Computational Biology and Bioinformatics
2013; 1(4): 15-21
Published online August 30, 2013 (http://www.sciencepublishinggroup.com/j/cbb)
doi: 10.11648/j.cbb.20130104.11

An analysis of cumulative selection and random drift in the evolutionary origination of novel protein motifs

Joseph E. Hannon Bozorgmehr

School of Informatics, University of Manchester, Manchester M13 9PL, United Kingdom

Email address:
bozorgmehr@hotmail.co.uk

To cite this article:
Joseph E. Hannon Bozorgmehr. An Analysis of Cumulative Selection and Random Drift in the Evolutionary Origination of Novel Protein Motifs. *Computational Biology and Bioinformatics*. Vol. 1, No. 4, 2013, pp. 15-21. doi: 10.11648/j.cbb.20130104.11

Abstract: Despite the fact that all-important protein motifs encoded in gene sequences are generally considered to have been generated solely through a gradual and undirected process of change, the idea lacks empirical evidence and also suffers from theoretical difficulties. More recently, sudden developments such as through the exonization of non-coding DNA, as well as frameshift mutation, have been suggested as another source of evolutionary innovation. A mathematical model was used here to describe the combined action of both cumulative natural selection and random drift, relative to that of an alternative mechanism of artificial selection, in the origination of a hypothetical multi-residue sequence motif. A computer simulation of the model quantifiably demonstrates the marked inefficacy of standard forces in developing these molecular features when compared to the power of a directed or self-organizing process. An examination of how natural population shifts, including migration and isolation, as well as intragenic recombination, may serve to facilitate this particular case is also explored. An evolutionary accretion for novel motifs is concluded as being eminently feasible, although not one wholly reliant on the outcome of chance and differential reproduction.

Keywords: Cumulative Selection, Protein Motifs, Functional Adaptation, Directed Evolution

1. Introduction

Perhaps one of the most interesting and challenging issues in molecular biology and evolution is in accounting for the origins of the highly conserved protein motifs/domains that are at the heart of virtually all cellular activities, and which contribute to organismal complexity [1]. These syntactically specific and distinctive amino acid sequences comprise the key functional units that define the nature and behavior of gene products. Whether they are found in such classes and families of protein as transcription factors, dyneins or kinases, their biochemical functionality is often found to be essential [2]. Allowing for certain variations in the peptide sequence itself, most protein domains have been stringently conserved by purifying selection across diverse and ancient phylogenies.

At the genomic level, it has been proposed that, through a process of *cumulative* natural selection, major evolutionary innovations within organisms can take place with beneficial substitutions becoming fixed and accumulating over time, thus leading to improved and adapted organismic features [3]. But at the intragenic level, however, cumulative natural selection is a far more uncertain means with which to explain innovation except when it is limited to the optimization or the readjustment of protein activities in response to environmental pressures [4].

Creating a novel domain *ex novo*, as must have happened at some point in the evolutionary past, would have entailed a truly expansive fitness landscape consisting of a comprehensive and diverse array of adaptive valleys and exaptive pathways. It is thus difficult to suppose that the sequences could have been built up stepwise and incrementally over such a rough fitness terrain. It is for this reason that sudden developments such as frameshift mutations [5], or the wholesale exonization of non-coding DNA [6], have been proposed as an alternative explanation to gradualism even though they are both highly problematic.

Artificial selection, however, refers to the deliberate intervention by humans in Nature to ensure that certain traits are represented and preserved. It works with few of the aforementioned constraints. Breeders have been successful in a very short time by selecting spontaneous variations, often regardless of their effect on fertility (e.g. seedless fruit), that increase the utility of animals and plants to society [7]. More generally, any evolutionary process can be directed towards an intended target [8] by availing itself of the imperfect nature of the molecular replication system.

The difficult problem of putative incipient and intermediate states, especially deleterious ones in any development would no longer represent a major obstacle [9]. This is because any selected variation need not necessarily confer an immediate reproductive advantage or fitness benefit. Directed evolution and mutagenesis, whereby researchers engineer novel proteins through their own experiments [10], is indeed considered to be analogous with that of *artificial* selection rather than *natural* selection [11].

But instead of being an entirely adaptive or completely neutral paradigm of biological change, molecular evolution actually entails a mix of both deterministic selection as well as stochastic processes [12]. Explaining the origin of novelty in function has in fact involved invoking the complementary actions of both natural selection and near neutral drift [13]. If (short) sub-sequences of novel functionality can evolve more freely, under a relaxed regime of purifying selection, random drift can thus become a potentially creative mechanism. This is because it can permit the fortuitous combination of certain characters that may together represent a novel sub-motif [14].

Any reasonable hypothesis for the evolutionary origins of protein domains would therefore typically entail a synthesis of both random drift and natural selection working *in tandem* to generate specific sequences, with selection preserving any productive outcomes that result from random drift. Nature would thus reduce, but not necessarily eliminate, the improbabilities associated with chance whilst still availing itself of the immense possibilities that can be thrown up in conjunction with it.

2. Model

2.1. Gene Duplication as a Mechanism for Innovation

Gene duplication, caused by mistakes in either mitosis or meiosis, is widely believed to play an important role in the evolution of novel protein functions [15]. The initial functional redundancy among paralogous genes offers the prospect of a respite in stringent purifying selection by providing a molecular substrate that can be used for natural experimentation by trial and error [16]. This does not, however, preclude more limited functional change within a singleton gene. But the masking/buffering effect offered by gene duplication can provide a possible mechanism for the development of novelty as it allows change to occur due to a significant relaxation of functional constraint.

The model proposed here envisages the emergence of a novel protein motif within a region of a duplicated gene, and evolving as the representative type for a population of any size. Specifically there exists an initial sequence, a region within both copies of the gene, consisting of a string of 30 codons. For the sake of the operability of the simulation, mutations in this region are limited to that of single nucleotide substitutions - by far the most common genetic variation [17] - and with one mutation occurring per iteration. This does not reflect any natural mutation rate as such, but rather those *de novo* changes as and when they happen – which may sometimes be concurrent as predicted by a Poisson distribution [18]. Also, no transition-transversion ratio bias [19] is assumed here: this is another simplification made to expedite the model/simulation but it does not have a bearing on the main objective since it is applicable to both of the cases examined.

Mutations are also equi-probable, in terms of both where they occur in the sequence and the particular nucleic base that replaces the previous one. The simulation ends a run when the pre-determined sequence of the same amino acid / codon length is successfully reached: the model also insists that all of the codons must be substituted to produce an entirely new peptide sequence. Of course, any two unrelated sequences will inevitably share a significant degree of nucleotide homology but it is required that there should be a complete change made in order to evolutionarily traverse from A to B, with no prior advantage. The likelihood of chancing upon the final sequence through an entirely random search is exponentially remote. Assuming there are on average 3 synonymous codons per amino acid residue, the number of rounds expected to achieve this would be $(64/3)^{30} = 7.44 *10^{39}$. Clearly, this is extremely implausible, and this is why any degree of selection must be effective *throughout the course* of the sequence's evolution, and not at the end.

2.2. Cumulative Evolution by Artificial Selection

In the case of cumulative artificial selection, as applied here to the case of incremental molecular evolution, those random mutations in the original sequence that match the evolved target are retained until eventually all have been selected. This happens when both are identical in terms of their transcribed nucleotide and translated amino acid arrangements. The only issue for investigation concerns the number of mutational rounds needed for this to take place. Assuming there is no initial homology between them, as explained previously, this could be estimated as the number of substitutions per nucleotide site, K, multiplied by the length of the sequence, L. Since there are a total of 4 nucleic bases, there are 3 potential replacements available at every site. However, it is not as simple as this because random mutations are still liable to occur at sites where the required substitution has already been selected. These will be discarded, and fail to become fixed, but this does impose an additional cost in the number of rounds expended.

For example, if the base triplet (codon) *AAA* is to evolve to become *TTT*, then the probability of a single substitution matching any particular nucleotide in this target is 1/3 * 1/3 = 1/9. The probability that any one of them is correctly guessed is 3 * 1/9 = 1/3. The number of rounds required for this to happen is simply the inverse, i.e. 3. Once one of the letters has been selected, the mutational probability for that letter remains the same since changes can still occur at the site of the correctly matched substitution. However, as there are only 2 letters now remaining to be altered, it would now become 2/9 and so take 4.5 rounds on average. The final

letter will then require 9 rounds. The number of rounds, R, expected with which to iteratively reach the specific sequence through a process of cumulative selection and random mutation, therefore amounts to 16.5 on average. In general, this can be represented by the formula below:

$$R = \sum_{m=1}^{L} (KL/m) \quad (1)$$

It follows from this that there exists an expansion for *m* of the order: 1+1/2+1/3… This is, of course, the divergent harmonic series [20] and it can be determined as:

$$R = \sum_{m=1}^{L} (KL/m) = KL\,(ln(L) + \gamma) \quad (2)$$

(Where γ is the Euler-Mascheroni constant and ε is the error factor, inversely related to *L*, but which can for practical purposes be ignored especially when *L* is large.)

Therefore the average number of rounds expected for a sequence to evolve under cumulative *artificial* selection, where individual nucleotides are preserved, amounts to the following:

$$R = \sum_{m=1}^{L} (KL/m) = KL\,(ln(L) + \gamma) \quad (3)$$

2.3. Functional Reducibility and Evolvability

Alternatively, evolution may proceed according to that of cumulative *natural* selection and random drift. Instead of each particular base pair being selected, as above, the model requires that only multi-nucleotide elements of adaptive value, when translated as amino acids, can be preserved. It has been suggested that protein motifs may be multifunctional, and that any constituent elements may even compete for sequence space amongst each other [21]. For example, although the DNA-binding homeobox domain consists of about 60 residues, only six of these amino acids actually make contact with the major groove of the DNA molecule [22].

Because of these observations, it is quite possible that protein motifs may *not* be irreducibly complex and so can be divided into more basic elements, each with its own separate function, but which nonetheless synergistically combine with others to become a complete entity. In this regard, natural evolution may adopt a divide and conquer approach that allows random drift to chance upon islands of functionality within sequence space whilst negative selection then preserves these as part of a cumulative process of adaptation and evolutionary accretion. As such, this can offer the possibility that the problem of originating a complex and specific motif can thus be broken down.

Moreover, research has demonstrated that elements consisting of only a few key residues can have a significant effect as far as biochemical functionality is concerned [23]. For example, transaldolase is an enzyme that is part of the pentose phosphate pathway that is involved in the production of ribose. Only three residues within its active site actually provide the chemical means for catalysis [24]. Moreover, due to the possibility of exaptation, some elements may originally have served one particular function but subsequently came to serve the one that ultimately survived. Therefore, the model for an entirely natural evolutionary development proceeds by partitioning the 30 residue (90nt) target into 10 distinct and contiguous (though not necessarily contiguous) elements, each being three residues long (9nt), as is evident in the representation of a translated sequence shown below where each functional element is separated with hyphens:

RVQ-EFL-PYW-MNP-AGT-EFD-SHK-EMQ-ASL-IYC

For the initial sequence in the region of the duplicated gene to change substantially, a large measure of relaxed selection is assumed, as previously stated. This allows mutated sites to freely and fortuitously combine to arrive upon particular tri-residue sub-motifs within the sequence. All substitutions, except for nonsense mutations that lead to the truncation of the open reading frame, are treated as neutral and allowed to fix consecutively. It should be noted that neutral variations may not actually have to reach fixation in order to survive but can instead "tunnel" across states, albeit at the risk of being lost from the gene pool [25]. However, the model does not account for any heterogeneity – new mutations at the nucleotide sites simply replace the previous ones. Consecutive fixation by neutral evolution does, of course, happen and so the presence of allelic diversity within a real population does not detract from the overall objective of this study.

Once all of the codons representing each of the three amino acid characters are in sync with each other, any further changes can thereafter be discarded by negative selection which preserves the functional coded element. This process will continue until the translated sequence matches that of the target motif. No speculation as to the putative selection coefficients needs to be considered, only that the adaptive benefit bestowed by the element can be assumed to be strong enough so as to guarantee eventual fixation in the population, and also in a relatively short time.

2.4. Probabilistic and Structural Constraints

In breaking down the problem this way, however, another issue then immediately arises. Instead of taking $(64/n)^3$ iterations, where *n* is the average number of synonymous codons in the sequence, all of the changes have now become partitioned among ten separate elements. The number of rounds expected for each tricodon element's evolution increases tenfold, as is represented below:

$$R \approx 10\,(64/n)^3 \quad (4)$$

In this respect, the 9nt elements (i.e. 3 combined sets of nucleotide triplets) evolve in pseudo-parallel, each one assuming a "slice" of mutational activity just as concurrent tasks would share resources in the case of an operating

system. As shown in equation 4, due to the fact that mutations are liable to occur across the full length of the sequence, the entire process would therefore need at least 10 times as many rounds as a result of splitting up the inherent complexity of the motif sequence.

Compounded to this is the fact that the expected probability distribution for the number of rounds required to evolve each of the constituent elements would be typically Gaussian in nature [26] – i.e. like that of a bell curve. If there is a near zero kurtosis (peakedness), then the coefficient of variation should be about 1.0 - with both the mean (μ) and standard deviation (σ) at parity. The upper and lower limits of the distribution are usually within 2-3 standard deviations from the mean. For example, if the average expected number of rounds to evolve one of the elements is 20,000, then the longest one might be as much as 80,000. This could have the effect of holding up the development of the sequence if the order of the stepwise process is a necessary factor.

The evolution of a motif that is compartmentalized in nature would not be a haphazard development devoid of any coordination and its own specific context. A cumulative process that improves reproductive fitness necessarily entails a gradual build-up, i.e. a stepwise progression, rather than one in which each constituent element is physically and functionally detached from one another. There exists the issue of molecular epistasis, and also the possibility of an adaptive conflict, affecting the complex biophysics of protein stability and folding that typically involves the synergistic interaction of amino acids [27]. While a large degree of functional independence is already assumed, the evolution of these elements may be achieved only as part of an additive and orderly process in response to adaptive requirements. This degree of *relation* between all of the constituent elements naturally imposes a greater cost in terms of the amount of trials expended to achieve the final outcome. If a sequence of K possible characters and of length L, as previously mentioned, is developed in series through random substitutions the rounds needed to guess just one of the letters is simply the product: KL. If cumulative *artificial* selection were to be proceed in a series order, it follows that the number of rounds required to obtain all of them is $KL(L)$, i.e. KL^2. It is thus propitious to determine the ratio between this and the process involving no order whatsoever, given in equation (3), because this should reveal the extent to which the number of rounds taken is amplified due to the need for a successive process:

$$KL^2 / (KL(ln(L) + \gamma)) \quad (5)$$

This ratio can then be simplified, with K eliminated, as is shown below:

$$L / (ln(L) + \gamma) \quad (6)$$

The formula could be also used if, instead of individual nucleotides, the characters now represent each of the tricodon elements. Therefore, it is expected that the more natural series expansion should be proportionately greater than the purely artificial one and also dependent on the length of the sequence. Generally, an approximation of the average number of rounds expected in order to evolve a nucleotide string of size L characters should be possible but with it now divided into functional elements of length M codons. If the number of constituent elements is $L/3M$, then the rounds required without any constraint of order can be derived from equation (4). If the upper limit of the distribution is determined as being no more than 4 times greater than the mean it then becomes:

$$R \approx (4\,(64/n)^M)\,(L/3M) \quad (7)$$

However, if the motif sequence is generated through an orderly process, then the number of rounds required involves multiplying equation (7) by equation (6) but substituting $L/3M$ for L.

$$R \approx \frac{(4\,(64/n)^M)(L/3M)\,(L/3M)}{(ln\,(L/3M) + \gamma)} \quad (8)$$

This can then be re-arranged more simply as:

$$R \approx \frac{(4L^2)\,(64/n)^M}{9M^2\,(ln\,(L/3M) + \gamma)} \quad (9)$$

As L, and particularly as M increases, it becomes apparent that the number of rounds required to produce the motif becomes substantially greater. In Table 1, the predicted outcomes of both natural and artificial selection, based on equations (3) and (9), are respectively calculated.

3. Methods

The simulation algorithm was written in GNU C++ and entailed iteratively substituting randomly selected positions within a pre-determined 90 character string, as explained in the model. Nucleic bases A, C, G and T were represented as the characters and made equally probable to occur by mutation using a default pseudo-random number generator. The program was set up to run until all of the constituent elements of the specified target motif sequence had been matched when translated as a peptide chain. Exactly 1000 runs were used to determine the average number of rounds taken. In the case of natural selection, both the mean and standard deviation for the distribution of the individual elements were calculated based on a count of the number of iterations taken for each of the encoded elements to reach its (translated) tri-residue part of the motif sequence.

4. Results and Discussion

4.1. Simulation Outcomes for Both Mechanisms of Selection

For the given sequence of size 90nt, the number of

mutational rounds required to select it artificially was simulated to be 1370 – exactly as predicted by equation (3). This is certainly slow, and it takes on average 15 rounds for each letter in the sequence to match its corresponding target character. Conversely, the average number of rounds taken to evolve the target through a solely natural process of selection and drift was simulated to be 1,089,724. The mean (μ) for all of the 9nt elements was 268,464 while the standard deviation (σ) was 333,437 – i.e. somewhat greater. The overall number of rounds till completion was equal to that of the particular tri-residue element that took the most number of substitutions with which to emerge. This upper limit was found to be exactly 2.5 standard deviations distant from the mean which is not unusual.

When the individual elements had to evolve within the confines of a successive order, i.e. one after the other, the number of rounds became 4,522,374 on average – approximately four times more than for that without any necessary order. This is the value predicted by the model when n=2 as is shown in Table 1, below. It therefore takes as many as 50,000 substitutions for just one nucleotide, in a very short sequence space, to correctly match its corresponding target character and so be retained. As well as the extensive number of attempts required to reach just one tricodon sequence through random drift, the dispersive effect associated with the normal distribution, and the complicating factor of an expansion in series, can serve to amplify this process by up to a hundred times.

As Table 1 clearly indicates, moreover, if the number or the size of the functional elements (as with those consisting of 6 codons) increases, then the number of rounds required to evolve them will increase exponentially. The last entry predicts that it will take almost one trillion rounds to evolve a 540nt sequence, equivalent in size to the DNA-binding T-box domain, where each functional element is 18nt long. This compares with just 10,000 rounds predicted for the alternative artificial manner of selection. The simulation is also generous in assuming that the motif can be reduced to such fractionally small and distinctive components. However, one thing not considered was the possibility that chemically similar residues could, at least initially, serve as substitutes in place of the more exact amino acid. In this way, the number of rounds expended would be less than expected but this would still depend on whether the specific context would allow a sub-optimal residue to confer a selective advantage. In some instances, this might be permissible but in others it would definitely not. Another factor is that of slippage replication [28] which may account for the occurrence of any tandem repeats in motifs much more easily than is so for single nucleotide substitutions though such repetitions would be fortuitous.

The overall difference determined between the natural and artificial processes is thus observed to be very substantive indeed. In the example above, the mechanism of artificial selection was found to be >3000 times more efficient than for the combined action of natural selection and drift, and this difference becomes far greater still as the size of the motif sequence increases. Although the simulation has not been population-specific, the relative size of the group, and the rate of recombination, is likely to change only particularities and not the core finding itself. As the population size decreases, the diffusion approximation [29] predicts that random drift becomes stronger whilst selection tends to be weaker. This means that sub-optimal, nearly neutral, changes have a better chance of surviving and becoming fixed [30]. Conversely, larger populations tend to exhibit greater genetic variation and diversity than smaller ones [31], although the precise population dynamics involved do not have any real bearing on the comparative analysis performed here.

Table 1: *The predicted number of rounds to reach the target sequences*

Selection type	L (length in base pairs)	N (synonymous codons per site)	M(no. codons in element)	R (no. rounds to target)
Artificial	90	-	-	$1.4 * 10^3$
Artificial	180	-	-	$3.1 * 10^3$
Artificial	270	-	-	$5.0 * 10^3$
Artificial	360	-	-	$7.0 * 10^3$
Artificial	450	-	-	$9.0 * 10^3$
Artificial	540	-	-	$1.1 * 10^4$
Natural	90	2	3	$4.6 * 10^6$
Natural	90	2	6	$4.9 * 10^{10}$
Natural	180	2	3	$1.5 * 10^7$
Natural	180	2	6	$1.5 * 10^{11}$
Natural	270	2	3	$3.0 * 10^7$
Natural	270	2	6	$2.9 * 10^{11}$
Natural	360	2	3	$4.9 * 10^7$
Natural	360	2	6	$4.8 * 10^{11}$
Natural	450	2	3	$7.3 * 10^7$
Natural	450	2	6	$7.1 * 10^{11}$
Natural	540	2	3	$1.0 * 10^8$
Natural	540	2	6	$1.0 * 10^{12}$

This comparative analysis shows the calculated number of rounds needed to evolve motif sequences of varying length for both mechanisms of selection, natural and artificial. Increasing the size of L significantly in the case of the latter does not impose a greater cost since there is a logarithmic relationship. However, increasing the size of M in the former incurs an exponential increase in the number of rounds expended to reach the target.

4.2. The Importance of Isolation and Recombination

Despite being evidently more proficient in nature, the pertinent question as to how any self-directed process of selection in Nature would actually work in practice on such a grand scale is less obvious. From human experience with experiments in directed evolution, the physical isolation and screening of mutants is deemed to be a necessary step [32]. These can then become part of a process involving subsequent bouts of investigation and exploration, with

some excluded and shielded from the rest of the group. While this is easy enough to achieve in the controlled environment of a scientific laboratory, it is perhaps not so appropriate for evolution occurring in the wild where there exists no degree of human involvement and supervision. Extensive studies on group isolation and migration in ecology may possibly provide some insight here: the apparent artificial selection of random mutations may well have at least a partly natural dimension that accounts for it. In order to achieve the survival of specific mutations, it is deemed necessary to isolate them since they may be lost either due to drift or because of a differential viability.

Migratory events, whereby a subset of a group becomes either reproductively or geographically isolate, may explain how this could happen [33]. This would, however, require developments corresponding to each individual nucleotide variation. Speculatively, mutants may be separated and then reintroduced into the gene pool through a constant flow of alleles. In this respect, the intragenic recombination [34] of nucleotide polymorphisms might help to speed up the process since any such allelic variation can then become linked up. Eventually, since the final product should be reproductively advantageous, it can be fixed by natural selection acting alone. Therefore, any directing or self-organizing principle within Nature [35] could avail itself of certain shifts in the group and so "select" those variations necessary to the evolution of the novel motif's sequence.

5. Conclusion

Although both models used here relied exclusively on chance mutations, one was organized and directed towards a specific goal whereas the other was not: It was found to be affected by the limitation of a distinct lack of coordination and synchronization that made it extremely inefficient. Dividing the functional complexity of the motif diminished the role of chance in the equation, but did not lower it to an extent where it was no longer problematic. If enough time is allowed, this may become less of an issue, but this also depends on the rate at which *de novo* mutations occur – something believed to be quite infrequent [36]. In a very large population, the number of variations among its members may, however, compensate for this shortcoming. Even so, random mutations may not always be repeatable – at a sufficiently high mutational rate – and could easily become lost from the gene pool. As previously mentioned, the model was flexible in assuming that the composite and holistic property of an independently folding protein domain could be so divisible. While it may be true in some instances, such a degree of functional reducibility is unlikely to be the case for most. What the model and simulation used here actually demonstrates, as a proof of principle rather than as a predictive paradigm, is the marked efficacy of a more directed mechanism for the evolutionary origination of novel motifs in contrast to one that is entirely dependent on the natural outcome of differential reproduction and chance.

References

[1] Kanapin, A.A, Mulder N, Kuznetsov V.A. [2010]. Projection of gene-protein networks to the functional space of the proteome and its application to analysis of organism complexity.BMC Genomics.;11 Suppl 1:S4.

[2] Malek, J.A. [2001]. Abundant protein domains occur in proportion to proteome size. *Genome Biology*. doi:10.1186/gb-2001-2-9-research0039

[3] Ewens, W.J, Wilf, H.S. [2010]. There's plenty of time for evolution. *Proceedings of the National Academy of Sciences*, Vol. 107, No. 52. pp. 22454-22456. doi:10.1073/pnas.1016207107

[4] Salthe, S.N.[2008]. Natural selection in relation to complexity. Artif Life.;14(3):363-74.

[5] Scherer, S.W, Feuk, L, Marquès-Bonet, T, Navarro, A, Okamura, K. [2006]. Frequent appearance of novel protein-coding sequences by frameshift translation. Genomics; 88:690–697

[6] Knowles, D., McLysaght, A [2009]. A. Recent de novo origin of human protein-coding genes. Genome research, 2009; Vol. 19, No. 10, pp. 1752-1759.

[7] Gregory, T. [2009]. Artificial Selection and Domestication: Modern Lessons from Darwin's Enduring Analogy. Evolution: Education and Outreach. Volume 2, Number 1, 5-27, DOI: 10.1007/s12052-008-0114-z

[8] Dougherty, M.J, Arnold, F.H. [2009]. Directed evolution: new parts and optimized function. Curr Opin Biotechnol.;20(4):486-91.

[9] Lynch, M., Abegg, A. [2010]. The rate of establishment of complex adaptations Mol Biol Evol.;27(6):1404-14.

[10] Jäckel C, Kast P, Hilvert D. [2008] .Protein design by directed evolution.*Annu Rev Biophys*. ;37:153-73.

[11] Romero, P.A., Arnold F.H., [2009]. Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol.;10(12):866-76.

[12] Gillespie, J.H.[1991]. *The Causes of Molecular Evolution*. Oxford University Press, New York.

[13] Ohta T .[1993].Interaction of selection and drift in molecular evolution. Jpn. J. Genet. 68, pp. 529-537

[14] Lynch, M. [2010]. Scaling expectations for the time to establishment of complex adaptations.Proc Natl Acad Sci U S A.;107(38):16577-82.

[15] Ohno, S. [1970]. *Evolution by Gene Duplication* (Springer, Heidelberg).

[16] Zhang J.[2003]. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18 (6): 292-298. doi:10.1016/S0169-5347(03)00033-8 To insert individual citation into a bibliography in a word-processor, select your preferred citation style below and drag-and-drop it into the document.

[17] Johnson, A.D. ,[2009]. Single-nucleotide polymorphism bioinformatics: a comprehensive review of resources. Circ Cardiovasc Genet.;2(5):530-6.

[18] Balin SJ, Cascalho M.[2010]. The rate of mutation of a single gene. Nucleic Acids Res.;38(5):1575-82. doi: 10.1093/nar/gkp1119.

[19] Yang, Z., Yoder A., [1999]. Estimation of the *transition*/*transversion* rate bias and species sampling. *J. Mol. Evol*. 48, pp. 274–283

[20] Cusumano, A. (1998). The harmonic series diverges. American Mathematical Monthly 105(7), 608.

[21] Melzer, N., Villmann, C., Becker, K., Harvey, K., Harvey, R.J., Vogel, N., Kluck, C.J., Kneussel M., Becker CM. [2010]. Multifunctional basic motif in the glycine receptor intracellular domain induces subunit-specific sorting. J Biol Chem.;285(6):3730-9.

[22] Rorick M.M, Wagner G.P. [2010]. The origin of conserved protein domains and amino acid repeats via adaptive competition for control over amino acid residues. J Mol Evol.;70(1):29-43.

[23] Ledneva, R.K., Alexeevskii, A.V., Vasil, S.A., Spirin, S.A., Karyagina A.S.[2001]. Structural aspects of interaction of homeodomains withDNA. Mol Biol 35(5):647–659

[24] Thorell S, Gergely P, Banki K, Perl A, Schneider G [2000]. "The three-dimensional structure of human transaldolase". *FEBS Lett.* 475 (3): 205–8.

[25] Gokhale, C.S, Iwasa, Y., Nowak, M.A, Traulsen, A. [2000]. The pace of evolution across fitness valleys J Theor Biol. 2009 7;259(3):613-20.

[26] McPherson, G. [1990]. *Statistics in scientific investigation: its basis, application and interpretation*. Springer-Verlag

[27] Gromiha, M.M, Selvaraj, S (2004). Inter-residue Interactions in Protein Folding and Stability. Prog. Biophys. Mol. Biol. 86, 235-277.

[28] Viguera E, Canceill D, Ehrlich S.D.[2001]. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO* J.;20(10):2587-95.

[29] Kimura, M.,[1962]. On the probability of fixation of mutant genes in a population. Genetics. 47(6): 713–719.

[30] Ohta T. [1972].Population size and rate of evolution. J Mol Evol.;1(3):305-14.

[31] Frankham, R.(1996). Relationship of genetic variation to population size. *Conservation biology* pages 1500-1508; volume 10, no. 6.

[32] Arnold,F.H., Georgiou,G. [2003]. *Directed Enzyme Evolution*: *Screening and Selection Methods*. Humana Press, Clifton, NJ.

[33] Lande, R. [1980]. "Genetic Variation and Phenotypic Evolution During Alliopatric Speciation". *The American Naturalist* 116: 463-479. http://www.jstor.org/pss/2460440.

[34] Stadler, D.R, [1973]. The Mechanism of Intragenic Recombination. Annual Review of Genetics. Vol. 7: 113-127 doi: 10.1146/annurev.ge.07.120173.000553

[35] Batten, D., Salthe, S., Boschetti, F. [2009].Visions of Evolution: self-organization proposes what natural selection disposes. *Biological Theory* 3 (1):17-29.

[36] Lang, G.I., Murray, A.W.[2008]. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae.Genetics. ;178(1):67-82.