

Research article

Open Access

# Molecular clock in neutral protein evolution

Claus O Wilke\*<sup>1,2</sup>

Address: <sup>1</sup>Keck Graduate Institute of Applied Life Sciences, 535 Watson Drive, Claremont, California 91711, USA and <sup>2</sup>Digital Life Laboratory, California Institute of Technology, Mail Code 136-93, Pasadena, California 91125, USA

Email: Claus O Wilke\* - wilke@kgi.edu

\* Corresponding author

Published: 27 August 2004

Received: 09 June 2004

BMC Genetics 2004, 5:25 doi:10.1186/1471-2156-5-25

Accepted: 27 August 2004

This article is available from: <http://www.biomedcentral.com/1471-2156/5/25>

© 2004 Wilke; licensee BioMed Central Ltd.

This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

**Background:** A frequent observation in molecular evolution is that amino-acid substitution rates show an index of dispersion (that is, ratio of variance to mean) substantially larger than one. This observation has been termed the overdispersed molecular clock. On the basis of *in silico* protein-evolution experiments, Bastolla and coworkers recently proposed an explanation for this observation: Proteins drift in neutral space, and can temporarily get trapped in regions of substantially reduced neutrality. In these regions, substitution rates are suppressed, which results in an overall substitution process that is not Poissonian. However, the simulation method of Bastolla et al. is representative only for cases in which the product of mutation rate  $\mu$  and population size  $N_e$  is small. How the substitution process behaves when  $\mu N_e$  is large is not known.

**Results:** Here, I study the behavior of the molecular clock in *in silico* protein evolution as a function of mutation rate and population size. I find that the index of dispersion decays with increasing  $\mu N_e$ , and approaches 1 for large  $\mu N_e$ . This observation can be explained with the selective pressure for mutational robustness, which is effective when  $\mu N_e$  is large. This pressure keeps the population out of low-neutrality traps, and thus steadies the ticking of the molecular clock.

**Conclusions:** The molecular clock in neutral protein evolution can fall into two distinct regimes, a strongly overdispersed one for small  $\mu N_e$ , and a mostly Poissonian one for large  $\mu N_e$ . The former is relevant for the majority of organisms in the plant and animal kingdom, and the latter may be relevant for RNA viruses.

## Background

Kimura has argued that the majority of nucleotide substitutions that accumulate in genes over time are selectively neutral, and go to fixation purely by chance [1]. One major prediction of Kimura's neutral theory is that the substitution process should be a Poisson process, with the mean number of substitutions per unit time equal to the variance. In contrast to this theory, empirical studies often find the variance to be significantly larger than the mean [2-8]. This observation has been termed the "overdis-

persed molecular clock". (For an excellent review of both empirical evidence and mathematical theories, see Ref. [9].) It is possible to reconcile Kimura's theory with the overdispersed molecular clock via Takahata's fluctuating neutral space model [10-12]. If the neutral mutation rate fluctuates slowly, then the substitution process ceases to be Poissonian, and becomes indeed overdispersed. However, the problem with the fluctuating neutral space model is that it does not offer any argument for why the

neutral mutation rate should fluctuate, and thus ultimately fails to explain the observed substitution patterns.

An explanation for fluctuations in neutral mutation rate was recently proposed by Bastolla et al. [13-16]. Different proteins with identical structure naturally vary in their neutrality, that is, in the fraction of single-point mutants that are viable. Therefore, as a gene slowly drifts through sequence space, the neutral mutation rate will fluctuate in correlation to the changing neutrality, and this fluctuation alone could be sufficient to explain the overdispersed molecular clock. Bastolla et al. studied the substitution process in a variety of models of neutral protein evolution *in silico*, and found significant overdispersion in all cases they considered.

However, the simulations that Bastolla et al. carried out were limited to cases in which the product of mutation rate  $\mu$  and population size  $N_e$  is small (because Bastolla et al. used only a single sequence as the representative of the whole population, an approach that is justified for  $\mu N_e \lesssim 1$ ). Since population size and mutation rate can be substantial in some species (most notably in RNA viruses), it is justified to ask how general this result is for arbitrary values of  $\mu N_e$ . It is known that large populations and populations evolving under high mutation pressure experience a strong selective pressure to avoid regions of low neutrality, an effect that has been termed "evolution of mutational robustness" [17-20]. In equilibrium, such populations settle in areas of sequence space that have above-average neutrality. As a result, regions of low neutrality are not represented in the population, and the distribution of neutralities in the population is much narrower than the total distribution of neutralities in sequence space. Therefore, we should expect that the neutral mutation rate does not fluctuate strongly under these conditions, and that the molecular clock will not be significantly overdispersed.

For the present paper, I have studied the behavior of the substitution process under neutral protein evolution as a function of mutation rate  $\mu$  and population size  $N_e$ . I have found that the accumulation of non-synonymous mutations is substantially overdispersed for small  $\mu N_e$ , in agreement with the results of Bastolla et al., but approaches a Poisson process when  $\mu N_e \gg 1$ . The accumulation of synonymous substitutions is always Poissonian, regardless of the value of  $\mu N_e$ .

## Results

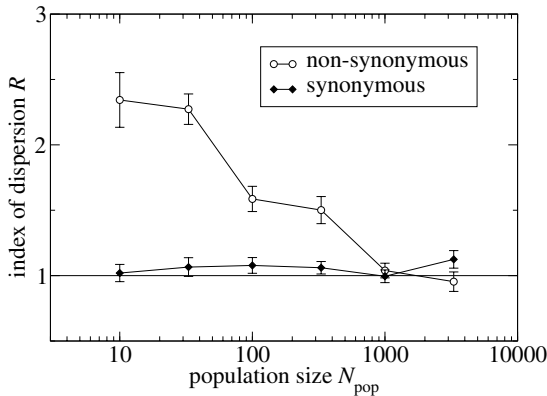
I carried out simulations with DNA sequences of length  $L = 75$ . I determined the fitness of a DNA sequence by translating it into the corresponding amino-acid sequence, and determining its native fold within the framework of a lattice-protein model. (A sequence would have fitness 1 if it

folded into a pre-determined target structure, and fitness 0 otherwise.) I used a simple model of maximally compact proteins on a  $5 \times 5$  lattice. This protein-folding model is much simpler than the ones used by Bastolla et al. [13,14], but has been shown to produce realistic distributions of folding free energies and neutralities [21-23]. The advantage of the simpler model is that entire populations of evolving sequences can be simulated, instead of just individual sequences.

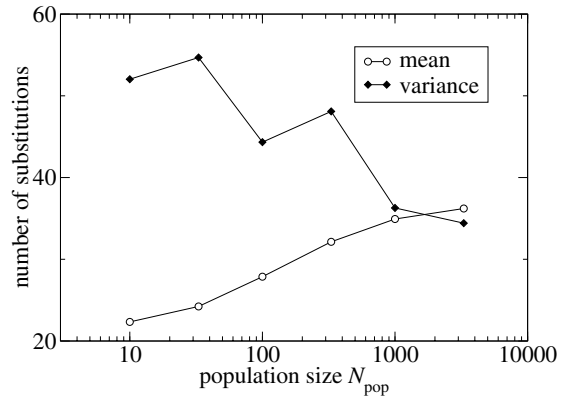
First, I have found that my model produces overdispersion (that is, an index of dispersion  $R$  substantially above 1) for non-synonymous substitutions, but not for synonymous substitutions. The finding for synonymous mutations is not surprising, because changes in the protein's neutrality do not affect the probability with which a synonymous mutation is neutral (which is always one). Neutral evolution could produce overdispersion in the synonymous substitutions only if the number of synonymous sites in the sequence were undergoing significant fluctuations. While these fluctuations do occur, they are apparently not large enough to affect the index of dispersion.

Second, I have found that for non-synonymous substitutions,  $R$  decays quickly with increasing population size  $N_e$  at fixed  $\mu$  (Fig. 1). Since one reason for a decaying index of dispersion could be a reduced number of accumulated mutations, I have studied how the mean number of accumulated mutations behaves as a function of population size. Instead of staying constant or decreasing, the mean increases with increasing  $N_e$ , while the variance decreases (Fig. 2). This result shows that the reduction in  $R$  is not caused by a mere reduction in the accumulated mutations, and that the substitution process does indeed shift from overdispersed to Poissonian as the population size increases.

For non-synonymous substitutions,  $R$  decays with  $N_e$  because of evolution of mutational robustness. However, mutational robustness is caused by large  $\mu N_e$ , rather than large  $N_e$  alone, and the parameter region in which mutational robustness becomes relevant is  $\mu N_e \gtrsim 10$  [17]. Therefore, it is more instructive to plot  $R$  as a function of  $\mu N_e$ . The only problem with a naive plot of that sort is that  $R$  increases as a function of  $\mu\tau$ , where  $\tau$  is the length of the time window during which mutations accumulate [9]. Thus, in Fig. 3, I show  $R$  for constant  $\mu\tau$  as a function of  $\mu N_e$ . Note that in this figure, instead of the sequence-wide mutation rate  $\mu$ , I use the non-synonymous mutation rate  $\mu_n = 0.76 \mu$ , which is corrected for the fact that only approximately 76% of mutations hit non-synonymous sites. (76% is the expected fraction of non-synonymous sites in a random DNA sequence.) Figure 3 shows that the transition from an overdispersed to a Poissonian



**Figure 1**  
Index of dispersion as a function of population size  $N_e$  for synonymous and non-synonymous substitutions ( $\tau = 1000$ ,  $\mu = 0.075$ ).



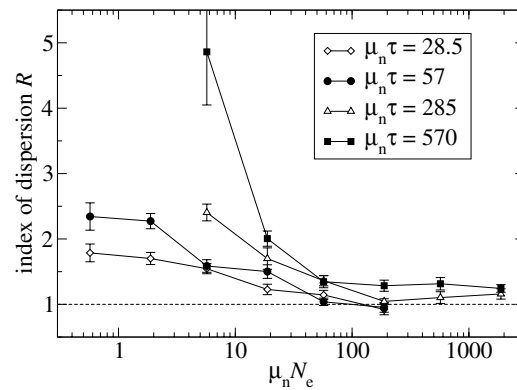
**Figure 2**  
Mean,  $\langle \tilde{N}_d \rangle$  and variance,  $\text{Var}(\tilde{N}_d)$  of lineage-adjusted number of non-synonymous substitutions as a function of population size  $N_e$  ( $\tau = 1000$ ,  $\mu = 0.075$ ). Quantities were calculated from all 500 replicates at each population size.

substitution process occurs for  $\mu N_e$  between approximately 10 and 100, in agreement with Ref. [17], and that the transition region seems to be largely independent of the value of  $\mu\tau$ .

Figure 3 also shows that  $R$  increases with  $\mu\tau$ . This dependency becomes clearer in Fig. 4, where I display  $R$  as a function of  $\mu\tau$  for fixed  $\mu N_e$ . The figure shows substantial increase in  $R$  with increasing  $\mu\tau$  for small to moderate  $\mu N_e$ . However, even for  $\mu N_e$  well above 50, there is still a slight increase in  $R$  with  $\mu\tau$ . Therefore, my results do not settle the question of whether the substitution process becomes truly Poissonian for sufficiently large  $\mu N_e$ , or whether it just approaches a Poisson process but always remains slightly overdispersed. To settle this question, one would have to carry out simulations with much larger  $\tau$  and  $N_e$ . Unfortunately, the protein folding model I use is still too computationally intensive to permit such simulations with current computational resources.

**Discussion**

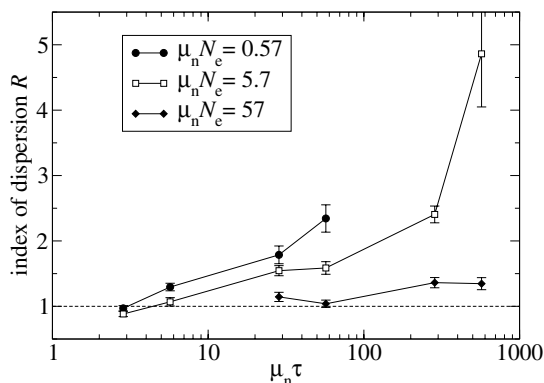
My results show that the size of the product  $\mu N_e$  has a substantial effect on the index of dispersion under neutral evolution. The substitution process is strongly overdispersed for small  $\mu N_e$ , but approaches a Poisson process as  $\mu N_e$  grows large. Therefore, the next question is which of the two regimes has more biological relevance. As discussed by Cutler [9], the biggest problem in explaining the overdispersed molecular clock is not to come up with mechanisms that produce overdispersion, but to find a



**Figure 3**  
Index of dispersion for non-synonymous mutations as a function of the product of non-synonymous mutation rate  $\mu_n$  ( $= 0.76 \mu$ ) and population size  $N_e$ .

general mechanism that does not depend on special conditions or finely-tuned parameters.

To assess the likelihood that fluctuations in neutrality contribute to the overdispersed molecular clock, we have



**Figure 4**  
Index of dispersion for non-synonymous mutations as a function of the product of non-synonymous mutation rate  $\mu_n$  ( $= 0.76 \mu$ ) and divergence time  $\tau$ .

to know the mutation rate and population size for the species of interest. It is notoriously difficult to obtain accurate data for these parameters, and only a few species have been studied in depth. One of the best data sets available is probably the one for *Drosophila*. Keightley and Eyre-Walker estimated the per-nucleotide substitution rate in *Drosophila* to be  $u = 2.2 \times 10^{-9}$  [24]. If we assume that the average gene in *Drosophila* is 1770 bp long [24], and that 76% of the nucleotides are non-synonymous (this number stems from averaging the number of non-synonymous sites over all codons with equal weight), then the average number of non-synonymous sites per gene is 1345 bp. Thus, the average rate of non-synonymous mutations per gene is  $\mu_n = 3.0 \times 10^{-6}$ . With an effective population size of approximately  $3 \times 10^5$  [25], we get a product of population size and per-gene-non-synonymous mutation rate of approximately 1. Since selection for mutational robustness starts to take effect when this product is substantially larger than 1, *Drosophila* lies well within the parameter region in which we expect overdispersion to be caused by neutral evolution. For other higher organisms, in particular mammals, which tend to have comparatively small population sizes, we can expect that the product  $\mu_n N_e$  falls into the same parameter region. On the other hand, for microorganisms, which can have very large population sizes, mutational robustness may play a role in their evolution. In particular, RNA viruses have genomic mutation rates on the order of one [26,27] and their genomes consist typically of only a handful of genes. Because RNA viruses undergo severe bottlenecks on a regular basis, their effective population size  $N_e$  is much smaller than the number of virus particles

in infected individuals (which can exceed  $10^{12}$ ), and is more closely related to the number of infected individuals. For HIV-1,  $N_e$  has been estimated to be approximately  $10^2$  for subtype A, and  $10^5$  for subtype B [28].

The preceding paragraph shows that neutral evolution of proteins is probably one source of overdispersed non-synonymous substitutions in *Drosophila* and other organisms. However, overdispersion has been observed in synonymous substitutions as well. For example, Zeng et al. [29] found an index of dispersion  $R$  significantly above one for synonymous, but not for non-synonymous substitutions in *Drosophila*. For mammals, some studies found  $R$  significantly above one for both synonymous and non-synonymous substitutions [8], while others found only the non-synonymous substitution process to be overdispersed [30]. Therefore, it is likely that other processes than neutral protein evolution also contribute to overdispersion. Such processes can be selection for optimal codon usage in the case of synonymous mutations, and positive selection on the amino acid level in the case of non-synonymous mutations.

I have demonstrated that large  $\mu N_e$  results in a substitution process with little overdispersion. However, I have not yet given an explanation for how overdispersion is reduced in populations with large  $\mu N_e$ . There are two elements: First, selection for mutational robustness reduces the fraction of sequences with low neutrality, and increases the fraction of sequences with high neutrality, thus making the population more homogeneous and reducing the overall range of neutralities [17-20]. Second, a sequence with low neutrality will experience a real selective disadvantage in comparison to a sequence with high neutrality for large  $\mu N_e$ , and will therefore have a reduced probability to end up on the line of descent. While this selective disadvantage is often small, it can nevertheless determine the evolutionary fate of a sequence in a large population. The larger the population, the more sensitive it becomes to small fitness differences, so that in a very large population a sequence with only a moderate reduction in neutrality will have a small probability to end up on the line of descent. (The fact that the mean substitution rate increases with the population size, as seen in Fig. 2, is also consistent with this reasoning. The larger the population size, the more high-neutrality sequences end up on the line of descent, which is reflected in the increase in the mean substitution rate.)

Throughout this paper, I have considered only neutral or lethal mutations. It is a reasonable question to ask if and how deleterious mutations would change my results. The answer is that they probably have only a minor impact, and the less so the larger  $N_e$ , unless they are very slightly deleterious. In order to affect the molecular clock, the

deleterious mutations must end up on the line of descent, that is, they must go to fixation. The probability of fixation  $p_{\text{fix}}$  of deleterious mutants drops exponentially with the population size,  $p_{\text{fix}} = [1 - \exp(2s)] / [1 - \exp(2sN_e)]$ , where  $s$  is the selective disadvantage of the deleterious mutation [31]. Therefore, for reasonable population sizes, only very slightly deleterious mutations can go to fixation and thus affect the molecular clock. This reasoning is independent of the size of  $\mu N_e$ , as long as  $N_e$  is large in comparison to  $s$ .

**Conclusions**

The present study supports the following conclusions:

- Neutral drift of proteins can lead to an overdispersed substitution process for non-synonymous mutations, but not for synonymous mutations.
- The amount of overdispersion in the non-synonymous substitution process depends strongly on the product of mutation rate and population size. As this product increases, the substitution process becomes more and more Poissonian. The transition region starts at  $\mu N_e \approx 10$ , and extends to values well above 100.
- It is not clear whether there are any species that have a sufficiently large population size and mutation rate to prevent overdispersion through neutral drift. In *Drosophila*, the product of mutation rate and population size is close to one, which is well below the parameter region in which the substitution process turns Poissonian.

**Methods**

**Lattice protein model**

I implemented a version of the  $5 \times 5$  lattice protein model put forward by Goldstein and coworkers [21-23,32]. In this model, proteins are sequences of  $n = 25$  residues that fold into a maximally compact structure on a two-dimensional grid of  $5 \times 5$  lattice points. There are 1081 distinct possible conformations in this model, and the partition function can be evaluated exactly by summing over the contact energies of all distinct conformations.

The contact energy of a conformation  $i$  is

$$E_i = \sum_{j < k} \gamma(\mathcal{A}_j, \mathcal{A}_k) \Delta_{jk}^i, \tag{1}$$

where  $\gamma(\mathcal{A}_j, \mathcal{A}_k)$  is the contact energy between amino acids  $\mathcal{A}_j$  at location  $j$  and  $\mathcal{A}_k$  at location  $k$  in the sequence, and  $\Delta_{jk}^i$  is 1 if the two amino acids are in contact in conformation  $k$ , and 0 otherwise. The partition function is

$$Z = \sum_i \exp(-E_i / kT), \tag{2}$$

where the sum runs over all 1081 conformations. A sequence folds into conformation  $f$  if the contact energy for that conformation is lower than the contact energies of all other formations,  $E_f < E_i$  for all  $i \neq f$ , and if the free energy of folding, which is defined as

$$\Delta G_{\text{folding}} = E_f + kT \ln [ Z - \exp(-E_f / kT) ], \tag{3}$$

is smaller than some cutoff  $\Delta G_{\text{cut}}$ . Throughout this study, I used  $kT = 0.6$  and  $\Delta G_{\text{cut}} = 0$ . The contact energies  $\gamma(\mathcal{A}_j, \mathcal{A}_k)$  were taken from Table VI in Ref. [33].

**Sequence evolution**

I simulated the evolution of populations of DNA sequences in discrete, non-overlapping generations. Population size is denoted by  $N_e$ . The fitness of a sequence was 1 if the DNA sequence translated into a peptide sequence that could fold into a chosen target structure, and had a free energy of folding smaller than  $G_{\text{cut}}$ . Otherwise, the fitness of the sequence was 0. All sequences had length  $L = 75$ . In each successive generation, sequences with fitness 1 were randomly chosen to reproduce, until the new generation had  $N_e$  members. At reproduction, the sequences were mutated, with an average of  $\mu$  base pair substitutions per sequence. I let each population evolve for several thousand generations, and kept track of the full genealogic information of all sequences in the population. In order to measure the molecular clock of fixed mutations only, I studied the pattern of base substitutions in a window of  $\tau$  generations along the line of descent backwards in time, starting from the most recent common ancestor of the final population.

I varied the parameters  $N_e$  (10, 33, 100, 330, 1000, 3300),  $\mu$  (0.0075, 0.075, 0.75), and  $\tau$  (500, 1000). For each set of parameters, I carried out 500 replicates (each with a different, randomly chosen target structure), to obtain a distribution for the number of synonymous and non-synonymous substitutions  $S_d$  and  $N_d$ . Since there was some variation in the number of synonymous and non-synonymous sites across different target structures (on the order of approximately  $\pm 5\%$  variation from the mean), I then applied a correction factor to  $S_d$  and  $N_d$  to bring them into comparable units: I calculated the corrected number of synonymous substitutions  $\tilde{S}_d$  as  $\tilde{S}_d = S_d \times \langle S \rangle / S$ . Here,  $S$  is the mean number of synonymous sites for the given replicate, and  $\langle S \rangle$  is the average of  $S$  over all 500 replicates. Likewise, I calculated  $\tilde{N}_d = N_d \times \langle N \rangle / N$  (Indices of dis-

persion calculated without this correction factor are slightly larger than the ones reported here, because the variation in  $S$  and  $N$  creates additional variance in  $S_d$  and  $N_d$ ). Similar correction factors have been used in sequence analysis [7], and are generally referred to as lineage adjustments. They control for differences among lineages that are primarily related to the expected number of substitutions in a lineage, and thus should not enter the index of dispersion.

To obtain an estimate for mean and standard error of the index of dispersion, I subdivided the 500 results into 10 blocks of 50 each, and calculated mean and variance of the number of substitutions for each block. The ratio of variance to mean for a given set of substitutions (synonymous or non-synonymous) in a block is the index of dispersion for this data set. I then calculated mean and standard error for the index of dispersion from the individual results of the 10 blocks.

The total CPU time needed to carry out all simulations was several months on a small cluster of Pentium II 500 MHz machines.

#### Calculation of synonymous and non-synonymous substitutions and sites

I calculated the number of synonymous and non-synonymous sites  $S$  and  $N$  and the number of synonymous and non-synonymous substitutions  $S_d$  and  $N_d$  according to the method proposed by Nei and Gojobori [34]. In short, under this method the number of synonymous sites  $s_i$  of a codon  $i$  is the fraction of possible substitutions to that codon that leave the residue unchanged. The number of non-synonymous sites  $n_i$  for the same codon is  $n_i = 3 - s_i$ . For the complete sequence,  $S$  and  $N$  are calculated as  $S = \sum_i s_i$  and  $N = \sum_i n_i$  where  $i$  runs over all codons in the sequence. The number of synonymous or non-synonymous substitutions  $s_{d,i}$  or  $n_{d,i}$  between two codons is the average number of such substitutions, where the average is taken over all paths that lead from one codon to the other. The total number of synonymous or non-synonymous substitutions between two sequences is the sum over all individual contributions,  $S_d = \sum_i s_{d,i}$  and  $N_d = \sum_i n_{d,i}$  (again,  $i$  runs over all codons in the sequence).

To calculate the number of synonymous or non-synonymous substitutions along the line of descent, I simply summed up all synonymous or non-synonymous substitutions that occurred from generation to generation. Because the full evolutionary history was known, a correction for multiple mutations such as the Jukes-Cantor cor-

rection [35] was not necessary. I also averaged the number of synonymous and non-synonymous sites over all sequences along the line of descent, to get the mean number of synonymous and non-synonymous sites for the given evolutionary trajectory.

#### Authors' contributions

COW carried out all aspects of this study.

#### Acknowledgements

This work was supported in part by the NSF under Contract No. DEB-9981397. I thank Ugo Bastolla for helpful comments on an earlier version of this paper.

#### References

1. Kimura M: *The Neutral Theory of Molecular Evolution* Cambridge: Cambridge University Press; 1983.
2. Ohta T, Kimura M: **On the constancy of the evolutionary rate of cistrons.** *J Mol Evol* 1971, **1**:18-25.
3. Langley CH, Fitch WM: **An estimation of the constancy of the rate of molecular evolution.** *J Mol Evol* 1974, **3**:161-177.
4. Gillespie JH: **The molecular clock may be an episodic clock.** *Proc Natl Acad Sci USA* 1984, **81**:8009-8013.
5. Gillespie JH: **Natural selection and the molecular clock.** *Mol Biol Evol* 1986, **3**:138-155.
6. Gillespie JH: **Variability of evolutionary rates of DNA.** *Genetics* 1986, **113**:1077-1091.
7. Gillespie JH: **Lineage effects and the index of dispersion of molecular evolution.** *Mol Biol Evol* 1989, **6**:636-647.
8. Ohta T: **Synonymous and non-synonymous substitutions in mammalian genes and the nearly neutral theory.** *J Mol Evol* 1995, **40**:56-63.
9. Cutler DJ: **Understanding the overdispersed molecular clock.** *Genetics* 2000, **154**:1403-1417.
10. Takahata N: **On the overdispersed molecular clock.** *Genetics* 1987, **116**:169-179.
11. Takahata N: **Statistical models of the overdispersed molecular clock.** *Theor Popul Biol* 1991, **39**:329-344.
12. Cutler DJ: **The index of dispersion of molecular evolution: slow fluctuations.** *Theor Popul Biol* 2000, **57**:177-186.
13. Bastolla U, Roman HE, Vendruscolo M: **Neutral evolution of model proteins: Diffusion in sequence space and overdispersion.** *J Theor Biol* 1999, **200**:49-64.
14. Bastolla U, Porto M, Roman HE, Vendruscolo M: **Lack of self-averaging in neutral evolution of proteins.** *Phys Rev Lett* 2002, **89**:20801.
15. Bastolla U, Porto M, Roman HE, Vendruscolo M: **Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution.** *J Mol Evol* 2003, **56**:243-254.
16. Bastolla U, Porto M, Roman HE, Vendruscolo M: **Statistical properties of neutral evolution.** *J Mol Evol* 2003, **57**:S103-S119.
17. van Nimwegen E, Crutchfield JP, Huynen M: **Neutral evolution of mutational robustness.** *Proc Natl Acad Sci USA* 1999, **96**:9716-9720.
18. Bornberg-Bauer E, Chan HS: **Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space.** *Proc Natl Acad Sci USA* 1999, **96**:10689-10694.
19. Wilke CO: **Adaptive evolution on neutral networks.** *Bull Math Biol* 2001, **63**:715-730.
20. Wilke CO, Adami C: **Evolution of mutational robustness.** *Mutat Res* 2003, **522**:3-11.
21. Taverna DM, Goldstein RA: **The distribution of structures in evolving protein populations.** *Biopolymers* 2000, **53**:1-8.
22. Taverna DM, Goldstein RA: **Why are proteins so robust to site mutations?** *J Mol Biol* 2002, **315**:479-484.
23. Taverna DM, Goldstein RA: **Why are proteins marginally stable?** *Proteins* 2002, **46**:105-109.
24. Keightley PD, Eyre-Walker A: **Deleterious mutations and the evolution of sex.** *Science* 2000, **290**:331-333.

25. Li YJ, Satta Y, Takahata N: **Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method.** *Genes Genet Syst* 1999, **74**:117-127.
26. Drake JW: **Rates of spontaneous mutation among RNA viruses.** *Proc Natl Acad Sci USA* 1993, **90**:4171-4175.
27. Drake JW, Holland JJ: **Mutation rates among RNA viruses.** *Proc Natl Acad Sci USA* 1999, **96**:13910-13913.
28. Grassly NC, Harvey PH, Holmes EC: **Population dynamics of HIV-1 inferred from gene sequences.** *Genetics* 1999, **151**:427-438.
29. Zeng LW, Comeron JM, Chen B, Kreitman M: **The molecular clock revisited: the rate of synonymous vs. replacement change in *Drosophila*.** *Genetica* 1998, **102/103**:369-382.
30. Smith NG, Eyre-Walker A: **Partitioning the variation in mammalian substitution rates.** *Mol Biol Evol* 2003, **20**:10-17.
31. Kimura M: **On the probability of fixation of mutant genes in a population.** *Genetics* 1962, **47**:713-719.
32. Buchler NEG, Goldstein RA: **Effect of alphabet size and foldability requirements on protein structure designability.** *Proteins* 1999, **34**:113-124.
33. Miyazawa S, Jernigan RL: **Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation.** *Macromolecules* 1985, **18**:534-552.
34. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3**:418-426.
35. Jukes TH, Cantor CR: **Evolution of protein molecules.** In *Mammalian protein metabolism III* Edited by: Munro HN. New York: Academic Press; 1969:21-132.

Publish with **BioMed Central** and every scientist can read your work free of charge

*"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."*

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

