Platt *et al. BMC Evolutionary Biology* (2018) 18:17
https://doi.org/10.1186/s12862-017-1085-x

BMC Evolutionary Biology

REVIEW Open Access

Protein evolution depends on multiple distinct population size parameters

Alexander Platt†, Claudia C. Weber† and David A. Liberles*

Abstract

That population size affects the fate of new mutations arising in genomes, modulating both how frequently they arise and how efficiently natural selection is able to filter them, is well established. It is therefore clear that these distinct roles for population size that characterize different processes should affect the evolution of proteins and need to be carefully defined. Empirical evidence is consistent with a role for demography in influencing protein evolution, supporting the idea that functional constraints alone do not determine the composition of coding sequences. Given that the relationship between population size, mutant fitness and fixation probability has been well characterized, estimating fitness from observed substitutions is well within reach with well-formulated models. Molecular evolution research has, therefore, increasingly begun to leverage concepts from population genetics to quantify the selective effects associated with different classes of mutation. However, in order for this type of analysis to provide meaningful information about the intra- and inter-specific evolution of coding sequences, a clear definition of concepts of population size, what they influence, and how they are best parameterized is essential. Here, we present an overview of the many distinct concepts that "population size" and "effective population size" may refer to, what they represent for studying proteins, and how this knowledge can be harnessed to produce better specified models of protein evolution.

Keywords: Protein evolution, Effective population size, Mutation-selection models

Background

Understanding how proteins evolve under the influence of natural selection is a central goal of evolutionary biology, as it provides insight into functional constraints and the diversification of genomes. Although it may be tempting to study protein evolution from a predominately biophysical perspective, considering how mutational processes generate variation and population level processes modulate selection is necessary to fully explain extant coding sequences. Counter to the notion that amino acid sequences are determined by functional requirements alone, some of the observed variation is a consequence of the limits of natural selection in finite populations.

The functional synthesis of protein evolution and population genetics has shown that the size of a population (N) modulates amino acid sequence divergence as well as the rates and patterns of adaptation [1, 2]. In simple models of population genetics, the standing pool of genetic diversity, the probability of fixation of a new neutral mutation, and the fixation probability of a selected mutation all relate to N and take on values of $2N\mu$, $1/N$ and $2Ns$, respectively. Here, s represents the relative selection coefficient of the mutant allele (that is, how fit the mutant is compared to the wild type).

In accord with this view, comparative genomics has linked changes in effective population size to differences in observed properties of proteins, such as stability and other features subject to selection, including binding specificity and avoiding misinteraction [2–4]. The rate of accumulation of neutral changes is not affected by N, as the probability of introduction of a mutation is inversely proportional to the neutral probability of fixation. However, all changes subject to selective constraint or adaptive pressure are influenced by the population size [5].

Beyond the simplest models, however, the number of individuals in a population is rarely the correct scalar for all of these parameters of interest. It is standard practice,

*Correspondence: daliberles@temple.edu
†Equal contributors
Department of Biology and Center for Computational Genetics and Genomics, Temple University, 19121 Philadelphia, USA

therefore, to employ a variation of it, the *effective* population size (N_e) to account for deviations introduced by any and all of a host of complications such as inbreeding, unequal sex ratios, linked selected sites, population substructure, life-history patterns, or high-variance reproductive strategies [6].

Given its importance in influencing sequence variation, what precisely do we mean when we refer to N_e in the context of protein evolution? There are multiple definitions of effective population size in use. For instance, in population genetics, N_e might be treated as a convenience parameter reflecting the extent of genetic drift (stochastic changes in allele frequency) inferred from a sequence, or as a constant summarizing an unruly history of fluctuating demography or complicated social structure. However, to understand how protein structure and function drive amino acid substitution, we require models that disentangle the factors contributing to neutral and adaptive sequence divergence and describe the underlying biological processes accurately [7, 8]. We refer to these as mechanistic models in this work (though the term is used differently elsewhere [9, 10]). The goal of the modeling effort becomes to connect real physical observations and processes with parameters in models. This can include experimentally determined mutation rates, selection coefficients, and recombination rates, among others.

Here we discuss how to define N_e and how this parameter can be augmented to capture information about mechanistic processes that are distinct from natural selection, a prerequisite to realistically characterizing the evolution of proteins in some scenarios.

Origins and historical applications

The concept of effective population size has its origins in the Wright-Fisher model, which describes the change in allele frequency through time in a single randomly mating population of constant size. This model has a single parameter: the size of the population. *All* predictions from this model could be interpreted as functions of the population size, giving a direct map between any predictable population measurement and a population size that would produce it. Where the model assumptions of Wright-Fisher were applicable, it was reasonable to refer to this parameter in its original sense as the unchanging, panmictic, neutral, asexual, non-overlapping, population size N [11, 12].

This model proved to be both tractable and powerful for deriving many important properties of evolving gene pools such as quantifying genetic drift, the probabilities of allele loss and fixation, and allele sojourn times within specified frequency ranges. Real populations, and even most interesting population models, however, violate many of the restrictive assumptions made in the Wright-Fisher model. Deviations include population size change, population structure, organisms divided into sexes, assortative mating, selection acting both directly on individual loci and indirectly on linked neutral loci (which we may not wish to include as part of N_e), and non-overlapping generations. Still, for any observation predictable under the Wright-Fisher model, it is possible to ask what population size in the Wright-Fisher model maps to the observed value in the more complicated model. This parameter was given the name "effective population size", while the actual size of the population is designated as the census population size or N_c.

Defined this way there is no requirement that the effective population size need be the same for any two different observations from the same population. In a Wright-Fisher population there is only a single value of N that is used to make all predictions about the behavior of the model. In a non Wright-Fisher population, the N_e that corresponds to the probability of fixation of a newly arising allele might not be the same value as the N_e that describes, for example, the rate of change in frequency that said allele exhibits or the probability that two randomly sampled individuals share a recent common ancestor. In each case, one must independently compare the observed value to a corresponding Wright-Fisher model, and it is always necessary when talking about the effective size of a population to qualify it with what observation is being described. Careful usage, then, was to explicitly label an effective population size as to what it was describing [6, 13]. Common variants included the inbreeding effective population size (describing the probability that two individuals shared a common ancestor in the previous generation) [6, 13, 14], the variance effective population size (describing the variance in reproductive success among individuals) [6, 13, 14], and the eigenvalue effective population size (describing the leading non-zero eigenvalue of the allele frequency transition matrix) [6, 13]. These concepts (like inbreeding structure or variable reproductive success in the population over a lineage of a phylogenetic tree) are directly related to individual protein-specific selective pressures and probabilities of amino acid fixation through the effects of the broader population acting on all proteins.

The more recent introduction of another useful effective population size, the coalescent effective population size, has gone a great way in elucidating the underlying mechanism of the frequent similarities of effective population sizes as well as when more subtle distinctions are required [15–18]. Many parameters of interest to a population geneticist can be described as properties of the genealogy of a random sample of individuals. Furthermore, many different population models converge to a common underlying form of this genealogy (often referred to as Kingman's coalescent), each differing only in a single scaling parameter. This includes models with deviations from Wright-Fisher such as unequal sex ratios, structured life-

stages, or uniformly elevated or depressed reproductive variance. For population models that converge to Kingman's coalescent, this scaling factor serves the purpose of the effective population size for *all* of the population genetic statistics determined by sample genealogy (largely by definition – under the conditions in which a population converges to a Kingman coalescent, the scaled Kingman coalescent fully describes the properties of the genealogy of the sample). It is compelling both as an explanation of why a great many definitions of effective population size often produce very similar numbers, as they are derived from a common rescaling of the underlying genealogy, and as an illustration of when there is *no* single parameter that can meaningfully compare the general behavior of a particular population to one evolving under Wright-Fisher rules. As illustrated in Fig. 1, subdivided populations offer a particularly clear example of this type of violation. For any large, random sample of individuals, the genealogy will have an excess of short tips (individuals in small sub-populations sharing recent common ancestry with each other) and longer internal branches (individuals in different sub-populations sharing only distant ancestors) than described by the Kingman coalescent. As the distortions in different parts of the tree are in opposite directions there can be no single scalar that adjusts this genealogy to match one produced by a Wright-Fisher population with the ratio of inter-coalescent times shifting along the timescale of the sample even if nothing changes about the size, structure, or behavior of the population [15, 19, 20].

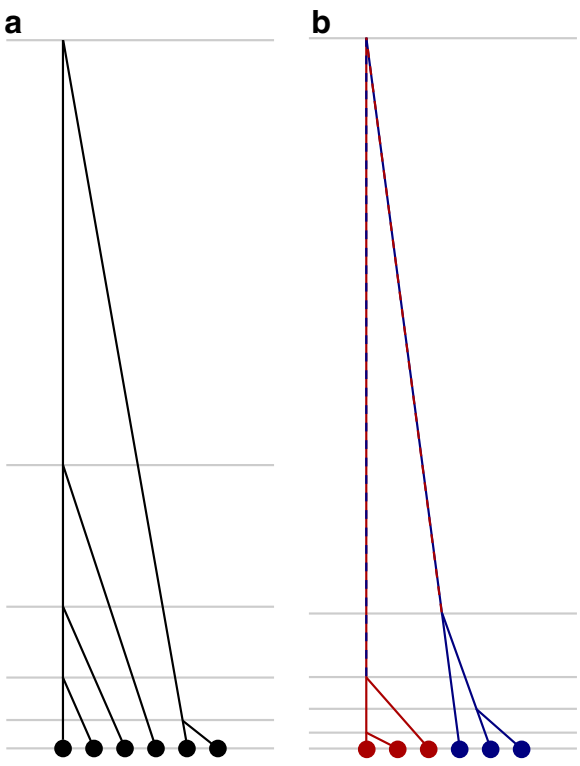

Fig. 1 Example coalescent trees. Representative coalescent trees from (**a**) an unstructured population and (**b**) a population with two sub-populations with limited migration between them. Both trees have been scaled to the same total depth, but the distribution of expected coalescent times (grey lines) differ dramatically between the trees

Multiple important population size parameters

Removed from the cozy confines of the Wright-Fisher model, the simple multiplicative relationship between population size and the context-dependent conditional probability of acquiring a beneficial mutation, the genetic diversity, and the probability of fixation of a favored allele becomes more complex and requires different treatments of population size. If any individual born into a population has some chance μ of carrying a new mutation, the population will acquire new mutations at the rate of $N_c\mu$ ($2N_c\mu$ for diploids). The population size scalar here does not depend on any properties of the relations among individuals within the population or how they came to be that way and is simply a reflection of the mutational target size. The loss of variation through random genetic drift, however, *does* depend on the nature of the population. A new neutral mutation enters a diploid population at frequency $\frac{1}{2N_c}$. The expected time until fixation or loss of this mutation is $-8N_eN_c\left(1-\frac{1}{2N_c}\right)ln\left(1-\frac{1}{2N_c}\right)$ [13, 21, 22], a function of both N_e *and* N_c. For mutations with an additive selection coefficient of s, the probability of fixation is $\left(1-e^{-2N_es/N_c}\right)/\left(1-e^{-4N_es}\right)$ [5],or approximately $2sN_e/N_c$ [5, 23], and is driven both by the intensity of selection on the particular variant and the ratio of the variance in reproductive output (as reflected in N_e) and the mean reproductive output (as reflected in N_c). In populations where N_e cannot be properly defined, such as those with persistent subdivision, these probabilities become more dependent on additional population size measures [24, 25]. Mutations arise within each sub-population with probability proportional to the census size of the sub-population and then drift to loss or fixation according to a complex meta-population dynamic that may include considerable time spent fixed within some sub-populations and absent in others. Individual effective population sizes come into play across sub-populations with a timescale-variant global effective population size accounting for a complex migration process [15, 19, 20, 26]. While there will always be *some* probability distribution of time to fixation or loss of an arbitrary new mutant it is unlikely in these cases to be able to assign values N_c and N_e such that the distribution matches anything produced by a population that converges to the Kingman coalescent at a biologically relevant timescale. These concepts need to be considered in the context of models for understanding selection in proteins.

In cases where N_e *is* well-defined, N_c cancels out of some key properties for molecular evolution, but often only after making other, further assumptions. For instance, in a randomly mating population of constant size permitting only mutations of small effect, while the probabilities of fixation or loss of a new mutation depend on both N_c and N_e, the total *rate* of substitution will depend only on N_e. Comparing two such populations of equivalent effective population size, one with $N_c = N_e$ and another with $N_c = 2N_e$, the one with the doubled census size will experience twice as many mutations but is half as likely to fix any one of them. This reassuring phenomenon breaks down, however, when the fitness effects of new mutations are drawn from a distribution with appreciable mass in its tail. A population with a large census size is more likely to encounter a mutation of sufficiently large effect that the diffusion approximations from which the N_c terms cancel [27] are no longer applicable.

Practical principles of effective population size for protein evolution studies

Effective population size impacts the distribution of fitness effects

In addition to influencing the probability with which a mutation with a given selection coefficient fixes, it has been suggested that N_e affects the distribution of selective effects itself [3, 28]. An important source of constraint on proteins arises from the requirement to fold stably into the correct structure [8, 29, 30]. The free energy of folding (ΔG) or stability of a given protein under stabilizing selection is determined by mutation-selection-drift balance. This describes the equilibrium at which the rate of deleterious alleles being removed by selection equals the rate at which they arise due to mutation [31, 32]. As a result, proteins in nature are only marginally stable [33]. Because N_e affects the efficacy of selection against deleterious, destabilizing variants, one might expect proteins to be more stable for organisms with large effective population sizes [4]. These differences in stability, then, are expected to impact the extent to which new mutations are stabilizing or destabilizing [34–37]. In other words, the starting point affects the mutant's fitness [38].

That the change in stability for new mutations, $\Delta\Delta G$, depends on ΔG is consistent with larger steps in phenotype space being more likely to land further from an optimum. For example, at one extreme, if a protein were optimal, all possible mutations would be deleterious or neutral; at the other extreme, for a protein furthest from the optimum all possible mutations would be advantageous or neutral; at intermediate values, the relative fractions in the advantageous and deleterious categories are expected to vary. If all possible sequences are ranked by fitness as approximated by the fraction of protein folded into the active state, the probability density takes a form with highest mass in the middle and lowest at the extremes, stemming from results in statistical physics [39].

When a system is at equilibrium with constant Ne and constant selection, any changes that fix are largely neutral and independent of N_e. A compensatory seascape where fitness fluctuates about an equilibrium driven by combinations of slightly deleterious changes and compensatory ones [40, 41] is nevertheless possible for changes of small fitness effect. Compensatory evolution is expected to affect a larger fraction of changes at small N_e, where there is weaker selection against deleterious changes of equivalent magnitude and where mutations segregate for shorter periods of time before fixing and therefore have less chance of fixing together with interacting mutations [3, 42, 43]. The dynamics of fixation themselves are dependent upon N_e. In small populations, mutations fix one at a time, creating more rugged movement on the fitness landscape. With larger populations, mutations and their compensatory changes may fix together due to stochastic tunneling [42]. This not only leads to less expected variance in ΔG, but also to faster neutral walks across sequence space, leading to stronger observed epistatic effects [44].

Further, Goldstein and Pollock as well as others [45] have implied that given the context dependence of fitness effects of mutations, termed epistasis, allowed and forbidden (unfit) amino acid states play a greater role than the relative fitnesses of allowed states. This would result in a model where changes are either neutral or impermissible without a major contribution of positive selection to the compensatory process, corresponding to a substitution model with a neutral rate plus a shifting set of invariant substitutions. This is an interesting extension of a covarion model (involving transitions between variable and invariant states over time) [46] embedded in a substitution matrix. This relates to a model proposed by Usmanova and coworkers [47], where states transition between allowed and disallowed substitution states as a Markov process. The small differences in fitness values for different allowed amino acid states and the corresponding selective coefficients for amino acid transitions may be an artifact of the substitution process model used in the simulations that generated these results, however. Future work will therefore be needed to evaluate if such a model actually explains protein sequence evolution well. The probability of fixation of an introduced mutation would be proportional to N_e under such a model.

Inferring fitness effects with mutation-selection models

Mutation-selection models allow the selective coefficients of amino acid changes to be inferred from sequence data. Given a set of observed codon substitutions over a phylogeny, we can assign each character state a fitness

parameter based on the rate at which that state fixes once it has arisen. The relationship between the probability of fixation and inferred fitness of a character is based on the principles outlined above. The probability P of a substitution occurring therefore depends on μ and N_c together, the selective coefficient s, and N_e. In contrast to many standard phylogenetic methods, the processes of mutation and selection can therefore be examined separately in this framework. This approach has recently become more commonly used in protein evolution studies, particularly with the availability more efficient computers [48–53]. It is of particular interest because it allows putatively adaptive or deleterious substitutions to be identified and separated from background noise in order to carefully characterize selection on proteins. Crucially, in order to obtain meaningful results from genomic data in this manner, we argue that a clear understanding of what is meant by N_e is required and that N_e and N_c are treated distinctly where necessary.

By contrast, the most commonly used statistic to infer selection in coding sequences is the ratio of nonsynonymous to synonymous substitutions (*dN/dS*) [54, 55]. However, its limitations in terms of capturing evolutionary processes have become increasingly clear [45, 56]. For instance, at a given constant selection coefficient for amino acid replacements, the ratio will vary proportionally with the effective population size, and therefore cannot provide information about selection coefficients without additional inference of N_e. In models where N_e and s vary differently across sites (s) and phylogenetic lineages (N_e) and are independently parameterized, they may in fact become identifiable. Furthermore, the dN/dS framework does not consider variable fitness effects conferred by different amino acid changes that can vary in terms of the severity of their impact on the protein structure. It also does not consider that exchanges between a given pair of amino acids may be favored in one direction, from a less fit to a more fit state, but disfavored in the other. These models are therefore consistent only with diversifying but not directional selection [45, 57, 58]. On the other hand, the mutation-selection framework models the probability of introducing a particular kind of mutation by multiplying the per site, per individual mutation rate by the census population size. Once introduced, a mutation's probability of fixation depends on the selective coefficient, the frequency at which it's introduced into the population, and the population's effective size. Applied to real populations with real sizes, structures, and complexities, the concept of the effective population size and its interpretation should be carefully considered. As populations deviate from abstract models, the notions of population size relevant for the introduction of a new mutation and its subsequent probability of fixation are not the same.

To describe the number of new mutations that may become available to a population as a forward looking measure, the relevant parameter is the current census population size. New mutations are introduced at frequency $1/2N_c$ in a diploid population. The historical effective population size is important as a backward looking measure to describe the past effects of selection on mutations in a population [31]. In this framework, the probability of fixation of an amino acid replacement relative to a neutral variant is approximately described by Kimura's diffusion equation [5], where the selection coefficient s and N_e determine amino acid substitution properties. In this case, N_e is the effective size of the population over a specific period of history and specific to lineages of phylogenetic trees where selection has been acting on the system in question. Where population sizes have changed rapidly, this may be a very different quantity than is relevant to describe the trajectories of existing or future variation. Furthermore, the relevant timescales for determining N_e will not be constant across mutations. As selection acts more rapidly on variants conferring large fitness effects, the properties of these variants will reflect a more recent N_e than neutral variants which will have experienced the effects of population parameters for a longer history [59, 60]. Generally, the probability of fixation of any particular mutation will depend on N_c at the time it arose and the range of N_e values during the period of time during which it existed at low frequency.

Local variation in N_e

Demographic factors such as time-varying population sizes, inbreeding, or unequal sex ratios typically cause N_e and N_c to deviate from N_c in a uniform manner across the entire genome. Other factors may create additional, local variation in specific parts of the genome. For instance, in species with heterogametic sexes, sex chromosomes have reduced effective population sizes *and* census population sizes relative to autosomes given their mode of transmission. Organellar genomes, such as those of mitochondria, show diverging locus-specific population sizes. While N_c for mitochondria will be higher than that for autosomes due to the high copy number of mitochondria in each cell, small inter-generational bottlenecks and exclusive transmission through the female germ line leads to a *smaller* N_e than seen in the autosomal genome. In accord with these predictions, rates of both heterozygosity and divergence have been shown to differ between autosomes and sex chromosomes (see, for instance, the fast-X effect [61]). Further, sex differences in the variance of reproductive success (a typical consequence of anisogamy) can lead to additional differences in N_e between sex chromosomes and autosomes that are not reflected in differences in N_c.

According to the population genetics perspective, loci linked to selected sites exhibit locally reduced effective population sizes due to genetic hitchhiking, background selection, and Hill-Robertson interference [62–66]. For neutral loci, this is of little concern as the probability of fixation of a neutral mutation is $1/2N_c$ and does not depend on N_e at all. However, the effects of selection on linked sites drive up the variability in reproductive success at the focal site, increasing the rate of genetic drift and locally decreasing the depth of the local genealogies [67]. This leads to regions of the genome experiencing higher rates of recombination exhibiting lower levels of linkage and therefore less local reduction in N_e due to selection on linked sites ([68–70] and reviewed in [71]). As the fixation of a given mutation depends on the product of the selection coefficient and the ratio N_e/N_c, high rates of recombination can increase the relative rates of fixation of beneficial mutations (and decrease deleterious ones) compared to functionally equivalent mutations in regions of the genome with lower recombination rates or higher densities of constrained loci [72–75].

Where it is feasible to model this local rescaling of the coalescent rate due to selection at linked loci with an explicit model based on the biological process it may be advantageous to our understanding of protein evolution to do so. To account for interference between linked sites, the fixation process can be modeled in linked blocks. Here, the probability of fixation for linked (but not functionally interacting) sites derives from the additive selective coefficients across the set of linked loci and a measure of N_e, call it $N_{e,demographic}$. This parameter scales the underlying Kingman coalescent tree of the population to account for all of the selective neutral processes that create deviations from the Wright-Fisher expectations. There are several possible solutions to implement the computation, including the use of Approximate Bayesian Computation based on simulation [76], inferring probable observed changes in a phylogenetic context, or approximating the effects of linkage through a sampling of the expected number of co-segregating changes coupled to a background distribution of s values. For longer evolutionary timescales, recombination can be incorporated into models of sequence evolution [77]. This would then cleanly separate N_e and s in mutation-selection models and prevent parameter bleed [78] where N_e and s would otherwise have co-linear effects on the shape of underlying local genealogies. At present, such approaches have only been implemented on a limited scale, and there exists ample scope to develop models that describe the processes driving protein evolution in a more elaborate manner. The identifiability of mutation-selection model parameters under different sets of assumptions is a current research topic [56].

Further issues related to model realism

Pitfalls in interpretation of N_e

Great care must be taken in estimating population size parameters from one aspect of observed data with the goal of using it to infer or predict other evolutionary parameters. Though they may often be correlated, estimates of N_e must be considered independently of measures of N_c. If N_e is estimated as a parameter from a model, it is not simply the mean N_e over the branch, but exhibits more complex dynamics, and when derived from an observable statistic of a natural population that may not have a genealogy well-represented by a Kingman coalescent may not be serve as an appropriate estimate for N_e in other contexts. Observations of segregating diversity can be used to estimate forward looking N_e [79], but are best considered more strictly as estimates of future segregating diversity.

Unlike N_e which is less a physical characteristic of a population than a descriptive one, N_c may in theory be unambiguously observed in nature. Outside of highly proscribed settings, however, actually doing so is difficult. Common methods include mark-and-recapture studies [80], plot sampling [81], or simply using body mass as a proxy for the inverse of the population size [27]. Proper estimation of N_c is critical when mutation rates are estimated independently as a per base, per replication rate, as this rate is only useful in an evolutionary setting when scaled by the census population size.

Ratios of statistics involving common definitions of N_e such as dN/dS can be particularly useful as backward looking estimators [82]. Where dN is approximately $2\mu_N sN_e/N_c$ and $\mu_S/(2N_c)$, the N_c terms cancel. With external estimates of the relative rates of non-synonymous and synonymous mutation (μ_N/μ_S) this gives us a direct estimate of $2sN_e$, a population-scaled selection coefficient often referred to as S in the protein evolution literature and γ by population geneticists. Any difficulties in correctly parameterizing or estimating N_e in this case will result in complementary problems in estimating s, and dN/dS has been known to perform poorly as an estimator in certain cases [82–85].

Genomic models not accounting for heterogeneity in N_e will compound difficulties in disentangling s from S or γ. Differences in recombination rate between loci may be more challenging to account for than the reductions in N_e that sex chromosomes experience. When N_e is treated as a mechanistic parameter, it is important that it is accurately defined and reflects the effective population size without absorbing mis-specifications in the model (misparameterizing one parameter with effects that should be fit with a different parameter) [78]. This becomes particularly acute when N_e is used in combination with s, a critical parameter for hypothesis testing in molecular evolution, for example when studying molecular adaptation.

A model that does not account for genomic heterogeneity in the mutation rate (associated with variables such as replication timing, heterozygosity and recombination rate) might also lead to incorrect inference of local N_e. In order to describe protein evolution realistically, an appropriately complex model of selection that is robust to mutation, linkage, epistasis, and covarion-like behaviors that are induced is therefore likely necessary. Here, epistasis is conceptualized as a discrete biochemical process where a change at one position in a protein directly affects amino acid fitnesses at other positions in the same or other proteins, changing probabilities of fixing introduced mutations. When modeled as a site-independent process, this gives rise to covarion-like behavior, where rates of change at a position shift over time [86, 87]. In the simplest form ([46]), this involves a shift between a substitutable position and an invariant site.

Conclusions

We have laid out how multiple distinct parameters associated with population size can jointly be used in mechanistic models of protein evolution. The goal of this discussion is to frame an understanding of population size that is cleanly separable from selection, and that has mechanistic meaning for the process of protein evolution. With this, models that capture the appropriate level of biological complexity to describe observed protein evolution data can be developed, enabling characterization of lineage-specific selective coefficients in comparative genomics.

Abbreviations

dN: The non-synonymous nucleotide substitution rate; *dS*: The synonymous nucleotide substitution rate; ΔG: The Gibbs free energy (of protein folding); $\Delta\Delta G$: The change in the Gibbs free energy (of protein folding); *N*: The size of a population; N_c: The census population size; N_e: The effective population size; *s*: The selective coefficient; *S* or γ: The population scaled selective coefficient; μ: The mutation rate

Acknowledgements

We thank Joanna Masel for discussions during the formulation of ideas that led to this work. We also thank Claus Wilke and two anonymous reviewers for helpful comments on the manuscript.

Funding

This work was supported by NSF grant DBI-1515704.

Availability of data and materials

Not applicable.

Authors' contributions

AP, CCW, and DAL all contributed to the concepts and text of this work. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

David Liberles is a section editor for this journal. The authors declare that they have no other competing interests.

Publisher's Note

Received: 25 May 2017 Accepted: 20 November 2017
Published online: 08 February 2018

References

1. Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A. 2005;102(27):9541–6.
2. Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci. 2007;104(suppl 1):8597–604.
3. Goldstein RA. Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability. Genome Biol Evol. 2013;5(9):1584–93.
4. Wylie C, Shakhnovich E. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci U S A. 2011;108:9916–21.
5. Kimura M. On the probability of fixation of mutant genes in a population. Genetics. 1962;47(6):713.
6. Ewens W. On the concept of the effective population size. Theor Popul Biol. 1982;21(3):373–8.
7. Benner SA, Sassi SO, Gaucher EA. Molecular paleoscience: systems biology from the past. Adv Enzymol Relat Areas Mol Biol. 2007;75:1–132.
8. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, De Koning A, Dokholyan NV, Echave J, et al. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 2012;21(6):769–85.
9. Yang Z, Nielsen R, Hasegawa M. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol. 1998;15(12):1600–11.
10. Doron-Faigenboim A, Pupko T. A combined empirical and mechanistic codon model. Mol Biol Evol. 2007;24(2):388–97.
11. Fisher RA. Xxi.–on the dominance ratio. Proc R Soc Edinb. 1923;42:321–41.
12. Wright S. Evolution in mendelian populations. Genetics. 1931;16(2):97–159.
13. Ewens WJ. Mathematical Population Genetics 1: Theoretical Introduction vol. 27. New York: Springer; 2012.
14. Crow JF. Breeding Structure of Populations. II. Effective Population Number. Iowa: Iowa State Coll. Press; 1954.
15. Nordborg M, Krone SM. Separation of time scales and convergence to the coalescent in structured populations. In: Modern Developments in Theoretical Population Genetics: The Legacy of Gustave Malécot. Oxford: Oxford Unvirsity Press; 2002. p. 194–232.
16. Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M. On the meaning and existence of an effective population size. Genetics. 2005;169(2):1061–1070.
17. Sagitov S, Jagers P, et al. The coalescent effective size of age-structured populations. Ann Appl Probab. 2005;15(3):1778–97.
18. Wakeley J, Sargsyan O. Extensions of the coalescent effective population size. Genetics. 2009;181(1):341–5.
19. Wakeley J. Nonequilibrium migration in human history. Genetics. 1999;153(4):1863–71.
20. Wilkins JF. A separation-of-timescales approach to the coalescent in a continuous population. Genetics. 2004;168(4):2227–44.
21. Kimura M, Ohta T. The average number of generations until fixation of a mutant gene in a finite population. Genetics. 1969;61(3):763.
22. Maruyama T, Kimura M. A note on the speed of gene frequency changes in reverse directions in a finite population. Evolution. 1974;24:161–3.
23. Haldane JBS. A mathematical theory of natural and artificial selection, part v: selection and mutation. In: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 23. Cambridge: Cambridge University Press; 1927. p. 838–44.
24. Whitlock MC. Fixation probability and time in subdivided populations. Genetics. 2003;164(2):767–79.
25. Rousset F. Genetic Structure and Selection in Subdivided Populations (MPB-40). Monographs in Population Biology. Princeton: Princeton University Press; 2013.
26. Nordborg M. Structured coalescent processes on different time scales. Genetics. 1997;146(4):1501–14.

27. Lanfear R, Kokko H, Eyre-Walker A. Population size and the rate of evolution. Trends Ecol Evol. 2014;29(1):33–41.
28. Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI. The influence of selection for protein stability on dn/ds estimations. Genome Biol Evol. 2014;6(10):2956–67.
29. Chi PB, Liberles DA. Selection on protein structure, interaction, and sequence. Protein Sci. 2016;25:1168–78.
30. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 2016;17:109–21.
31. Crow JF, Kimura M, et al. An Introduction to Population Genetics Theory. New York: Harper & Row, Publishers; 1970.
32. Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins Struct Funct Bioinforma. 2011;79(5): 1396–407.
33. Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins Struct Funct Bioinforma. 2002;46(1):105–9.
34. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is facilitated by indirect paths. Elife. 2016;5:16965.
35. Dellus-Gur E, Elias M, Caselli E, Prati F, Salverda ML, de Visser JAG, Fraser JS, Tawfik DS. Negative epistasis and evolvability in tem-1 β-lactamase-the thin line between an enzyme's conformational freedom and disorder. J Mol Biol. 2015;427(14):2396–409.
36. Tufts DM, Natarajan C, Revsbech IG, Projecto-Garcia J, Hoffmann FG, Weber RE, Fago A, Moriyama H, Storz JF. Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude pikas. Mol Biol Evol. 2014;32:311.
37. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary stokes shift. Proc Natl Acad Sci. 2012;109(21):1352–9.
38. Tenaillon O. The utility of fisher's geometric model in evolutionary genetics. Ann Rev Ecol Evol Syst. 2014;45:179–201.
39. Lau KF, Dill KA. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules. 1989;22(10):3986–97.
40. Whitlock MC, Griswold CK, Peters AD. Compensating for the meltdown: The critical effective size of a population with deleterious and compensatory mutations. Ann Zool Fenn. 2003;40(2):169–83.
41. Orlenko A, Teufel AI, Chi PB, Liberles DA. Selection on metabolic pathway function in the presence of mutation-selection-drift balance leads to rate-limiting steps that are not evolutionarily stable. Biol Direct. 2016;11(1):31.
42. Lynch M, Abegg A. The rate of establishment of complex adaptations. Mol Biol Evol. 2010;27(6):1404–14.
43. Cherry JL. Should we expect substitution rate to depend on population size? Genetics. 1998;150(2):911–9.
44. Bastolla U, Porto M, Eduardo Roman H, Vendruscolo M. Statistical properties of neutral evolution. J Mol Evol. 2003;57:103–19.
45. Goldstein RA, Pollock DD. The tangled bank of amino acids. Protein Sci. 2016;25(7):1354–62.
46. Tuffley C, Steel M. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci. 1998;147(1):63–91.
47. Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. A model of substitution trajectories in sequence space and long-term protein evolution. Mol Biol Evol. 2015;32(2):542–54.
48. Jones C, Youssef N, Susko E, Bielawski J. Shifting balance on a static mutation-selection landscape: a novel scenario of positive selection. Mol Biol Evol. 2017;34:391–407.
49. Tamuri AU, Goldman N, dos Reis M. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. Genetics. 2014;197(1):257–71.
50. Rodrigue N, Lartillot N. Detecting adaptation in protein-coding genes using a bayesian site-heterogeneous mutation-selection codon substitution model. Mol Biol Evol Molecular Biology Reports Mol Biol Rep. 2017;34:204–14.
51. Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 1998;15(7):910–7.
52. De Maio N, Schlötterer C, Kosiol C. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Mol Biol Evol. 2013;30(10):2249–62.
53. Spielman SJ, Wilke CO. Extensively parameterized mutation-selection models reliably capture site-specific selective constraint. Mol Biol Evol. 2016;33(11):2990–3002.
54. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding dna sequences. Mol Biol Evol. 1994;11(5):725–36.
55. Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 1994;11(5):715–24.
56. Spielman SJ, Wilke CO. The relationship between dn/ds and scaled selection coefficients. Mol Biol Evol. 2015;32(4):1097–108.
57. Anisimova M, Liberles D. Detecting and Understanding Natural Selection in Codon Models (ed:Cannarozzi and Schneider). Oxford: Oxford University Press; 2012, pp. 73–96.
58. Kryazhimskiy S, Plotkin JB. The population genetics of dn/ds. PLOS Genet. 2008;4(12):1–10.
59. Barton N. Understanding adaptation in large populations. PLOS Genet. 2010;6(6):1–3.
60. Weissman DB, Barton NH. Limits to the rate of adaptive substitution in sexual populations. PLOS Genet. 2012;8(6):1–18.
61. Mank JE, Vicoso B, Berlin S, Charlesworth B. Effective population size and the faster-x effect: empirical results and their interpretation. Evolution. 2010;64(3):663–74.
62. Gossmann TI, Woolfit M, Eyre-Walker A. Quantifying the variation in the effective population size within a genome. Genetics. 2011;189(4): 1389–402.
63. Gillespie JH. Genetic drift in an infinite population: the pseudohitchhiking model. Genetics. 2000;155(2):909–19.
64. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10(3):195–205.
65. Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974;78(2):737–56.
66. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966;8(03):269–94.
67. Birky CW, Walsh JB. Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci. 1988;85(17):6414–8.
68. Charlesworth B. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res. 1994;63(03): 213–27.
69. Charlesworth D, Charlesworth B, Morgan M. The pattern of neutral molecular variation under the background selection model. Genetics. 1995;141(4):1619–32.
70. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genet Res. 1996;67(02):159–74.
71. Comeron JM, Williford A, Kliman R. The hill–robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity. 2008;100(1):19–31.
72. Akashi H, Osada N, Ohta T. Weak selection and protein evolution. Genetics. 2012;192(1):15–31.
73. Weber CC, Hurst LD. Protein rates of evolution are predicted by double-strand break events, independent of crossing-over rates. Genome Biol Evol. 2009;1:340–9.
74. Campos JL, Halligan DL, Haddrill PR, Charlesworth B. The relation between recombination rate and patterns of molecular evolution and variation in drosophila melanogaster. Mol Biol Evol. 2014;31(4):1010–28.
75. Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. Adaptive evolution is substantially impeded by hill-robertson interference in drosophila. Mol Biol Evol. 2015;33:442–55.
76. Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring coalescence times from dna sequence data. Genetics. 1997;145(2):505–18.
77. Arenas M, Posada D. The influence of recombination on the estimation of selection from coding sequence alignments. In: Fares MA, editor. Natural Selection: Methods and Applications. Boca Raton: CRC Press; 2015. p. 112–25.
78. Liberles DA, Teufel AI, Liu L, Stadler T. On the need for mechanistic models in computational genomics and metagenomics. Genome Biol Evol. 2013;5(10):2008–18.
79. Lynch M, Conery JS. The origins of genome complexity. Science. 2003;302(5649):1401–4.
80. Besbeas P, Freeman SN, Morgan BJ, Catchpole EA. Integrating mark–recapture–recovery and census data to estimate animal abundance and demographic parameters. Biometrics. 2002;58(3): 540–7.
81. Borchers DL, Buckland ST, Zucchini W. Estimating Animal Abundance: Closed Populations vol. 13. London: Springer; 2002.

82. Nabholz B, Uwimana N, Lartillot N. Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino-acid replacement in mitochondrial genomes of mammals and birds. Genome Biol Evol. 2013;5:1273–90.
83. Weber CC, Nabholz B, Romiguier J, Ellegren H. Kr/kc but not dn/ds correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. Genome Biol. 2014;15(12):542.
84. Lartillot N. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. Mol Biol Evol. 2013;30(2):356–68.
85. Hua X, Bromham L. Darwinism for the genomic age: connecting mutation to diversification. Front Genet. 2017;8:12.
86. Wang HC, Spencer M, Susko E, Roger AJ. Testing for covarion-like evolution in protein sequences. Mol Biol Evol. 2007;24(1):294–305.
87. Miyamoto MM, Fitch WM. Testing the covarion hypothesis of molecular evolution. Mol Biol Evol. 1995;12(3):503–13.