for its progeny of getting harmful mutations is decreased. In this case genetic redundancy might be enough a reason for a duplication to be maintained (8).

1.5 Duplicate gene preservation by subfunctionalization

For the fixation of duplication to occur in a population most researchers still believes that one of the copies be gradually improved or at least that it diverges (9). But then one has to face the conundrum how improvements by mutation can take place, since these are almost always harmful. Ohno predicts that a diversification of duplicated gene expression in time and space or under other conditions when a modular protein is expressed can explain why a duplicated gene is maintained (4). The theory of subfunctionalization is a related explanation, where a complementary degenerative mutation in different regulatory elements is assumed to make easier double-sided maintenance of a duplicated gene (10). In this way harmful mutations are turned into something diversifying instead of being something destructive.

It is not impossible in a situation of subfunctionalization, that it could be affected by the so-called Hill-Robertson effect. This effect describes the two-locus situation in which selection favours one allele at each locus and in which there is linkage between the loci. The term linkage applies when two genes are prone to be inherited together into the next generation. According to Hill and Robertson selection on one locus hinder the probability of fixation of the beneficial allele in the other locus (5). In the case of subfunctionalization this could eventually hinder differently subfunctionalized alleles to exist in the same time, one in each locus, if they are linked. On the other hand it will also make linked deleterious genes interfere with each other's selective elimination. All these linkage effect are poorly understood under the process of subfunctionalization.

Genetic linkage is a common phenomenon that affects genes that are not fixed; therefore it is mainly discussed in terms of linkage disequilibrium (LD), since in equilibrium all genes are presumed to diffuse freely. If the way of changing of genetic linkage disequilibrium is studied during subfunctionalization, the same behaviour could eventually be identified in nature. Succeeding with this would be strong evidence for the plausibility of the theory. Therefore my purpose with this report is to create a model for evaluating how the theory of subfunctionalization affects genetic linkage.

2. Theory

In the sequel I will describe the theory of subfunctionalization, explain how it could be realised in a mathematical simulation of molecular evolution, then add some words on how the theory of linkage disequilibrium (LD) works and finally comment on how and why it should be affected by subfunctionalization. LD might in fact be a very good instrument for measuring the evolutionary impact of subfunctionalization.

2.1 Subfunctionalization

The theory of subfunctionalization grew out of the notion that a mutation doesn't necessarily have to be a totally devastating incident. Many genes, especially those involved in development, have multiple, independently mutable subfunctions with respect to timing and tissue specificity of expression. This means that if one of these subfunctions is struck by a mutation it can still be well working with respect to its other subfunctions. However there must be some gene compensating the subfunction that was mutated. The duplicate fulfils this charge. Next step in the process leading to subfunctionalization is the duplicate getting mutated also, but in some other subfunction. The individual's fitness hasn't though been affected since the two gene copies are able to handle all their tasks together. But at this stage they have become dependent on each other, and suddenly there is a gain to make in their double-sided maintenance (Figure 1). Lynch and Force believe that the question of how to retain both copies of a duplicated gene could be explained in this way. According to their definition subfunctionalization is the fixation of complementary loss-of-function alleles that result in the joint preservation of duplicate loci (10).

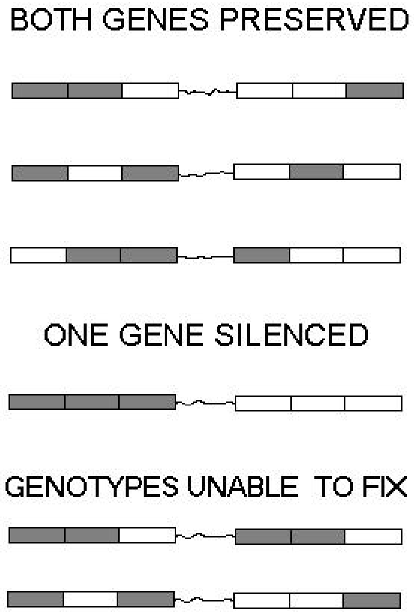

Figure 1. Schematical representation of genotype classes under subfunctionalization.

So subfunctionalization could explain how duplicated genes can be preserved in the genome and moreover does it only by the use of harmful mutations. Beside these two main issues there are other interesting consequences that come with the theory.

Subfunctionalization reduces pleiotropy so that natural selection can more closely tune the duplicate members of a pair to their specific subfunctions (10). Subfunctionalization may provide a mechanism for the development of reproductive incompatibility, i.e. speciation.

2.2 A model for molecular evolution

In a mathematical simulation of molecular evolution it is common to follow the development of a genetic population for many generations. Such a model could be individually based and follow a certain individual chromosome, but could also be statistically based on samples from type-frequencies. The process is partly decided by its initial conditions, but is also affected by some random processes (drift, mutation, and recombination).

In my model a pair of loci with a certain number of subfunctions represents each chromosome. Selection, recombination and mutation are processes that work on the chromosome-frequencies. The fitness of a certain genotype remains intact as long as there are no subfunctions that are totally struck out, i.e. mutated in all of its four loci.

This kind of model is however very resource-demanding when realised, since the number of combinations of alleles grows very fast with the number of subfunctions. If there are n subfunctions in each locus, there will be 2^{2n} different kinds of gametes (allele combinations) and $(2^{2n-1}(2^n + 1))^2$ zygotic combinations, of whom only $2^{2n}(2^{2n} - 1)/2 + 2^{2n}$ differ in genotype. This yields 64 kinds of zygotes for $n = 2$, but in reality there might be many more subfunctions in a gene.

Because of this rapidly growing degree of complexity an individually based model very soon becomes too computationally heavy. Thus, my model has been based on statistics rather than the individual following of each character in a small population.

2.3 Linkage disequilibrium

An important feature of a model is in what way it presents its answers. This can be done in many ways, and we have chosen a way in which the model's impact on a parameter called linkage disequilibrium (LD) is measured. Another important feature of a model is what kind of results from it that should be considered as significant, but this I'll save to a later discussion.

LD is the non-random association of alleles at different loci and the lower its value, the more correlated are the loci. If the LD value is 0.5 the two loci are totally uncorrelated. In the case of a single chromosome organism like the one to be described in my model, LD becomes identical to the conception of physical linkage. Physical linkage determines how close on a chromosome two loci are, and the lower its value, the closer the loci. If the linkage value is 0.5, then the two loci are so far away from each other that they could just as well be positioned on different chromosomes.

Two closely linked loci are likely to be inherited together, hence a duplicated gene that's undergone subfunctionalization and which' gene copies are now dependent on each other should have high a LD value that is deviating. This value of LD could be measured in general or on each kind of gamete, and over all generations or after each and every one. The result that is chosen could then be tested in some appropriate way, for instance with a ?^{2}-test. The null hypothesis is the LD expected in the classical model.

3. The model

3.1 Basics of the model

In this project I have produced a computer model that simulates how the processes of molecular evolution might operate on duplicated genes, according to the theory of subfunctionalization. Initial there is a gene perfectly duplicated and totally unaffected by mutations in either locus. Its way of developing is interpreted in terms of linkage disequilibrium. Further on, the model is zygotic with respect to selection, deterministic with respect to mutation and uses haplotype frequencies that are statistically based. When drift is simulated it operates by random sampling from the statistical haplotype frequency pool.

The simulation was programmed in Matlab because of its wealth of handy commands and since it provides an easy way of avoiding pointer problems, by its total lack of use of pointers. The number of subfunctions *s* can be chosen, just like the number of generations run *g*, the recombination factor *r*, the population size *N* and the mutation constant *u*.

For each generation all the relevant evolutionary processes are passed through, and these are organised in different Matlab functions that are reached from main.

The allele combinations that make up each kind of genotype are represented as numbers in a vector. These numbers represent, when transferred into binary numbers, the genotype's collection of subfunctions in both loci, with a zero standing for a subfunction that's been mutated, and one standing for intact subfunction. In this way all possible combinations of alleles can be represented in what makes up a pool of haplotype frequencies.

3.2 Fitness

The fitness is one for all zygotes that do not have any of the subfunctions mutated at all four loci. For those zygotes the fitness is zero. Random mating and selection is assumed for in the selection phase. Thus each type of gamete is multiplied by all other types of gametes including its own, and with the fitness of the zygote that they together make up. These frequencies are then gauged with the total fitness that is the sum of all gamete frequencies. Thus the selection phase can be summarised by the formula

$$P_i' = \frac{P_i \sum_j^{2^{2n}} P_j W(P_i, P_j)}{\sum_k^{2^{2n}} \sum_j^{2^{2n}} P_k P_j W(P_k, P_j)}$$

where P_i is frequency of genotype *i*, *n* is the current number of subfunctions in each locus and $W(P_{i,Pj})$ is the zygotic fitness.

3.3 Recombination

When it comes to recombination the new frequency of each haplotype is deterministically calculated by first subtracting the fraction *r* that recombine into other haplotypes. Then the fraction r from all other haplotypes that happen to recombine into the first sort of haplotype is added. Thus the formula for the calculation of every new haplotype frequency can be stated

$$P_{ij}' = P_{ij}(1-r) + r\sum_{k}^{2^{2n}} \sum_{l}^{2^{2n}} P_{ik}P_{lj}$$

3.4 Mutation

The mutational step was inspired by that one used by Nowak *et al.* (11). This model for mutation just like the one for recombination and selection is deterministic instead of being based on random mutations, since randomness demands very much computer power. Instead a small part u of all haplotypes is mutated in each generation and distributed over those kinds of haplotypes that carry the new amount of mutations. Thus every kind of mutated haplotype's frequency is given a small contribution in each generation. Double mutations are ignored because they occur with a rate of only μ^2, just like back mutations are. The latter is because the occurrence of a mutation can easily harm a gene, but a mutation back to the original situation demands that the new mutation should hit exactly the same base pair as the first one did, and there are millions of base pairs. Nevertheless the risk for a given haplotype to have a second subfunction mutated is considered only half of it what it was before, if we take the two allele situation as an example. The genotype that already has all its subfunctions mutated of course is excepted from this rule since it cannot be more mutated at all.

3.5 Genetic drift

Because of genetic drift a model simulating evolutionary processes is not totally deterministic. Thus genetic drift is modelled as follows. Sampling N (the size of the population) new individuals multinomially from the haplotype frequency vector does this. These samples are gauged into new frequencies by dividing by N, and these new frequencies are what make up the haplotype frequency vector of the next generation.

3.6 Deterministic model based on statistics

An important question here is whether this statistically made up version of a population can really be equivalent to an individual based model. In a frequency vector representing different sorts of haplotypes the population can be assumed to be effectively infinite. It is because of this that recombination and mutation can be treated as deterministic processes in the production of the gamete pool for the next generation. According to Lynch and Force this kind of measures should give indistinguishable results from those obtained with an individual-based model (10).

3.7 Parameters

When it comes to LD it is calculated as the difference between the frequency of each type of gamete and the product of its allele frequencies in the whole population

$$\mathrm{LD} = \mathrm{P}_{\mathrm{ij}} - \mathrm{p}_{\mathrm{i}}\mathrm{q}_{\mathrm{j}}.$$

Thus LD can be interpreted as the covariance of the gamete frequency for each type of gamete.

An important property of a model is the kind of termination criterion it uses, i.e. what determines if it has fulfilled or failed its goals. My model uses LD as its main guide; therefore the most obvious criteria to be fulfilled is if a gamete becomes fixed, i.e. becomes the only gamete left. The most important criterion though is if the simulation has reached equilibrium and its genotype frequencies have remained the same for many generations. It might be wise though to study this after the simulation has run, in order not to slow the program down.

4. Results and discussion

4.1 Reality: the system depicted by the model

The theory of subfunctionalization was designed to explain why duplicated genes persist for such a long time in genomes. Indeed to prolong the lifetime of duplicated genes is also its effect when modelled. But a model is always a reflection of its system, and in this case the system is reality. Let us therefore take a look at the kind of situations it depicts.

The duplicated genes that we find in reality usually aren't new but have persisted on their places in their genomes for ages. In a model we might be able to deduce this from the very moment of duplication, but in nature what we have is the look of our immediate and contemporary samples, to guide us. A duplicated gene might of course have evolved to some extent because of mutation and other natural processes, but you can still expect that it has either become fixed in the population or that it is under some equilibrium that helps it remain from century to century.

If the process of subfunctionalization has led to fixation of the duplicated gene the situation could be interpreted as a case of extreme LD, but at this stage this will not show since there are no alternatives to the fixed gene. But when the process is still going on it is very probable that there will be LD between different kinds of alleles that cannot go together. Here is one of the main purposes with my model; to study these relationships. LD, which is an important parameter in molecular evolution, hasn't been studied under the theory of subfunctionalization. We hope to be able to show or at least get some estimate of how much LD the subfunctionalization process is likely to cause, by using simulated data, that genes in nature are really divided into subfunctions.

Whatever the situation expected to prevail in nature it should be reflected in the model so that when the model finally interrupts this is because one of the genotypes has become fixed, or because an equilibrium has been reached. In my model only a subfunctionalized genotype can become fixed, since other less mutated genotypes will still be under the mutational pressure. The record of the model has now to be analysed in a way that can help interpreting cases of duplicated genes in nature, as well as the model itself. Under what circumstances is subfunctionalization relevant as an explanation?

To analyse naturally occurring duplication one has to know what information can be taken from natural data. The LD value is easy to get, but is not at all as nuanced as the one received from the model. The model gives a LD value for each existing type of haplotype, i.e. allele combination, after each generation. A duplicated gene in nature only gives the LD value for one haplotype after one generation. What is worse is that relevant haplotypes are not easily identified.

Because a subfunction in the model relates to a certain stretch of DNA, this one of course can suffer from very many different kind of mutations. But all of these are classed the same by the model, since in both cases it takes only one mutation to strike the subfunction out. To cope with this the different sequences will have to be ordered in classes. LD will then have to be measured between the classes, to make the values comparable to the model’s. To the worse, finding the relevant classes might be a very difficult task.

Then when natural data has been stapled in these piles the model data has to be processed, so that it can give us a “LD picture” of what we are looking for. This means that the data of the model should probably be weighted in some way to fit the natural data. How to do this seems to be a delicate matter. Here one will have to experiment. To start with one could of course take the mean over all genotypic LD values. And when taking the mean one should remember that it should be the mean over means from many simulation replicates that have gone to their equilibria.

4.2 Validity of the model

Of course even with weighted LD values it is not obvious at all that the model can always be applied. Therefore an important part of its tuning should aim at determining under what conditions subfunctionalization is supposed to occur. Lynch and Force’ model which should be quite similar implies that this is most likely when $\mu N < 0.1$, where μ is the mutation rate and N the effective population size. In my model the effective population size equals the population size, since there is only one chromosome. When μN is this small, fixation can occur because new loss of function mutations will occur very seldom. The allele in the other locus becomes quite unimportant for selection and the main cause of gene frequency change will be random genetic drift. For vertebrate populations these conditions are fairly often met (10).

In the range $0.1 < \mu N < 10$ the details of selection becomes more important, for instance linkage, which will increase the probability of preservation for a pair of genes according to the Hill-Robertson effect, whereby linked deleterious genes interfere with each other's selective elimination. Also the number of independently mutable subfunctions will increase the probability of gene preservation, since this simply leads to a greater number of paths by which a gene pair can be subfunctionalized (10). Thus in this μN interval maybe the most interesting model datas are to be collected.

When μN becomes greater than 10 these details of the model will no longer make any difference, since mutational conversion becomes such a dominant factor. This will prevent the fixation of subfunctionalized alleles (10). Yet $\mu N > 10$ is not necessarily something common in nature since this means the effective size of the population should be on the order of 10^6 to 10^7 or greater.

All these suggestions remain however to be tested in my own model. The time scale of this study is limited and not enough data has yet been gathered for analyse. Further work must be done to confirm statistically what these suggestions from the Lynch and Force model indicate.

4.3 Realism and assumptions of the model

More generally my model could perhaps be criticised for being idealised. Its starting point with two identical copies of a gene, each one perfect, is of course idealised since it depicts a very special case of duplication. The most realistic thing to do would be to start with only one locus, at equilibrium between selection and mutation, and then double this locus. In this way this equilibrium will not have to be reached. However this will mainly affect the time until fixation, but our purpose here is to study LD. A good side of the starting situation chosen here is that it doesn't force the development into some direction.

Other experiments closely related to mine that could nuance this set of problems could be to imply different mutation rates for the two loci, or to create partial recessivity on certain subfunction combinations. But in the present study then again, the limited time scale of the present study does not allow this.

Are there any assumptions of importance that have been made in this model? The deterministic behaviour of mutation and recombination has already been accounted for when the model was described. Besides this there is no recombination within the loci, only between them. This is because the distance between two loci is so much bigger than the length of a locus. Therefore there is a much bigger risk for a recombination to occur between the loci than within them.

Using random mating is also a kind of assumption. However investigating other kinds of mating is a whole research area and for the results that we obtain here it might be wise to chose random mating, since this makes the results comparable to what other research groups have obtained under this very common assumption.

A slightly more important assumption made is the excluding of gain-of-function mutations. This of course could be viewed as one of the virtues of the subfunctionalization theory: it doesn't need neofunctionalization to explain why the duplicated genes are able to persist. Yet some evidence suggests that gain-of-function mutations might be quite common; maybe even as common as loss-of-function mutations (10). The extent to which my results would have to be modified because of this depends on these mutations' influence on fitness, and hence on natural selection.

However neofunctionalization appears to be ineffective at preserving gene duplicates under the conditions when subfunctionalization according to Lynch and Force is supposed to occur. Neofunctionalization only becomes a plausible explanation for gene preservation when the effective population size is on the order of 10^6 or greater (10).

The eventually most important assumption made is the complete recessivity of mutated subfunctions. This assumption is likely to reduce the probability of duplicate gene preservation. This is because the probability that one member of a pair struck by a mutation that has only a small degenerative effect will be totally nonfunctionalized before the other has also been struck is small (10).

4.4 Experiments

The experiments to be undertaken at first should have the purpose of confirming if the values given by the model are correct. When this is taken cared of one should design experiments that explain when the model apply and when there might be other plausible explanations than subfunctionalization to account for the preservation of a pair of duplicated genes. Since this model in many ways fill the same purpose as the

DDC model proposed by Lynch and Force (10), one has good reason to believe that it will be valid under the same conditions as their's is. This will make a good lead for designing these experiments, but of course here the emphasis should be on watching the LD behaviour. Finally when the model is well tuned in, so that can give a somewhat covering picture of the linkage effects under different conditions, it should be applied to real problems as for instance those faced by the researchers working with *B. nigra,* the black mustard.

At present we are only at the beginning of this venture. Thus the first kind of experiments that should be made are those that confirm the correctness of the model's values and that it is correctly behaving. For the simple case of having only one subfunction in each locus, the model will work as an ordinary model of duplicated genes. Much research has been made with this kind of models and so it will be easy to confirm the model's guiltiness. Christiansen and Frydenberg (12) suggested that the frequency of the totally mutated haplotype in a population would be $\sqrt{\mu}$ at equilibrium. Running the simulation to equilibrium sufficiently many times could easily verify this. These researchers also suggested that the quota P_A/P_B would evolve along a straight line towards a parabola. At equilibrium only drift can alter the quote's value, and only along this parabola. Thus if run many times the model's P_A/P_B quotes at equilibria will form this parabola (Figure 2).

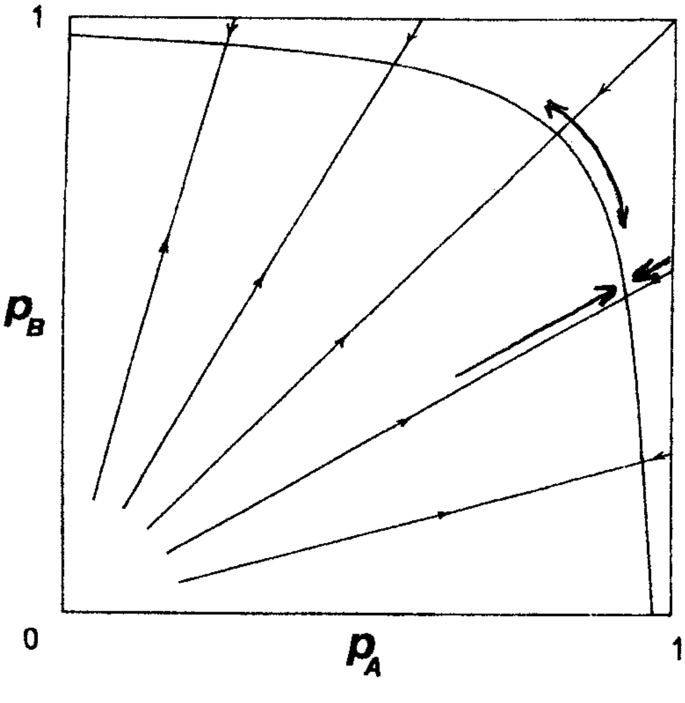

Figure 2. The change in gene frequency through time. Trajectory of any population is restricted to a straight line through origin and point of initial gene frequencies. Population will converge to a point on the hyperbola arching between the points (1 - vμ, 0) and (0, 1 - vμ) (1).

After these basic experiments the models working area should be examined, by varying different parameters, namely the recombination factor r, the effective population size N, the mutation constant μ and most important of all, the number of subfunctions. Primarily the conditions suggested by Lynch and Force concerning different μN values will be investigated. A series of experiments is presented in Table

1. Yet this table of possible experiment certainly is neither an exhausting one, nor a definitive one. Experiments will always have to be outlined along the way.

# subfunctions	r	mut. Factor	N	# replicates
1	0,5	0,001	100	50
1	0,5	0,001	1 000	50
1	0,5	0,001	10 000	50
2	0,5	0,001	100	50
2	0,5	0,001	1 000	50
2	0,5	0,001	10 000	50
4	0,5	0,001	1000	50
1	0	0,001	100	500
1	0	0,001	1 000	500
1	0	0,001	10 000	500
2	0	0,001	100	500
2	0	0,001	1 000	500
2	0	0,001	10 000	500
4	0	0,001	1000	500

Table 1. An outcast of possible experiments.

Other mutation factors could be tried as well. Much more replicates are needed to assure the results when linkage is complete than when it is completely random. However what finally governs the number of replicates needed is how big the differences between the simulations run are.

Each simulation run will have to be replicated many times. It is not impossible that it will be necessary to do as many as thousand replicates under some of the conditions. Others again won't need that many runs since for instance when r is 0.5 LD in each new generation will be independent from the previous one. Therefore these simulations should be undertaken first.

A ?^{2}-test has been tried for testing LD values. If the LD values deviate from the expected, this kind of test will sensor that and rate the deviation's significance. In this way the LD behaviour during subfunctionalization can be charted.

When it comes to reality-adapted experiments, the first thing one should do is to change the above mentioned parameters to those that are found in the species under examination. A big problem is the impossibility of knowing what kind of subfunctionalization lies behind the data given. Therefore this program has been specially created in such a way that the number of subfunctions can be changed, for closer adaptation to linkage data. Unfortunately simulations with more than two subfunctions are very time consuming. Even though making the code slimmer might work, duration time will always be a problem because of the great number of possible genotypes when there is also many subfunctions. However, the behaviour of the model when the number of subfunctions increases is probably generalisable.

4.5 Further development

This model is composed of functions that can be changed independently of each other. This opens for many modification of the model. Different mutation rates for the two loci, or creating partial recessivity has already been mentioned. It would also be possible to infer beneficial mutations under some criterion. It would be interesting to see how strong preservation by the subfunctionalization theory is, compared to beneficial mutations in the same simulation.

Finally, to make the model produce other kinds of data than LD and genotype frequencies can also be easily made with this versatile tool for going deeper into the questions of gene duplication.

5. Acknowledgements

This project was carried out October 2001 to March 2002 at the Department of conservation biology and genetics, Evolutionary Biology Center at the Uppsala University. I am most grateful to everybody at the department, making my stay very pleasant. I wish to thank the following persons in particular:

Per Sjödin for excellent supervision and support at all times.

Dr. Martin Lascoux for supervision and being a friendly boss. Vive la république!

Johan Fogelqvist for being a good and helpful roommate.

Niclas Gyllenstrand for unbelievable jokes and behaviour, setting the style of the department.

Working at The department of Conservation Biology and Genetics is a mere pleasure!

6. References

1. Holland, P. W. H (1999). Introduction: Gene duplication in development and evolution. *Sem.in Cell & Dev. Biol.* **10**, 515-516.

2. Lagercrantz, U. & Axelsson, T. (2000). Rapid evolution of the family of *CONSTANS LIKE* genes in plants. *Mol. biol. evol.* **17(10)**, 1499-1507.

3. Force, A., Lynch, M., Pickett, B. F., Amores, A., Yan, Y. & Postlethwait (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531-1545.

4. Holland, P. W. H (1999). Gene duplication: Past present and future. *Sem.in Cell & Dev. Biol.* **10**, 541-547.

5. Lascoux, M. (2001). Personal communication.

6. Lynch, M. & Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science.* **290**, 1151-1154.

7. Lynch, M. & Force, A. G. (2000). The origin of interspecific genomic incompatibility via gene duplication. *Am. Nat.* **156(6)**, 590-605.

8. Wagner, A. (1999). Redundant gene functions and natural selection. *J. Evol. Biol.* **12**, 1-16.

9. Krakauer, D. C., Nowak, M. A. (1999). Evolutionary preservation of redundant duplicated genes. *Sem.in Cell & Dev. Biol.* **10**, 555-559.

10. Lynch, M. & Force, A. G. (1999). The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459-473.

11. Nowak, M. A., Boerlijst, M. C., Cooke J & Smith, J. M. (1997). Evolution of genetic redundancy. *Nature.* **388**, 167-171.

12. Christiansen, F. B., Frydenberg, O. (1977). Selection-mutation balance for two nonallelic recessives producing an inferior double homozygote. *Am. J. Hum. Genet.,* **29**, 195-207.

13. Ohno, S. (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Sem.in Cell & Dev. Biol.* **10**, 517-522.

14. Wagner, A. (2001). Birth and death of duplicated genes in completely sequenced eukaryotes. *Trends in Genet.***17(5)**, 237-239.

Appendix

The Matlab code

1. Main program

```
clear;

for f = 1:50 %Programdel som används då man vill köra flera
simuleringar på raken.
f
%clf;
%Huvudprogram som anropar underprogrammen

[s,n,r,u,N,g] = matain(5); %5 står där bara för att det måste stå
något.
[A,B] = gametpool2(s,n);
%Xg = zeros(g,2^(s*n)); %X-variablerna har med X2-test att göra
%XT = zeros(1,g);
%Dt = zeros(g,2^(s*n));
SB = zeros(1,2^(s*n)); %Matris där frekvenserna lagras efter varje
generation.
SD = zeros(1,2^(s*n)); %Dito för LD-värdena.

avsluta = 0;
generation = 0;

%for q = 1:g  %Programdel som används då man bara vill köra ett visst
antal generationer.
while avsluta == 0

   [A,B] = selektion(s,n,A,B);
   [A,B] = rekombination(r,s,n,A,B);
   [A,B,sub] = mutation2(u,s,n,A,B);
   [A,B,avsluta] = drift(s,n,N,A,B,sub,avsluta);
   [D,X,Xt,Fgrad] = linkage(s,n,N,A,B);  %Jag har tagit bort g från
inparametrarna.
   %Xg(q,:) = X;  %X-variablerna har med X2-test att göra
   %XT(q) = Xt;
   %Dt(q,:) = D;
   %q
   generation = generation + 1;
   SB = [SB; B]; %Ytterligare ett B läggs till matrisen.
   SD = [SD; D]; %Dito ett D.
end

[D,X,Xt,Fgrad] = linkage(s,n,N,A,B);
D
nummer = num2str(f);
namn = 'datafest,s1,r0.5,u0.001,N100,nummer';
namnonummer = [namn,nummer];
save (namnonummer); %Ett filnamn som stegas upp med replikaten har
skapats.
%Dt(q,:) = D
%plot(Dt)
%Xg;
%XT;
%plot(XT)
%Fgrad;
end
```

2. Parameter input

```
function [s,n,r,u,N,g] = matain(x);

% Funktion som står för gränssnittet mot användaren.

n = 2;  %n är antalet lokus.
s = 1;%input('Mata in önskat antal subfunktioner: ');
r = 0.5;%3.6*10^-7;%input('Mata in rekombinationsfaktorn (värdet
skall ligga mellan 0 och 0,5): ');

if r<0 | r>0.5
   r = input('Gör om, gör rätt!  ==> ');
end

g = 5;%input('Mata in önskat antal simulerade generationer: ');
u = 0.001;%input('Mata in önskad mutationsfrekvens: ');
N = 100;%input('Mata in hur stor population du vill ha: ');
```

3. Gamete pool

```
function[A,B] = gametpool2(s,n);

% Skapar gametfrekvensvektorer F (A och B), av längden 2^(s*n).
Vektorerna innehåller
% frekvensen 1 på sista platsen 2^(s*n). Vektorn A innehåller värdena
före
% manipulering och vektorn B efter.

p = s*n;
A = zeros(1,2^p); % Indexet i binär form slutar ej på ettor om
vektorn börjar på 1!
B = zeros(1,2^p); % Därför har jag minskat binära indexet med -1 när
en omvandling sker.
B(2^p) = 1;
```

4. Selection

```
function[A,B] = selektion(s,n,A,B);
%S=sum(B)
C = B;  %Omvandling av B till A och vice versa skall ske före varje
funktion initieras.
B = A;
A = C;

% Selektion: Fitnessen för varje zygotkombination beräknas genom att
dess subfunktioner
% genomsökes efter "dubbelnollor". Frekvensen för varje element i A
multipliceras med
% varje annats samt med de båda faktorernas gemensamma fitnessvärde.
Därefter lagras
% det nya värdet i vektorn B. Den totala fitnessen Wtot är summan av
alla nya värden
% i B och B divideras nu med detta tal.

p = s*n;

for k = 1:2^p
   for l = 1:2^p

      bk = dec2bin(k-1, p);  %Omvandlar indexet för haplotyp k till
ett binärt tal.
      bl = dec2bin(l-1, p);  %Detta lagras som en textsträng.
      subfunk = 0;

      for q = 1:s
         for w = (q+0):s:(q+(n-1)*s)  %Stegar igenom de binära
indexen subfunktionvis.

            subfunk = subfunk + bk(w) + bl(w);
         end
            if (subfunk == 0)

               W = 0;  %Om fitnessen ska lagras kan man skapa en ny
matris att stoppa in den i.
            %  break

            else

               W = 1;
            end
      end

      B(k) = B(k) + A(k)*A(l)*W;
      W = 0;
   end
end

Wtot = sum (B);
B = B / Wtot;
%S=sum(B)
```

5. Recombination

```
function[A,B] = rekombination(r,s,n,A,B);
%R=sum(B)
C = B;  %Omvandling av B till A och vice versa skall ske före varje
funktion initieras.
B = A;
A = C;
K = zeros(1,2^(s*n));

% Vid rekombination beräknas den nya frekvensen av en viss
haplotypsort efter att rekom-
% binationsfaktorns r bråkdel dragits ifrån, och de nyrekombinerade
haplotyperna av
% sorten lagts till. För den del rekombinanter som också parar sig
med samma sort räknas
% hela summan.

for f = 0:2^s:2^(s*n)-2^s  %Den högra halvan av A's index binära form
(talen > 2^s-1) stegas igenom.
   for g = 1:2^s

      for l = 0:2^s:2^(s*n)-2^s  %Samma sak som ovan göres, men för
indexen för
         for k = 1:2^s           %rekombinanterna som bildar nya
B(f+g)

            K(f+g) = K(f+g) + r*A(f+k)*A(l+g); %Frekvenserna av de
rekombinanter
                                               %som bildar nya B(f+g)
beräknas.
         end
      end

      B(f+g) = K(f+g) + A(f+g)*(1-r); %Rekombinationsfaktorn
subtraheras.
   end
end
%R=sum(B)
```

6. Mutation

```
function[A,B,sub] = mutation2(u,s,n,A,B);

C = B;  % Omvandling av B till A och vice versa.
B = A;
A = C;
%M=sum(A)

p = s*n;
H = zeros(1,2^p); %Vektor där de haplotyper som muteras lagras
temporärt.
sub = zeros(2,2^p); %Vektor där de index vars binärforms summa är >=
s lagras på sin egen
                    %plats, samt i nästa rad denna summa av ettor.

% Mutationen simuleras genom att en bråkdel u för varje subfunktion
överförs till den genotyp
% man får om man byter ut den aktuella subfunktionens etta mot en
nolla i det binära indexet.

for k = 1:2^p  %Frekvensvektorernas index.

   mutanter = 0;
   subfunk = 0;
   bk = dec2bin(k-1,p);  %Omvandlar indexet för haplotyp k till ett
binärt tal.
   bk2 = dec2bin(k-1,p); %Detta lagras som en textsträng.

   for l = 1:p

      subfunk = subfunk + str2num(bk(l)); %Summan av ettorna i det
binära indexet beräknas.
      if bk(l) == '1'

         mutanter = mutanter + 1;
         bk2(l) = '0';
         h = bin2dec(bk2) + 1;
         H(h) = H(h) + u*A(k); %u*A(k) överförs till den nya
genotypen.
      end
   end

   B(k) = B(k) - A(k)*mutanter*u;

   if subfunk >= s %binära index vars summa är >=s

      sub(1,k) = k;
      sub(2,k) = subfunk;

   end
end

B = B + H; %Den temporära matrisen adderas till B.

%M=sum(B)
```

7. Genetic drift

```
function[A,B,avsluta] = drift(s,n,N,A,B,sub,avsluta);
%driv=sum(B)
C = B;  % Omvandling av B till A och vice versa.
B = A;
A = C;

% Driften skapas genom att N gameter samplas från frekvensvektorn.
Dessa skapar
% den nya frekvensvektorn som sedan delas med N så att frekvenserna
erhålls.

p = s*n;
E = zeros(1,2^p);
D = cumsum (A);

for k = 1:N

   gamet = rand; %Ett tal gamet, mellan 0 och 1 dras.
   if gamet <= D(1)

      E(1) = E(1) + 1;
   else
      for l = 2:2^p

         if D(l-1)<gamet & gamet<=D(l) %A's kumulativa summa beräknas
i D. Om gamet ligger i
                                %if-satsens intervall dras en individ
till typen med index l.
            E(l)= E(l) + 1;
         end
      end
   end
end
%L=sum(E)
B = E/N;

% Avslutningsvillkor: om fler än en genotyp med summan av det binära
indexet >= s finns
% kvar fortsätter simuleringen ytterligare en generation. Annars
avslutas programmet.

P = 0;
I = 0;
for t = 1:2^p

   if sub(1,t)>0 & B(sub(1,t))>0 %Om villkoren är uppfyllda ökas P
på.
      I = I + sub(2,t);
      P = P + 1;
   end
end

if P <= 1 & I == s  %Avslutningvillkoret för programmet är att det
bara finns en
                    %subfunktionaliserad genotyp kvar.
   avsluta = 1;
end

%driv=sum(B)
```

8. Linkage

```
function[D,X,Xt,Fgrad] = linkage(s,n,N,A,B);

C = B;  %Omvandling av B till A och vice versa skall ske före varje
funktion initieras.
B = A;
A = C;

% LD beräknas för varje gamettyps frekvens genom att produkten av en
gamets allelfrekvenser
% subtraheras från gametfrekvensen. D motsvarar alltså
gametfrekvensens kovarians. Varje binärt
% index delas upp i två allelsträngar.

p = s*n;
P = zeros(2,2^p);
D = zeros(1,2^p);
X = zeros(1,2^(n*s));
Fgrad = (n*s-1)^2;

for k = 1:2^p

   bk = dec2bin(k-1,p);  %Omvandlar indexet för haplotyp k till ett
binärt tal.
   G = strcat(bk(1:s));  %Första halvan av indexet lagras i G.
   H = strcat(bk((s+1):p));  %Andra halvan av indexet lagras i H.
   E(1,1:s,k) = G;
   E(2,1:s,k) = H;
end

% Allelfrekvenserna beräknas.
for l = 1:2^p
   for m = 1:2^p
      if E(1,1:s,l) == E(1,1:s,m)

         P(1,l) = P(1,l) + A(m);  %Summan av varje allels förekomst i
lokus 1 adderas.
      end

      if E(2,1:s,l) == E(2,1:s,m)

         P(2,l) = P(2,l) + A(m);  %Samma sak som ovan, men för lokus
2.
      end
   end
end

% LD beräknas och testas för varje enskild haplotypsort.
for d = 1:2^p

   if E(1,1:s,d) == E(2,1:s,d)

      D(d) = A(d) - P(1,d)*P(2,d);
   else
      D(d) = A(d) + P(1,d)*P(2,d);  %LD-värdena lagras i D.
   end

   %if P(1,d) ~=0 & P(2,d)~=0
```

```
    %  X(d) = (N*D(d)^2)/(P(1,d)*(1-P(1,d))*P(2,d)*(1-P(2,d))); %Det
individuella
                                                               %X2-
värdet beräknas.
   %end %Ska verkligen N användas? Ja, för N = Ne.
end

% Den övergripande hypotesen att inget D är skilt från noll testas.
Xt=0;
%for l = 1:2^p

 %  if P(1,l) ~=0 & P(2,l)~=0

  %       Xt = Xt + (N*D(l)^2)/(P(1,l)*P(2,l));
   %end
%end
%sum(D)
```