ELSEVIER

FEMS Microbiology Letters 178 (1999) 1–6

FEMS MICROBIOLOGY LETTERS

MiniReview

Toward an understanding of evolutionary potential

Barry G. Hall *

University of Rochester, Biology Department, Hutchison Hall, River Campus, University of Rochester, Rochester, NY 14627, USA

Received 27 January 1999; received in revised form 8 June 1999; accepted 8 June 1999

Abstract

The genome of each organism contains the potential to evolve novel functions that will allow it to thrive in alternative environments. There is not yet a sufficient understanding of the selective constraints on that potential to permit us to predict which genes are most likely to evolve a particular novel function or to predict the mutations that are most likely to lead to that function. Technological advances in the areas of rapid and massive DNA sequencing and in vitro evolution by sexual PCR (DNA shuffling) now make it possible to make an effective start on developing theory that will allow us to assess the evolutionary potential that is present in existing genomes.

Keywords: Evolutionary potential; Novel gene function; DNA shuffling

1. Introduction

The genome of each organism contains not only information for its functioning in the current environment, but the potential to evolve novel functions that will allow it to thrive in alternative environments. That potential information is subject to a variety of selective constraints that limit the organism's actual ability to evolve any particular new function. Currently we do not possess a detailed understanding of those constraints, and we do not even know what kinds of information must be obtained in order to develop that understanding. On paper, the sequence of any gene can be 'evolved' to the sequence of any other gene by a series of discrete mutations that include base substitutions, duplications and/or deletions. In reality, many or most of those mutations will result in gene products that lack any biological function and which will therefore not be selectively advantageous, with the result that the mutations will not be fixed into the population. A detailed understanding of those constraints would be necessary for predicting which genes might evolve a particular novel function, and for predicting the most likely series of changes that would lead to that novel function.

As an example, both the well known *Escherichia coli lacZ*-encoded β-galactosidase and the much less well known *E. coli ebgA*-encoded β-galactosidase enzymes hydrolyze the milk sugar lactose (4-*O*-β-D-galactopyranosyl-D-glucose). Indeed, most of the known β-galactosidases have been identified biochemically on the basis of lactose hydrolysis. A recent phyloge-

* Tel.: +1 (716) 275-0721; Fax: +1 (716) 275-2070; E-mail: drbh@uhura.cc.rochester.edu

0378-1097/99/$20.00
PII: S0378-1097(99)00292-X

netic analysis has shown that *lacZ* and *ebgA* are paralogs that resulted from a very ancient gene duplication [1]. Their divergence dates back to the root of all β-galactosidases, predating the 2.2-billion-year-old divergence of Gram-positive and Gram-negative Eubacteria [2], and thus vastly predates the 100-million-year-old appearance of mammals, the primary source of milk and hence lactose in the modern world. It is therefore clear that the β-galactosidases did not originally evolve to hydrolyze lactose, and indeed we do not know what the natural substrate of those ancestral β-galactosidases might have been. The appearance of mammalian lactose must have provided strong selection for lactose hydrolysis, but the details of that selection process are unknowable. For instance, did evolution of efficient lactose hydrolysis require a reduction or loss in the efficiency with which the original substrate was hydrolyzed? We do not know the natural substrate of the Ebg β-galactosidase, but we do know that it hydrolyzes lactose very inefficiently. We assume that it does have a natural substrate and that it provides some useful function for the cell because the Ebg active-site amino acids have been so well conserved that it is unlikely that Ebg is simply a useless pseudogene [1]. Ebg can evolve efficient lactose hydrolysis as the result of two specific base substitutions (reviewed in [3]), but without knowing Ebg's natural substrate we cannot assess the effect of those mutations on its natural activity. Like the modern Ebg β-galactosidase, the ancestral β-galactosidases included the *potential* to evolve the ability to hydrolyze lactose very effectively.

There are many important reasons to develop an understanding of evolutionary potential in addition to increasing our understanding of evolutionary processes in general and the molecular basis of adaptation in particular. For example, in medicine that knowledge could aid in suppressing the emergence of novel pathogens through the application of effective environmental controls.

Modern evolutionary biology attempts both to reconstruct the history of life and to elucidate the processes that account for that history; the sub-specialty of molecular evolution focuses on the history and causes of evolutionary changes in nucleotide sequences of genes, gene structure, and organization of gene on chromosomes. Recently there has been a distinct shift away from purely descriptive studies to those which attempt to develop a theoretical basis for the observed history. The current paradigm involves developing ever more refined scenarios that explain the evolutionary outcomes we observe as today's life. We require that our explanations are consistent with both what we observe today, and what we observe, albeit dimly, in the fossil and reconstructed DNA record of yesterday. The test of any theory lies not in its ability to explain, however plausibly, previous observations, but in its ability to accurately predict future observations. A stringent test of evolutionary theory is the prediction of evolutionary changes that occur as the consequence of defined selections. For instance, given a set of different microbial species, none of which can utilize resource X, which species is the most likely to be able to evolve the ability to do so? Which genes are most likely to mutate? Which regulatory circuits are most likely to change? At a finer level, which amino acids in each of the candidate proteins are most likely to change? These answers, and more, are required to determine the potential a particular organism has to evolve new functions.

Today many organisms are evolving rapidly as populations adapt to the huge increase in human populations and the associated environmental changes. It would be very helpful, for instance, to understand how the HIV virus is *likely* to respond to the selection imposed by the widespread use of new drugs or combinations of drugs. It would be very useful to understand the likelihood that an attenuated organism which is used in a live vaccine will re-acquire virulence. A pharmaceutical company must increasingly consider whether and how target organisms might develop resistance to drugs. As more manufacturing processes utilize microorganisms, it becomes more important to understand how those microbial populations are likely to evolve in response to the new ecological niche created by the manufacturing process. For instance, paper manufacturers are starting to develop closed systems for handling the waste water involved in making paper. When the input material includes recycled, microbially contaminated paper it becomes important to know not only the composition of that microbial community, but how that community will evolve in the closed loop system. As more and more rivers,

ponds, and lakes become burdened with toxic wastes it becomes essential to understand how the microflora and fauna in those systems will evolve to adapt to the wastes. We may not be able to depend upon existing microbial communities to eventually clean up those wastes, and may therefore have to accelerate the evolutionary process. Understanding the evolutionary potential that lies within the existing communities will allow us to decide on rational courses of action.

Microbial systems have been used to study evolution experimentally for over 30 years [4] because they are both simple and powerful. In each case the strategy has been to apply selection for a defined new function to large microbial populations and to see what happened. In no case has there developed a theoretical framework suitable for predicting a priori which genes would mutate and the nature of those mutations. There are two reasons for the absence of those predictions: (1) we are only now determining the complete sequences of enough microbial genomes to really know what constitutes the starting point, and (2) even with that information we really do not know what questions we should be asking in order to make those predictions. The matter is further complicated by our realization that there is often more than one molecular solution to any given functional problem.

Two powerful new technologies now make it possible to begin developing a theoretical basis for predicting evolutionary potential. The first is the set of technologies for rapidly sequencing massive DNA molecules. Those technologies are permitting the sequences of entire microbial genomes to be determined within the space of a few months. The second is sexual PCR, or DNA shuffling, that permits the simultaneous introduction of multiple mutations into DNA molecules, and the subsequent reassortment and recombination of those mutations to generate vast libraries of multiply substituted DNA molecules.

The problem of determining whether organism X is likely to be able to evolve the ability to utilize resource Y can be broken down into two separate problems: (a) determining which gene is most likely to evolve the required function; and (b) determining whether that gene can actually mutate to provide the required function. DNA shuffling (sexual PCR) provides the means to accomplish step b, but because it is time and labor intensive it can only be applied to a small number of genes; hence the requirement for step a.

2. What kind of information best predicts which gene will evolve?

At this point we do not know what information best predicts which genes are most likely to evolve a particular new function, but our present understanding does suggest several distinct kinds of information might be valuable.

2.1. Presence of the function in a related organism

At the crudest level, we might begin by asking whether a closely related organism possesses the function of interest. For instance, neither *E. coli* nor *Salmonella typhimurium* is able to utilize cellobiose, but their close relative *Klebsiella* can do so [5]. On that basis one might predict that both *E. coli* and *Salmonella* have the potential to evolve that capability, a prediction that is in fact borne out [6,7]. Such predictions are strongly dependent both on good taxon sampling and on construction of reliable phylogenetic trees. Where sampling is sparse, or tree topology is uncertain, the value of such predictions is minimal.

2.2. Genome sequence information

As of the spring of 1999 the complete sequences of 18 microbial genomes have been determined and another 32 are at various stages of completion. It is reasonable to expect that many more genomes will be added to the list in the next few years, and that those additions will be dominated by medically and environmentally important organisms.

If one or more genes encoding the function of interest are in the one of the databases, then those sequences may be used to screen the genome of interest for homologous genes. Indeed, in sequenced genomes only a minority of the open reading frames (ORFs) can be assigned functions on the basis of direct genetic, biochemical or physiological evidence. Most ORFs are assigned putative functions only on

the basis of sequence homology with genes whose functions have been determined experimentally.

In the organism of interest there may well be several homologs of the functionally related gene that was used as a probe. Which of those homologs is most likely to evolve the function of interest? For instance, the *malH* gene of *Fusobacterium mortiferum* encodes a phospho-α-glucosidase that acts on phosphorylated maltose [8]. If we are interested in whether *E. coli* can evolve a phospho-maltase, probing with the MalH protein sequence identifies three homologs in *E. coli*: a phospho-β-glucosidase (CelF), an α-galactosidase (MelA), and a truncated gene GlvG. Which is most likely to evolve a phospho-α-glucosidase function? Intuitively we would expect that the most closely related sequence would be the best candidate, but that is not necessarily the case. Although it is the closest relative of the *F. mortiferum* MalH, GlvG is an unlikely candidate because it is truncated at about the midpoint of the gene. CelF is 40% similar to MalH, while MelA is 33.4% similar, a difference that is too small to make a confident prediction.

2.3. Biochemical information

When sequence similarity fails to distinguish a single strong candidate gene biochemical information may help resolve the issue. For instance, both MelA and CelF of *E. coli* could be cloned into expression vectors and the expressed proteins assayed in vitro for phospho-α-glucosidase activity. The protein with the highest specific activity would be the preferred candidate for evolution of the new function.

3. Using sexual PCR to assess the evolutionary potential of a candidate gene

Once the best candidate gene is identified it is necessary to determine experimentally whether the assumed evolutionary potential can be realized. In simple cases, such as the Ebg system [3], that potential can be evaluated by in vivo selection for the new function because only one or two mutations are required to realize the potential of the Ebg gene. When more mutations are required practical considerations of time, labor and materials preclude using in vivo selection.

When simple in vivo selection fails to detect the assumed potential it means either that the assumed potential does not exist, or that more than one or two mutations are required to reveal that potential.

A recent, and, I think, much undervalued approach to in vitro experimental evolution involves using sexual PCR. Sexual PCR involves PCR amplification of a candidate gene under error-prone conditions to introduce a variety of base substitution mutations into the gene, followed by an error-prone shuffling or recombination process that results in a population of molecules that contains a large number of combinations of those substitutions [9]. Using that method Stemmer obtained a 32 000 fold increase in activity toward cefotaxime as the result of six amino acid substitutions in the TEM-1 β-lactamase [9]. Because the mutations were distributed over a large region, that increase in activity could not be obtained by the usual 'cassette' mutagenesis in which one small region is mutagenized. Stemmer has since used the same approach to evolve a β-fucosidase from the classical *lacZ* β-galactosidase, increasing the specificity for fucosides over 1000 fold by six amino acid substitutions [10], and Yano and his colleagues have directed the evolution of aspartate aminotransferase to increase its catalytic efficiency for β-branched chain amino acids 10^5 fold [11]

The power of sexual PCR is that it makes no assumptions concerning the changes in amino acid residues that are most likely to give the desired phenotype when modified. By selecting from the transformed populations those that grow fastest on lactose, it is possible to identify those substitutions that engender the maximum possible activity toward lactose, i.e., the highest peak in the in vivo fitness landscape.

The best enzyme is likely to include both those substitutions that confer maximum activity and additional neutral substitutions that are irrelevant to improved activity. The neutral substitutions can be eliminated by a 'shuffling back-cross' [9] by shuffling DNA from the 'best' isolate in the presence of excess wild-type DNA and again selecting the fastest growing isolates. A modification of the shuffling process results in high-fidelity shuffling that introduces essentially no new mutations [12], and the method has

been used to distinguish the two functional mutations among the 10 mutations present in an artificially evolved thermostable subtilisin gene [13].

Although sexual PCR can dramatically increase the number of variants that can be tested compared with in vivo selection, it does not even begin to explore all of the possible variants. Given a gene of 1000 base pairs there are over 10^{34} sequences that differ from the wild-type sequence by 10 or fewer mutations. Not only can we not explore all of those possible variants, life itself has barely had sufficient time to explore all of those possibilities. The mass of the earth's oceans is about 1.4×10^{24} g. Even if living cells constituted 10^{-4} of the mass of the oceans, given about 10^{12} bacterial cells per gram, a reproduction rate of about 1 cell generation per day and a mutation rate of about 10^{-9} per cell generation and 4 billion years of life there has been sufficient time to explore only 1.6×10^{34} variants of a single 1000-bp sequence. However, evolution does not proceed by exploring all possible variants but by incorporating single mutations, selecting the fittest of those variants, expanding the population of the fittest variants, and incorporating additional single changes.

Sexual PCR experiments generate effective libraries of about 10^7 variants. If there is the typical average of 7 or 8 substitutions per gene, then that library will include all single-base substitutions, about 2.5% of all possible double-mutation variants, and less than 10^{-18} of all possible seven-mutation variants. At the same time, it will include over 7×10^6 variants each of which has 5–10 base substitutions. No other method today permits screening such a large number of multiply mutated sequences. Given the tiny fraction of possible sequences that can be explored experimentally, how can we develop any confidence that the observed outcomes in any way resemble those that would arise by 'natural' evolution? We can accomplish that end by mimicking the natural evolutionary process, i.e., by sequentially introducing into the wild-type sequence each of the mutations that is present in the multiply mutated end product that was identified by selection during the sexual PCR experiment.

Let us assume that backcrossing by high-fidelity DNA shuffling has eliminated all those mutations that are irrelevant to the selected phenotype and that the 'evolved' sequence differs from the wild-type sequence by n mutations that result in effective amino acid replacements. If the 'evolved' sequence differs from wild-type at n sites then there are n possible first-step amino acid replacement mutants. Each of those single mutants can be created by site-directed mutagenesis and the effect on fitness determined by competition experiments. The best (fittest) of those amino acid replacements can be chosen and the $n-1$ possible second-step mutants created, the fittest double mutant chosen, the $n-2$ possible third-step mutations introduced, etc. The effect of this exercise is to mimic an evolutionary pathway in which the fittest single mutant is fixed into the population, that population expands, the fittest double mutant arises and is fixed into the population, etc. Orr [14] has recently shown on theoretical grounds that adaptive evolution is expected to proceed in exactly this fashion in which the first mutations to be fixed are those that have the greatest positive effect. Because real populations are large enough that in any given gene all possible single mutant *are* expected to arise, we can be confident that the 'best' first step mutation which we identified will arise in a real population. If the selective conditions we chose apply, that mutation will be favored, the population will expand, and the second mutation we identified will arise. We can therefore be confident that if we can construct a pathway from a sequence with function 'x' to an 'evolved' mutant with new function 'Y' in the laboratory, then gene X has the real potential to evolve function Y in nature. Because we have not explored even a significant fraction of all possible variants we cannot be confident that natural evolution will lead to the same sequence we obtained in the laboratory. Possibly better mutations than those we found will arise and be selected instead. We can, however, be confident that an endpoint as good or better than the one we obtained can evolve in nature.

What if we cannot demonstrate a pathway to the best evolved sequence we obtained? If that sequence involves six amino acid replacements, we might find that after introducing three replacements, each of which further improves fitness, none of the three remaining replacements improves fitness. Assuming that the final six-mutant sequence is significantly fitter than the triple mutant, that result means that two, or perhaps even three, of the remaining substitutions must be introduced simultaneously to further improve fitness. The simultaneous occurrence of two

or more specific mutations is obviously highly unlikely, but what about the possibility that one of the two mutations will arise, be selectively neutral, but be fixed into the population by drift? Were that to occur the second mutation would quickly be incorporated by selection. The probability that a newly arisen neutral mutation will be fixed into the population is the reciprocal of the population size. When populations are large enough that the probability of the occurrence of the mutation is very high, e.g., the population size approaches the reciprocal of the spontaneous mutation rate, then the probability of fixation is very low. Although neutral variants arise constantly, it is very unlikely that the particular neutral variant we require will be fixed into the population. Thus, unless each of the mutations confers a selective advantage relative to its parent, it is unlikely that the final six-mutant sequence would evolve naturally. In the example above we would conclude that the evolutionary potential may well be limited to the triple mutant.

Sexual PCR, then, offers one way to circumvent some of the limitations imposed by the usual constraints of populations size, mutation rates, and laboratory experimental time on in vivo experimental evolution. Coupled with pathway reconstruction it offers a realistic means to explore and predict evolutionary potential.

4. Conclusion

The primary purpose of this article is to point out that recent technological breakthroughs have now made it possible for evolutionary biology to start making specific predictions about the evolutionary outcomes that result from environmental selections applied to existing genomes. Ten years ago DNA sequencing was both tedious and expensive; now it is so fast and relatively inexpensive that sequences of several complete genomes are being reported each year. The existence of sexual PCR now offers a way to circumvent the limitations of in vivo experimental evolution to test predictions. It is not my intent to suggest that the approaches outlined in this article are the only approaches, or even necessarily very good approaches, to determining what kind of information is required. It is only my intention to suggest that the problem is tractable, that additional technological breakthroughs will make it more tractable, and that the time to **start** on the problem is now.

Acknowledgements

I am grateful to D. Presgraves and C. Jones for comments on the manuscript, and to an anonymous reviewer for thoughtful comments which greatly improved the manuscript.

References

[1] Hall, B.G. and Malik, H.S. (1998) Determining the evolutionary potential of a gene. Mol. Biol. Evol. 15, 514–517.
[2] Feng, D.-F., Cho, G. and Doolittle, R.F. (1997) Determining the divergence times with a protein clock: Update and reevaluation. Proc. Natl. Acad. Sci. USA 94, 13028–13033.
[3] Hall, B.G. (1999) Experimental evolution of Ebg enzyme provides clues about the evolution of catalysis and to evolutionary potential. FEMS Microbiol. Lett. 174, 1–8.
[4] Mortlock, R.P. (1984) Plenum Press, New York.
[5] Schaefler, S. and Malamy, A. (1969) Taxonomic investigations on expressed and cryptic phospho-β-glucosidases in *Enterobacteriaceae*. J. Bacteriol. 99, 422–433.
[6] Kricker, M. and Hall, B.G. (1984) Directed evolution of cellobiose utilization in *Escherichia coli*. Mol. Biol. Evol. 1, 171–182.
[7] Schaefler, S. and Schenjein, I. (1968) β-Glucoside permeases and phospho-β-glucosidases in *Aerobacter aerogenes*: Relationship with cryptic phospho-β-glucosidases in *Enterobacateraceae*. Proc. Natl. Acad. Sci USA 59, 285–292.
[8] Thompson, J., Robrish, S.A., Bouma, C.L., Freedberg, D.I. and Folk, J.E. (1997) Phospho-β-glucosidase from *Fusobacterium mortiferum*: Purification, cloning, and inactivation by 6-phosphoglucono-δ-lactone. J. Bacteriol. 179, 1636–1645.
[9] Stemmer, W.P.C. (1994) Rapid evolution of a protein in vitro by DNA shuffling. Nature 370, 389–390.
[10] Zhang, J., Dawes, G. and Stemmer, W. (1997) Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. Proc. Natl. Acad. Sci. USA 94, 4504–4509.
[11] Yano, T., Oue, S. and Kagamiyama, H. (1998) Directed evolution of an aspartate aminotransferase with new substrate specificities. Proc. Natl. Acad. Sci. USA 95, 5511–5515.
[12] Zhao, H. and Arnold, F.H. (1997) Optimization of DNA shuffling for high fidelity recombination. Nucleic Acids Res. 25, 1307–1308.
[13] Zhao, H. and Arnold, F.H. (1997) Functional and nonfunctional mutations distinguished by random recombination of homologous genes. Proc. Natl. Acad. Sci. USA 94, 7997–8000.
[14] Orr, H.A. (1998) The population genetics of adaptation. The distribution of factors fixed during adaptive evolution. Evolution 52, 935–949.