The Evolution of Multimeric Protein Assemblages

Michael Lynch*

1Department of Biology, Indiana University

*Corresponding author: E-mail: milynch@indiana.edu.

Abstract

Although the mechanisms by which complex cellular features evolve constitute one of the great unsolved problems of evolutionary biology, it is clear that the emergence of new protein–protein interactions, often accompanied by the diversification of duplicate genes, is involved. Using information on the levels of protein multimerization in major phylogenetic groups as a guide to the patterns that must be explained and relying on results from population-genetic theory to define the relative plausibility of alternative evolutionary pathways, a framework for understanding the evolution of dimers is developed. The resultant theory demonstrates that the likelihoods of alternative pathways for the emergence of protein complexes depend strongly on the effective population size. Nonetheless, it is equally clear that further advancements in this area will require comparative studies on the fitness consequences of alternative monomeric and dimeric proteins.

Key words: complex adaptation, dimer, genome evolution, heteromer, molecular evolution, random genetic drift.

Although the earliest cells must have been considerably simpler than any of today’s free-living organisms, the mechanisms by which complex cellular features emerge remain unclear. However, recent insights into the molecular architecture of protein complexes and the population-genetic conditions required for their establishment provide guidance as to the range of likely possibilities. Many of the protein complexes that comprise cellular features are assembled from subunits derived from the same gene or from loci related via gene duplication rather than from products of unrelated genes. This appears to be the case, for example, for the flagellum (Liu and Ochman 2007), centrioles (Carvalho-Santos et al. 2010), the nuclear pore complex (Alber et al. 2007), the spliceosome (Scofield and Lynch 2008), the cytoskeleton (Löwe and Amos 2009), the proteasome (Hughes 1997), chromatin-remodeling complexes (Monahan et al. 2008), ion channels (Dent 2010), nucleosomes (Malik and Henikoff 2003), the ribosome (Smits et al. 2007), and many other components of prokaryotic and eukaryotic cells. Thus, an essential first step for understanding the emergence of complex cellular adaptations is the development of a general theory for the evolution of protein–protein interactions.

Potential advantages to protein complex formation include increased structural size and diversity, reduced problems of folding single large proteins, and increased opportunities for allosteric regulation and protein activation (Marianayagam et al. 2004; Hashimoto et al. 2011). On the other hand, proteins with oligomerization potential can also be dangerous, human disorders involving the production of inappropriate protein aggregates (e.g., Alzheimer’s and Parkinson’s disease and amyotrophic lateral sclerosis) being prime examples (Chiti and Dobson 2009; Huntington 2011). Overexpression of genes encoding adhesion-prone proteins often encourages deleterious promiscuous protein–protein interactions (Semple et al. 2008; Vavouri et al. 2009), and such negative selection pressure seems to be reflected in the fact that many highly expressed genes have features that minimize the propensity for self-aggregation (Tartaglia et al. 2007).

This fine line between adaptation and potentially malfunctional protein aggregation raises the possibility that some oligomeric associations may not have arisen initially as de novo adaptations, but as simple compensatory mechanisms for ameliorating defects incurred by individual subunits. There is, for example, a significant tendency for the proteins of larger more complex organisms (with smaller effective population sizes) to be more adhesive and hence more likely to engage in promiscuous protein–protein interactions (Fernández et al. 2004; Fernández and Lynch 2011). This situation, which results from the accumulation of amino acid changes that reduce the protection of backbone hydrogen bonds from water attack, may secondarily promote the evolution of multimeric complexes with better overall wrapping. Relative to the situation in prokaryotes, eukaryotic proteins also exhibit substantial increases in the lengths of interdomain regions (Wang et al. 2011), which presumably influences the tendency to engage in intramolecular versus intermolecular interactions. It is also notable that universal mutation pressure toward A/T nucleotides (Hershberg and Petrov 2010; Hildebrand et al. 2010; Lynch 2010a) encourages a bias toward more hydrophobic residues with A/T-rich codons (Knight et al. 2001; Bastolla et al. 2004), which might further reduce the ability of species with relatively small population sizes to resist the accumulation of mutations that encourage protein adhesion.

Background

Approximately 65% of proteins in prokaryotes and 55% of those in eukaryotes exist as dimers or higher-order
complexes (not including very high-order structures such as the cytoskeleton, ribosomes, etc.) (fig. 1, left). Among complexes, homomers are about \(4\times\) more frequent than heteromers in unicellular species, whereas the two types are equally frequent in vertebrates. As a consequence, heteromers constitute \(\sim10\%\) of proteins in unicellular species but nearly \(30\%\) in vertebrates. In all phylogenetic groups, there is also a strong nearly negative exponential frequency distribution for the numbers of subunits within protein complexes (fig. 1, right). Although there are many possible explanations for such distributions, such patterns can arise as natural outcomes of a steady-state process in which subunits are stochastically gained and lost (Lynch 2007).

These observations are subject to bias, as they are derived from the protein structures that happen to be deposited into the Protein Data Base (the PDB), which need not be random. Nevertheless, the fact that very similar distributions are found for Eubacteria and Archaebacteria as well as for invertebrates and vertebrates, suggests that the results in figure 1 provide reasonable first-order approximations of the dispersion of protein-complex sizes in the major domains of life.

At least three issues are relevant to understanding the conditions under which multimeric structures might evolve. First, colocalization of gene products is an essential starting point for the coevolution of protein subunits. The frequent subcellular localization of specific mRNAs promotes spatial aggregation (Holt and Bullock 2009), and the innate tendency for proteins to self-aggregate generates additional oligomerization potential (Ispolatov et al. 2005; Wright et al. 2005; Lukatsky et al. 2007; Andre et al. 2008). By physically tying two loci together, gene fusion provides still another powerful way to facilitate mutual adhesion between two protein domains, with subsequent gene fission potentially leading to the evolution of a heterodimeric relationship (Kuriyan and Eisenberg 2007). Alternatively, loop shortening between two interfacing domains of a monomeric protein can promote homodimer formation when the ancestral binding contacts can no longer access each other (Bennett et al. 1994).

Second, the establishment of stable protein–protein interactions often involves the substitution of just a small number of amino acids or the insertion/deletion of a small stretch of residues, with individual mutations often having mildly deleterious effects unless compensated by changes at other locations (Jones and Thornton 1995; Bogan et al. 1998; Janin et al. 1998; Hashimoto et al. 2010). Protein complexes are generally stabilized by hydrophobic interactions and/or hydrogen bonding, and the addition of just a few pairs of appropriately spaced residues (such as Arg–Asp, Lys–Glu, or Cys–Cys) or the elimination of a few unfavorable contacts (such as Arg–Arg or Glu–Glu) may be sufficient to generate a functional interface. Likewise, the alteration of a few key residues can cause a multimer to revert to a monomeric state.

Third, as products of more than one genetic locus, heteromers are expected to incur elevated mutation rates to defective structures relative to homomers. Heteromers may also experience stoichiometric imbalance if the source genes are expressed at different levels, leading to the circulation of potentially harmful monomeric subunits. Thus, if a favorable homodimeric interaction is to be displaced evolutionarily by a heterodimeric structure, the latter must either enjoy a net advantage or the magnitude of genetic drift must exceed the net mutational and selective disadvantage (Lynch et al. 2001; Lynch 2007). Drawing from these observations, an
attempt is made below to identify the population-genetic conditions that most plausibly promote the emergence of multimeric protein assemblages. The focus will be on the evolution of dimers, although the approaches taken should have more general utility, as most higher-order complexes are derived from the dimerization of lower-order structures (e.g., tetramers are often dimers of dimers). It is assumed that a protein starts its evolutionary history as a functional monomer, with certain kinds of amino acid substitutions inducing structural modifications that cause the monomeric subunit to become more prone to adhesion with other localized proteins with similar features. Reliable aggregation with a complementary subunit may simply maintain the original level of protein efficiency, increase functionality beyond that of the initial monomer, and/or yield a multimer with entirely novel capabilities.

Although these are the general steps in the transformation of a monomer to a dimer, there are several dimensions to the problem. First, a newly arisen dimer may simply result from the aggregation of monomeric subunits of the original protein (homodimerization) or be composed of proteins from divergent genes (e.g., heterodimerization involving the members of a duplicate-gene pair). Second, the emergence of a dimeric structure may require the accumulation of one or more neutral or deleterious intermediate state mutations. Third, for situations involving gene duplication, the potential exists for one locus to initially evolve a neofunctionalized homodimer, while the original locus maintains the initial function conferred by the monomeric state, with still another gene duplication allowing the two subunits of the homodimer to secondarily diverge to a more nonsymmetric (heterodimeric) state (Ispolatov et al. 2005; Pereira-Leal et al. 2007; Reid et al. 2010).

Results

The initial focus will be on the origin of homodimers (the aggregation of two protein subunits derived from the same locus), both because of the high frequency of such aggregates and because such a condition provides a likely launching pad for the emergence of heterodimers following gene duplication and subsequent divergence. Three alternative models will be considered, with emphasis on the dependence of the mean time required for a population-level transition from a monomeric to a homodimeric state on population size, mutation rate, and the selective advantages/disadvantages of the stepwise mutations contributing to the final adaptation. Most of the derived expressions can be readily modified to describe the reversion of dimers to monomers.

The theory, which is supported by computer simulation results in the supplemental material (Supplementary Material online), is developed initially in a manner that treats population sizes and mutation rates as independent parameters. However, although such an approach follows past tradition, empirical evidence implies that mutation rates are strongly negatively correlated with effective population sizes across the tree of life, with the single-site rate of base-substitution mutation per generation being approximated by $0.000025N^{-0.6}$ in observed taxa, where $N$ is the effective population size (Lynch 2010c). Thus, to determine the natural scaling of the time to establishment with effective population sizes and also to reduce the dimensionality of the analyses, the results in the main text will rely on this relationship. Unless otherwise stated, it is assumed that the genetic effective population size is equal to the actual number of reproductive adults.

The Domain-Swapping Model

A frequently invoked mechanism for the origin of homodimers is encapsulated in the domain-swapping model (Bennett et al. 1994), whereby a monomeric protein with two interfacing domains is physically altered in such a way that binding between domains within the same polypeptide chain is prevented and can only be accomplished by reciprocal domain sharing between two monomeric subunits (fig. 2). An attractive feature of this model is the presence of complementary binding domains in the ancestral protein, and convincing examples of domain-swapping proteins exist (Liu and Eisenberg 2002), but the conditions required for such an evolutionary transition remain unclear. Moreover, it is plausible that the process is bidirectional, with mutations in domain-swapping dimers sometimes causing reversions to the monomeric condition.

Here, we consider the simplest case in which an allele for the domain-swapping protein arises by a single mutation that denies self-accessibility within the ancestral monomer (such as a deletion in a loop between the two domains of the ancestral protein). If the dimer is beneficial, such a mutant allele can readily proceed to fixation by positive selection in a haploid species. However, in a diploid outcrossing species, the mutant allele will initially be present exclusively in heterozygotes, where fitness may be compromised by the production of malfunctional composites of the two alternative monomeric subunits, for example, chimeras with unbound domains. Such heterozygote disadvantage will impose a barrier to the spread of the mutant allele, as this requires the population to pass through a bottleneck in mean fitness, a highly unlikely event unless the power of random genetic drift is substantially greater than the heterozygote disadvantage. The magnitude of the latter will presumably depend on the rate of folding of the ancestral monomer and the overall cellular concentration of both allelic products, as slow folding and/or high concentration should magnify the likelihood of encounters between the two alternative proteins.

In this and all remaining analyses, we will assume an idealized random-mating Wright–Fisher population containing $N$ diploid individuals (i.e., discrete generations with consecutive phases of selection, mutation, and random genetic drift), inquiring as to the time that is expected to elapse between a starting point of a population monomorphic for allele $A$, which produces a monomer, and a final state of monomorphism for domain-swapping allele $a$. This total time for establishment, denoted as $t_e$, can be viewed as the sum of the arrival time for the first mutant $a$ allele
destined to fixation and the subsequent time for such an allele to progress from initial frequency \(1/(2N)\) to 1.0. 

Letting the fitnesses of genotypes AA, Aa, and aa be 1, \(1 - \delta\), and 1 + \(s\), respectively, the probability of fixation of an allele a exhibiting heterozygote inferiority is given by

\[
\phi = \frac{\text{erf}\left[\left(\frac{p_0 - 0.5}{(1 + \omega)}\right)\sqrt{4\Phi(1 + \omega)} + \text{erf}\left(\frac{\theta}{(1 + \omega)}\right)\right]}{\text{erf}\left[\left(1 - 0.5/1(1 + \omega)\right)\sqrt{4\Phi(1 + \omega)} + \text{erf}\left(\frac{\theta}{(1 + \omega)}\right)\right]},
\]

where \(p_0\) is the initial frequency of \(a\), \(\theta = Ns\), \(\omega = s/(2\delta)\) (Walsh 1982), and

\[
\text{erf}(x) = \int_0^x e^{-y^2} dy,
\]

which is readily solved with numerical approximations in Abramowitz and Stegun (1964). For generations in which a mutation arises, the initial frequency of mutant a alleles has expectation \(p_0 \approx (1/(2N)) + u\), where \(u\) is the mutation rate from A to a, which reduces to \(\approx 1/(2N)\) when \(2Nu \ll 1\), and \(\approx u\) when \(2Nu \gg 1\). With \(2Nu\) new mutations arising per generation, the average arrival time of the first mutation destined to fix is \(\approx 1/(2Nu\phi)\), and the expected time to complete establishment of the domain-swapping allele is then

\[
\bar{t}_e \approx \frac{1}{2Nu\phi} + \bar{t}_f,
\]

where \(\bar{t}_f\) is the mean fixation time (approximated with the expression for a recessive mutation presented in the supplementary material [Supplementary Material online]). Results from computer simulations of newly arisen mutations in the classical Wright–Fisher framework demonstrate that these formulations are quite accurate (supplementary fig. 1, Supplementary Material online).

Of special interest is the population size \(N^*\) above which the efficiency of selection is so strong that there is effectively no possibility of passing through the fitness bottleneck imposed by heterozygotes. With heterozygotes having a fitness reduction of \(\delta\), homozygotes an advantage of \(s\), and \(p\) being the frequency of the domain-swapping allele, mean population fitness is defined as \(W = 1 - 2p(1 - p)\delta + p^2 s\). The latter reaches a minimum when \(p = \delta/(s + 2\delta)\), with \(p < p^*\) implying net selection against and \(p > p^*\) net selection in favor of allele a. Thus, the key issue is whether a mutant allele can drift from initial frequency \(p_0\) to \(p^*\). As a first-order approximation, when \(p\) is small and the frequency of aa homozygotes is negligible, allele a acts like a deleterious mutation being removed from the population at rate \(\delta\). Transition to the domain-swapping state then requires allele a to drift a distance \(p^*\) against a persistent selection gradient of \(\sim \delta\), the probability of which is given by the diffusion approximation.
Noting that \( \phi_m \to 0 \) as \( 4N\delta p \) becomes large, the population size barrier to the establishment of the domain-swapping protein is then
\[
N' \sim \frac{1}{\delta^2} = \frac{s + 2\delta}{\delta^2}.
\]

Although these results provide a framework for evaluating the mean time to establishment of an underdominant mutation for arbitrary magnitudes of drift, mutation, and selection, as noted above, the strong negative correlation between mutation rates and effective population sizes across the tree of life will cause the mean time to establishment in natural populations to scale more weakly with population size than would be the case with independent \( N \) and \( u \). Nevertheless, even when this issue is accounted for, it is still clear that \( T_c \) increases with population size under the domain-swapping model (fig. 2).

Under this model, a transition from a monomeric to a dimeric state is most plausible under two sets of conditions: 1) a haploid population in which heterozygote disadvantage is never experienced and 2) a diploid population in which selection against heterozygotes is inefficient, either because the effective population size is small or because the reduction in heterozygote fitness is negligible. In both cases, the mean time to establishment scales with \( \left( Nu\phi_f\right)^{-1} \) (preceded by a factor of 0.5 in the case of diploidy). Thus, assuming negligible selection on heterozygotes, if the selective advantage for homozygotes (or the beneficial haploid state) substantially exceeds the power of drift (\( Nu > 1 \)), then \( \phi_f \approx 2s \), and \( T_c \) scales simply with \( \left( Nus\right)^{-1} \), which implies proportionality to \( N^{-0.4} \) when the mutation-rate scaling noted above is employed. This result even holds when the effective population size \( (N) \) is unequal to the actual population size \( (N_0) \), as in this case the mutational input is proportional to \( N_0u \) and \( \phi_f \) to \( sN_0/N_0 \), the product again being \( Nus \). If, on the other hand, the power of drift overwhelms even positive selection for homozygotes, \( \phi_f \approx 1/(2N) \), and \( T_c \) scales simply with \( u^{-1} \), which is \( \propto N^{-0.6} \).

Notably, these rather different scalings are on a per-generation time scale, and a more meaningful comparator across the tree of life ought to involve absolute time units. Because organisms with small \( N \) generally have much longer generation times than those with large \( N \) (e.g., generation times of land plants and metazoans are typically on the order of weeks to years, whereas those for microbes are on the order of hours to days), the scaling of \( T_c \) with \( N \) in absolute time units must be more negative than that based on generations. The approximate scaling of generation length with \( N \) suggested in Lynch (2010b) implies a scaling of \( T_c \) in the above cases between \( N^{-1.2} \) and \( N^{-0.2} \). Thus, if an advantageous allele conferring domain-swapping ability is able to proceed to fixation with no significant inhibition from heterozygote disadvantage, such an architecture is expected to emerge most rapidly in species with large \( N \).

In contrast, if there is significant enough heterozygote disadvantage that \( N > (s + 2\delta)/\delta^2 \), it is virtually impossible for a domain-swapped allele to proceed to fixation. Clearly, a knowledge of the fitness consequences of mixtures of the products of ancestral and derived alleles is essential to resolving how readily domain swapping can evolve in diploid populations. There appears to be no direct evidence on the matter of whether domain-swapping dimers confer greater or lesser fitness than monomers, and if \( s = 0 \), the threshold-barrier to domain-swapping evolution is just \( N' = 2/\delta \).

The theory presented above is entirely general in that a simple change in definitions of terms is all that is required for estimating the reverse transition of homodimer to monomer. In principle, a lineage might wander back and forth between alternative states, with the long-term probability of being monomeric equaling \( r_{dm} / (r_{dm} + r_{md}) \), where \( r_{dm} \) and \( r_{md} \) are the rates of transition from dimer to monomer and vice versa.

### The Compensatory Mutation Model

A second scenario by which homodimers might arise involves two (or more) changes that are individually neutral or deleterious but together alter the monomeric structure in a way that encourages stable dimeric complexation. In contrast to the situation with the domain-swapping model, here the ancestral monomer is nonadhesive and therefore not compromised by the presence of derived alleles within heterozygous carriers. Consequently, even though more mutations are required to make the transition to a homodimeric state, the population need not experience a bottleneck in mean fitness because recurrently introduced intermediate-step alleles are either neutral or kept rare by selection while also serving as substrate for secondary mutations to beneficial final-step alleles, which then enjoy a clear path to fixation by positive selection.

We start with a two-site model with no recombination, with single-site mutations experiencing a reduction in relative fitness equal to \( \delta \), and dimerization of the double mutant causing a fitness increase of \( s \) per gene copy (fig. 3). A diploid random-mating population is again assumed, with the genotypic fitnesses being determined by the additive effects of the two alleles. Single mutations of relevance to the final adaptation are assumed to arise at rate \( u \), summed over all relevant sites. Back mutations are ignored, as we assume that multiple amino acid alterations can lead to the relevant (and functionally equivalent) first-step changes so that the forward mutation rate dominates the evolutionary process. Mutations with major deleterious effects are ignored as well, as these will remove all alleles from the population at equal rates. Finally, enough potential sites are assumed to be involved in the initial dimerization process that the mutation rate can be assumed to be the same at both steps in the process.

As in the case of the domain-swapping model, the rate of establishment of the homodimer under this model depends on the population size (Weissman et al. 2009; Lynch 2010b; Lynch and Abegg 2010). Starting with the case of neutral intermediates (\( \delta = 0 \)), if the population is sufficiently small, the evolutionary dynamics will proceed in two
Expected number of generations to establishment of a homodimer requiring two mutational changes under the assumption that the mutation rate per site scales negatively with the effective population size, as described in the text, with \( u \) being assumed to be equal to \( 10^x \times \) the per-site rate. Results are given for three levels of selective disadvantage of the first-step mutants (\( \delta \)) and four levels of advantage for the second-step mutants (\( s \)), using the expressions presented in the text. Note that the times to establishment at small population sizes are lower with \( \delta > 0 \) than with \( \delta = 0 \) because first-step alleles are assumed to be present at the frequency defined by selection–mutation balance in the former case but are required to arise by new mutations in the latter case. In the lower right, only monomeric subunits with two complementary alterations assemble into dimers. The small discontinuities in the upper left curves result from the use alternative approximations noted in the supplementary material (Supplementary Material online).

discrete steps, with a first-step mutant becoming fixed by drift prior to the arrival of a successful second-step mutation. The rate of establishment by this sequential path is equal to the reciprocal of the sum of the expected arrival times for the mutations contributing to the adaptation, which in the case of neutral intermediates is

\[
r_s \simeq \frac{u}{1 + (1/2N\phi_b)},
\]

where

\[
\phi_b = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}
\]

is the probability of fixation of a beneficial (second-step) mutation with additive effects, initially in single copy (Kimura 1962). This expression essentially fully describes the rate of establishment when \( N \ll 1/\sqrt{8u\phi_b} \), but this is a potentially narrow domain as with \( u = 10^{-7} \) and \( \phi_b = 0.01 \), \( N \) must be \( < 10^4 \).

With larger population sizes, prior to fixation, first-step alleles can lead to the establishment of second-step mutations by a process sometimes called stochastic tunneling or the rescue effect. Some of the basic results have been derived in different contexts by Walsh (1995), Lynch et al. (2001), Komarova et al. (2003), and Iwasa et al. (2004). Consider a newly arisen single-site mutation, assumed to be effectively neutral so that the initial dynamics are governed entirely by random genetic drift, and the probability of being lost is \( 1 - (2N)^{-1} \) unless a secondary mutation to a beneficial function can propel a sublineage to fixation. As the details for this rescue effect have been worked out previously, it is simply stated here that the probability that a first-step mutation acquires a beneficial (homodimerizing) secondary mutation destined to fixation is approximately equal to \( \sqrt{u\phi_b} \). Noting that \( 2Nu \) first-step mutations arise per generation, the expected rate of appearance of second-site mutations destined to fixation by stochastic tunneling is then \( 2Nu\sqrt{u\phi_b} \), with the overall rate being

\[
r_t \simeq \frac{u\phi_b}{\sqrt{N}} \left( \frac{1}{2Nu} + 1 \right)^{-1},
\]

where the second term (negligible when \( 2Nu \ll 1 \)) accounts for the additional waiting time for the second mutation (Weissman et al. 2009). For populations large enough to generate at least one single-site mutation per generation (\( 2Nu > 1 \)), a more appropriate approximation is

\[
r_t \simeq 2u\sqrt{N\phi_b}/\pi
\]

(Weissman et al. 2009).

In summary, the mean time to establishment of the homodimer via two mutations with a neutral intermediate state is equal to the sum of two terms: the reciprocal of the rate of appearance of double mutants destined to fixation and the time to fixation. The former can be approximated by the sum of the sequential and stochastic tunneling rates for \( 2Nu < 1 \), and otherwise by the semideterministic tunneling
rate, equation (7b). An expression for the time to fixation is derived in the supplementary material (Supplementary Material online). Taken together, these expressions provide a relatively simple and accurate description of a fairly complex process as illustrated by comparison with computer simulations (supplementary fig. 2, Supplementary Material online).

Ignoring the fixation time for populations large enough to experience the rescue effect, equation (7a) implies that provided \( 2Nu < 1 \), \( T_e \) scales with \( [2Nu(u\phi_b)_{0.5}]^{-1} \) for the two-site model with a neutral intermediate (fig. 3). This expression can be extended to allow for additional intermediate-step mutations en route to the final adaptation by simply substituting nested terms of \( (u\phi_b)_{0.5} \) for each previous \( \phi_b \), yielding, for example, \( [2Nu \cdot u^{0.5} \cdot (u\phi_b)_{0.25}]^{-1} \) for the three-site model. Assuming large enough population sizes that \( \phi_b \simeq 2s \), for two-, three-, four-, and five-step adaptations, this leads to a population size scaling for \( T_e \) of \( N^{-0.10}, N^{-0.095}, N^{-0.125} \), and \( N^{-0.1625} \), respectively, on a per-generation basis, and \( N^{-0.70}, N^{-0.75}, N^{-0.675} \), and \( N^{-0.6375} \) on an absolute time basis. When \( 2Nu > 1 \), the scaling is altered to \( [4Nu^2\phi_b/\pi]^{0.05}]^{-1} \) for the two-site model, which translates to a scaling of \( N^{-0.70} \) in absolute time. In this case, for larger numbers of intermediate states, we expect the exponent on \( u \) to increase accordingly, leading to an absolute time scaling for \( T_e \) of \( N^{-0.4}, N^{-0.17}, \) and \( N^{-0.025} \), respectively, for three-, four-, and five-site adaptations. Because the mutational input is proportional to the actual population size \( (N_s) \), which clearly scales more rapidly than linearly with \( N \) for species with large \( N_s \) (Lynch 2007; Neher and Shraiman 2011), actual scalings of \( T_e \) with \( N \) are likely to be even more negative than those given above.

This general approach is readily extended to situations in which the intermediate states are deleterious. Considering first the two-site model, the rate of establishment by the sequential pathway becomes

\[
r_s \simeq 2Nu/[((1/\phi_d) + (1/\phi_b))],
\]

where \( \phi_d \) is the rate of fixation of the deleterious intermediate, obtained by substituting \(-\delta\) for \( s \) in equation (6); and the rate of establishment by the rescue effect is

\[
r_t \simeq 2Nu^2\phi_b/\delta.
\]

This expression follows simply from the fact that prior to the arrival of a second-step mutation, deleterious first-step mutations will remain in approximate selection–mutation balance with frequency \( u/\delta \), with each copy having a probability \( u\phi_b \) of giving rise to a successful second-step mutation. The mean time to establishment can then again be approximated by adding the reciprocal of the sum of the two arrival rates to the fixation time (given in the second section of the supplementary material [Supplementary Material online]), an approach that yields an excellent fit to simulated data (supplementary fig. 3, Supplementary Material online).

If multiple \( (d > 2) \) equally deleterious intermediate states precede a successful dimer, the rate of stochastic tunneling is \( \simeq 4N(u/\delta)^d(s/\delta) \) per generation (Weissman et al. 2009; Lynch and Abegg 2010). Again, ignoring the final (and usually shorter) phase of fixation, the mean time to establishment in absolute time units then scales as \( N^{-0.6}, N^{-0.6}, N^{-1.12} \), respectively, for adaptations involving one, two, three, and four intermediate deleterious states. As shown in the supplementary material (Supplementary Material online), if the spatial clustering of mutations causes the rate of double mutation to be substantially greater than \( u^2 \), as the evidence suggests (Schrider et al. 2011), the rate of adaptation under the deleterious intermediate model can be greatly accelerated, although the scaling with \( N \) is unaltered.

Finally, it is worth considering the consequences of recombination, as the preceding results assumed complete linkage between selected sites. Although an intermediate level of recombination (\( \sim s/2 \) between sites for two-site adaptations) maximizes the rate of establishment, the effect is not great, and recombination has little influence on the scaling of \( T_e \) with \( N \) for neutral intermediates (Lynch and Abegg 2010; Weissman et al. 2010). However, in the case of deleterious intermediates, when the rate of recombinational breakdown exceeds the selective advantage, the consistent return of adaptive alleles to deleterious intermediate states strongly inhibits the rate of establishment at large \( N \).

With the average recombination rate in prokaryotes being on the order of \( 10^{-9} \) per nucleotide site (Lynch 2007), and average lengths of coding regions being on the order of 1 kb, the maximum rate of recombination between sites in the same protein is just \( 10^{-6} \) per generation. In this case, the scaling of \( T_e \) with \( N \) is expected to be nearly independent of the recombination rate unless selection is extremely weak. On the other hand, recombination rates in unicellular eukaryotes tend to be on the order of \( 10^{-7} \) to \( 10^{-6} \) per site, whereas those in metazoans and land plants generally range from \( 10^{-9} \) to \( 10^{-7} \), yielding more potential for recombinational interference in the establishment of epistatically interacting mutations, particularly for sites contained within different exons in species with long introns.

Taken together, these results suggest that under the compensatory mutation model, transitions to a homodimeric state will generally be either made more rapidly in large populations or at approximately equal rates at all population sizes. The only potential exceptions to this general trend are when the intermediate states are deleterious and the recombination rate exceeds the final selective advantage or when four or more linked intermediate states must be traversed to achieve the final adaptation (although in this case, the time to establishment may be so large as to make such a pathway highly unlikely). This negative to weak positive scaling of \( T_e \) with population size under the compensatory mutation model is quite different than the situation with domain swapping even though there is an ancestral precedent for domain interaction in the latter case, and even though the compensatory mutation route may involve deleterious intermediate alleles.
Gene Duplication Models

Although gene duplication may be the most common route to the origin of heterodimers, it may also potentiate the emergence of homodimers. To fulfill the definition of a homodimer, the products of a duplicate gene must interact entirely with each other rather than with those of the paralogous gene. In principle, this might be accomplished if the two paralogs were expressed at different subcellular locations, different times in development, etc., perhaps made possible by an initial step of subfunctionalization resulting from incomplete regulatory-region duplication (Force et al. 1999; Katju and Lynch 2006). We will assume that following such a duplication event, the pair of gene duplicates suffers one of two fates: 1) loss of one of the copies via the fixation of a nonfunctionalizing mutation or 2) permanent preservation of both copies, as one acquires beneficial mutations to a new function (in this case, embodied in the creation of a homodimer) at the expense of the ancestral essential function, which is retained by the other copy. This model does not deny the possibility of neofunctionalizing mutations that do not involve homodimeric construction but the concern here is primarily with this particular path. The following results rely on a number of methods developed in Lynch et al. (2001). Gene duplications arise at rate $2ND$ at the population level, with $D$ being the rate of duplication per gene per generation, and the initial frequency at a new locus being $1/(2N)$, with all other “absentee alleles” at the novel locus being effectively null. Starting with the case of complete linkage between duplicates, a newly arisen gene duplicate is initially destined to fixation with probability $1/(2N)$ and to loss with probability $1 - (2N)^{-1}$. Conditional on proceeding to fixation and assuming there are no intrinsic advantages to functionally redundant duplicates, either one member of the pair will be lost by a nonfunctionalizing mutation (which arise at rate $u_0$ per locus) or the pair will be preserved by a neofunctionalizing mutation in one copy (the probability of which is denoted $\rho_1$). Conditional on the haplotype carrying the duplicate pair being initially destined to loss, the possibility exists that it will be rescued and propelled to fixation by a neofunctionalizing mutation (the probability of which is denoted $\rho_2$). Letting $1 - (2N)^{-1} \simeq 1$, the rate of establishment of the new locus (which in this case specifically leads to dimerization) can then be expressed as

$$r_d \simeq D \left( \frac{\rho_1}{\rho_1 + 2u_0} + 2N\rho_2 \right). \quad (9)$$

The preservation probabilities, $\rho_1$ and $\rho_2$, may take on various forms depending on the evolutionary path to homodimerization, but unless the population size is extremely small, the second term will almost always dominate so that $r_d \simeq 2ND\rho_2$. For example, under the assumption that just a single mutation is required for dimerization at the new locus, and letting $u_1$ be the rate of origin of neofunctionalizing mutations, the rate of fixation of neofunctionalizing mutations conditional on fixation of the duplicate locus is $\rho_1 \simeq 2Nu_1\phi_0$; and following the logic outlined above for the rescue effect, $\rho_2 \simeq \sqrt{u_1\phi_0}$, where $\phi_0$ is the fixation probability defined by equation (6), with $s$ being the selective advantage of each copy of the dimerizing allele. The simplified function, $r_d \simeq 2ND\rho_2$, provides a generally good fit to data acquired by simulations (supplementary fig. 4, Supplementary Material online).

If more than one mutation is required for dimerization, the previous expressions must be modified to allow for rare sequences of mutational events that can insure fixation by positive selection. For example, if a neutral intermediate mutation is required prior to the construction of a dimer by a second mutation, $\phi_0$ in the previous expressions for $\rho_2$ must be replaced by the probability that a first-step mutation is rescued by a second-step mutation, $\sqrt{u_1\phi_0}$, where $u_2$ is the rate at which first-step alleles acquire second-step mutations, here assumed to equal $u_1$, leading to $\rho_2 \simeq \sqrt{u_1\phi_0}$ (Weissman et al. 2010). In addition, because of the longer time span involved in the acquisition of additional mutations, the probability that duplicate genes are hit with a nonfunctionalizing mutation prior to procuring otherwise adaptive mutations must be considered. These issues, as well as derivations for smaller population sizes, are developed in the supplementary material (Supplementary Material online), where it is again shown that the overall results closely approximate observations derived from computer simulations (supplementary fig. 4, Supplementary Material online).

A key feature of this gene duplication model is that although neofunctionalized (dimerizing) a alleles will almost certainly historically arise by mutation at the ancestral locus prior to gene duplication, because aa homozygotes are lethal due to the lack of the essential ancestral function, allele a cannot advance beyond the low frequency expected under balancing selection (heterozygote superiority). Tandem a–a duplicates that are completely linked also cannot contribute to homodimer origin because the absence of essential ancestral function in the linked pair would again prevent fixation.

On the other hand, there are two ways in which an A/a polymorphism at the ancestral locus can lead to rapid neofunctionalization if the duplicated gene is unlinked to its parental copy (Spofford 1969; Lynch et al. 2001). Consider the case in which just a single mutation is required for neofunctionalization. Provided $Ns^2 > 4$ and $u_0 < s^2$, the neofunctionalized allele will be maintained at the ancestral locus at equilibrium frequency $\hat{p}_n \simeq (s^2 - u_0)/s$ by balancing selection. Details are worked out in Lynch et al. (2001), and the overall consequences for homodimerization are shown in figure 4, but an especially simple limit for the probability of fixation of the homodimer can be obtained when $u_0 \ll s^2$, in which case the probability that the duplicate locus is randomly initiated from a neofunctionalized allele is $\hat{p}_n \simeq s$. As most individuals at the original locus will be AA, this neofunctionalized allele will experience the full selective advantage $s$, and with the probability of fixation at large $N$ being $\simeq 2s$ (the asymptotic value given by eq. 6), this path yields a probability of establishment of the homodimerized allele of $s \cdot 2s = 2s^2$. Alternatively, the new locus will be founded by the normal allele with probability

1360
The Evolution of Multimeric Protein Assemblages

DOI: 10.1093/molbev/msr300

MBE

F.

4.

doi:10.1093/molbev/msr300

MBE

Scaled probability of preservation by homodimerization of a newly arisen gene duplicate (relative to the neutral expectation of \(1/(2N)\)), derived from simulations of a Wright–Fisher population. The mutation rate for sites involved in the construction of dimerizing alleles follows the natural negative scaling with effective population size as described in the text, whereas the rate of mutation to defective gene copies is \(u_0 = 100u_1\). Complete linkage is assumed in all panels, but the upper right, where free recombination with the ancestral gene copy is assumed. Discontinuities in some of the curves result from the use of alternative approximations for different population size domains.

\[
\left(1 - s\right), \text{ and in this case a selective advantage is derived from the masking of null homozygotes at the original locus, whose frequency is } \sim s^2, \text{ leading to a probability of fixation of } 2s^2. \text{ Should the latter event occur, the ancestral locus will go to fixation for the already-present homodimerizing allele with probability close to 1.0, which is already established at a fairly high frequency, so the overall probability of establishment by this path is also } \sim 2s^2. \text{ Thus, for large populations, the rate of establishment of a homodimer via unlinked duplicates, } \sim 4NDs^2, \text{ is essentially independent of the mutation rate due to the fact that an ample supply of novel alleles is present at the outset.}

Finally, we consider the situation in which the evolution of a homodimer following duplication requires two mutations with the first-step allele being deleterious, the theory for which is presented and shown to agree reasonably well with simulations in the supplementary material (Supplementary Material online). In this case, the mean time to establishment scales inversely with \(1/\sqrt{\delta}\) to \(1/\delta\) depending on whether the population size is large or small (fig. 4), although the overall scaling with population size is not greatly different from that seen with neutral intermediates.

In summary, three gene duplication models for the origin of homodimers have been considered. First, in the simplest case of a single mutation with additive beneficial effects being required, when duplicates are completely linked, the time to establishment scales as \(2ND(u_0b)^{0.5} - 1\) in generations, implying a population size scaling of \(N^{-1.5}\) on an absolute time basis. If, however, such duplicates are unlinked, neofunctionalization can proceed without any mutational input in large populations that maintain neofunctionalizing alleles by selection–mutation balance, and the time to establishment scales as \(N^{-1.8}\) in generations and \(N^{-1.6}\) in absolute time units. Second, for the two-site model with neutral intermediates, when the duplicates are linked, the absolute time scaling varies from \(N^{-1.9}\) to \(N^{-1.35}\) for small versus large population sizes. Finally, for the two-site model with a deleterious intermediate state, the absolute time scaling for \(t_e\) is not much different from the neutral-intermediate case, ranging between \(N^{-1.6}\) and \(N^{-1.2}\) for small to large populations. Thus, we again conclude that the evolution of an advantageous homodimer is considerably more likely to evolve in populations with large effective sizes.

Homodimer to Heterodimer Transitions

Although heterodimers can, in principle, arise from promiscuous interactions among nonorthologous proteins, most seem to arise from interactions between paralogs arising from gene duplication, which will be the focus here. Homodimerization may precede gene duplication, providing a natural launching pad for heterodimerization following paralog divergence, or gene duplication may occur first, with
complexation of the paralogous products arising secondarily. Three general scenarios, not necessarily exclusive, are sketched out below.

Consider first the situation in which a locus encoding a homodimer harbors two alleles, such that the cross-product dimer created within heterozygotes elevates fitness beyond that for either of the two pure types produced in homozygotes. In this case, the two alleles will have been maintained in the ancestral (preduplication) population by balancing selection, with frequencies \( s_1/(s_1 + s_2) \) and \( s_2/(s_1 + s_2) \) for the \( a \) and \( A \) alleles under the assumption that the fitnesses of the \( AA, Aa, \) and \( aa \) genotypes at the ancestral locus are \( 1 - s_1, 1, \) and \( 1 - s_2, \) respectively, with \( 0 < s_1, s_2 < 1. \) As described above, gene duplication then provides an opportunity for each locus to fix an alternative allele, in which case every member of the population would have the expression pattern found in the ancestral heterozygote (fig. 5). Assuming the products of each locus randomly assemble at this early stage of duplicate-gene establishment, three types of dimers would be found within individuals in a 1:2:1 ratio (as is also true for heterozygotes in the ancestral single-locus state). Following the establishment of this complementing duplication state, subsequent mutational modifications at one or both loci might then lead to a pure heterodimer with the products of the individual loci no longer self assembling.

The selective advantage of a newly arisen duplicate under this model, derived in the supplementary material (Supplementary Material online), has a simple form. Provided the population size is large enough that the power of selection outweighs that of drift, the probability of fixation of the duplicate is

\[
\phi_{dup} \simeq \frac{2s_1s_2}{3(s_1 + s_2)},
\]

which reaches a maximum of \( s/3 \) when \( s_1 = s_2 = s. \) The expected time to transition (in generations) from a homodimer to a heterodimer under this model is then \( (2Nd\phi_{dup})^{-1}, \) which again implies a much shorter time to establishment in larger populations (\( \propto N^{-1.8} \) in absolute time units). This particular model is, of course, only relevant to diploid species, as haploids cannot harbor ancestral heterozygosity.

The remaining two scenarios by which a heterodimer may evolve are conceptually very similar to the subfunctionalization model of gene duplication (Force et al. 1999; Lynch and Force 2000; Lynch et al. 2001), with duplicate genes being reciprocally preserved when each copy loses a complementary essential subfunction (fig. 5). On the one hand, duplication of a locus already engaged in homodimerization may lead to a situation in which both loci acquire complementary mutations that together discourage homodimerization (enforcing the exclusive construction of heterodimers between the distinct monomers produced by the two loci). A mechanism similar to this has been suggested for the evolution of chaperonins (Ruano-Rubio and Fares 2007). Alternatively, an ancestral locus engaged in monomer production may become duplicated, in this case with subfunctionalization (or partial incapacitation) at the two loci resulting in heterodimer production. In both cases, the resultant heterodimeric structure may simply conserve the ancestral gene function, although it is possible that enhanced fitness may result from novel features associated with dimerization, either at the time of establishment or after the arrival of secondary advantageous mutations (Hughes 1994).

A key aspect of these subfunctionalization models is that if the fitness of individuals with complementary paralogs is no greater than that for the ancestral single-locus state, there is essentially no chance of joint preservation if \( Nu_0 > 1, \) where \( u_0 \) is the null mutation rate (Lynch and Force 2000; Lynch et al. 2001). There are two reasons for
this. First, because the mean time for an initially neutral allele to drift to fixation is $4N$ generations, at sufficiently large population sizes all descendants of such a duplicate locus will almost certainly be hit with a silencing mutation prior to fixation of the lineage. Second, a protein function that requires the products of two loci will have an elevated mutation rate to the null state equal to $\mu_Q$ relative to the single-locus case, which acts to maintain the single-locus state by positive selection. Thus, in contrast to the situation with fixation of adaptive ancestral heterozygosity, the latter two models predict that transitions from homodimers to heterodimers via gene duplication are much more likely in small than large populations if the driving force of duplicate gene preservation is complementary degenerative mutation.

**Discussion**

Heavy on theory and light on data, the previous results are offered as a starting point for discourse on the evolution of protein complexes. Although comparative biology provides a catalog of the historical products of biodiversification, the mechanisms by which complex cellular features evolve are constrained by fundamental principles of population genetics. If a proposed path to the acquisition of a specific molecular state by natural selection can be shown to be near impossible in common population-genetic environments, either the proposed molecular route of evolution is wrong, a constrained set of population-genetic conditions was involved, or the evolutionary transition to the novel state was driven by entirely nonadaptive forces. Consider, for example, the evolution of domain-swapping dimers, transitions to which are generally assumed to result from single deletion events. If the disadvantage of chimeric proteins in heterozygotes exceeds the power of random genetic drift, such changes cannot proceed to fixation in diploid populations, as the opposing gradient of natural selection would be overwhelming. Thus, the evolution of domain-swapping dimers appears to require either a haploid condition, or for diploidy, a situation in which the features of combined allelic products in heterozygotes do not cause deleterious effects of sufficient magnitude to offset the vagaries of random genetic drift.

For more complex adaptations requiring multiple mutations, it is often argued that adaptive evolution essentially never occurs via intermediate deleterious states. Such arguments derive from the assumption that evolutionary routes to adaptation involving deleterious early steps impose a bottleneck in mean population fitness. However, it is now clear that the rescue effect (stochastic tunneling) by secondary mutations provides a powerful mechanism for vaulting an adaptive valley while avoiding negative repercussions at the population level. With sufficiently low recombination rates, adaptive progress can be made without a population ever experiencing a decline in mean fitness because deleterious intermediate-state alleles never rise to high frequencies. The existence of such evolutionary pathways to adaptive exploitation raises significant challenges for laboratory studies that strive to reconstruct the historical order of events leading to the establishment of complex adaptations under the assumption that all intermediate steps must have been nondeleterious and procured in a stepwise fashion (Weinreich et al. 2006; Dean and Thornton 2007; Ortlund et al. 2007). Via the rescue effect, mutations that are individually deleterious not only can become fixed but do so simultaneously with the rescuing mutations, leading to episodic evolutionary change. A number of recent studies suggest that this sort of evolution is common with respect to the evolution of gene function (DePristo et al. 2007; Yokoyama et al. 2008; Field and Matz 2010; Carroll et al. 2011), and there is no reason to think that dimerizing interfaces cannot arise in a similar manner.

To close the loop between theory and observation on the evolution of protein complexes, substantial comparative work is needed on the features of orthologous proteins with monomeric versus multimeric forms but otherwise identical functions. To date, there appears to be no compelling evidence that multimeric proteins outperform their monomeric orthologs in other species, and some examples suggest otherwise. For example, the mismatch-repair system, which plays important roles in replication fidelity, DNA repair, and recombination is comprised of monomeric proteins in eubacteria but dimers in eukaryotes (Kunkel and Erie 2005; Iyer et al. 2006), yet the repair efficiency of eukaryotic systems appears to be lower than that in prokaryotes (Lynch 2011). As another example, the sliding clamps used in DNA replication are homodimeric in eubacteria but homotrimeric in eukaryotes, with both structures having very similar overall architecture (Kelman and O’Donnell 1995), yet replication-fork progression rates are nearly an order of magnitude faster in prokaryotes (Lynch 2007). In addition, although the ribosome has a much more complex protein repertoire in eukaryotes than in prokaryotes (Smith et al. 2008), there is no evidence that translation fidelity is elevated in the former, and although the data are limited, a number of observations suggest the opposite (e.g., Loftfield and Vanderjagt 1972; Buchanan et al. 1980; Parker 1989; Salas-Marco and Bedwell 2005; Kramer and Farabaugh 2007).

Many more comparative studies will be required to determine if shifts to higher-order complexes are typically unaccompanied by significant enhancement in functionality, but continued evidence of this nature would further support the idea that multimeric proteins often simply arise as compensatory responses to accumulated defects in monomers in species experiencing relatively small effective population sizes (Fernández and Lynch 2011). For the time-being, the preceding theory at least provides a plausible set of mechanistic explanations for the patterns in figure 1. Consistent with the approximately equal apportionment of monomers in prokaryotes and eukaryotes (fig. 1), the theory suggests that homodimers can evolve at least as readily in large as in small populations under a wide variety of conditions. On the other hand, the data in figure 1 suggest that once established, homodimers make transitions to heterodimers much more readily in small than in large populations. This pattern appears to be most...
consistent with a scenario in which such structures arise largely as a consequence of the accumulation of complementary degenerative mutations in duplicated genes.

These observations raise the possibility that, in the case of gene structure and genomic architecture (Lynch 2007), variation in the power of random genetic drift among phylogenetic lineages has contributed significantly to the emergence of many of the complex (and often arcane) features of eukaryotic cells. Nonadaptive arguments for the origin of cellular infrastructure have been made before. For example, invoking a process called constructive neutral evolution, Stoltzfus (1999), Gray et al. (2010), and Lukes et al. (2011) suggested that the establishment of molecular machines such as the ribosome and the spliceosome may have emerged via the fortuitous, neutral establishment of interactions among lower-level protein components, which in turn suppressed the effects of subsequent mutations that would have otherwise inactivated the individual parts. Under this model, after the establishment of such degenerative mutations, the components of such a complex would then be mutually interdependent. In a related exercise, Frank (2007) argued that any evolutionary step that leads to increased robustness of a cellular function will also magnify the likelihood that the underlying component parts will acquire mutational defects, again leading to the growth of complexity at the expense of the previously autonomous parts.

Although the latter arguments provide potentially plausible paths for the nonadaptive evolution of complex structures, key aspects of the evolutionary dynamics required to arrive at the postulated end points remain to be explored. In both cases, scenarios of stepwise fixation appear to have been assumed, which as noted above are unlikely to be realized except in very small populations (perhaps too small to avoid extinction). Also lacking is attention to the problematical issues of segregation and recombination in diploid populations, which can substantially alter the fates of pairs of mutations, especially when carried in different genes. Most notable, however, is the need to account for the fact that essentially all added layers of complexity impose a mutational burden relative to simpler structures that carry out the same function but comprise smaller targets for inactivating mutations. This differential vulnerability to mutation, which acts like a weak form of selection against added ( gratuitous) complexity, may be the key to explaining why so many higher-order structures that have evolved in eukaryotes maintain simpler forms in prokaryotes (Lynch 2007).

Supplementary Material

Supplementary Material is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work has been supported by the National Institute of Health (R01 GM036827 to M.L.) and W. K. Thomas, National Science Foundation (EF-0827411 to M.L.), and US Department of Defense (ONRBA 10-002 to M.L., P. Foster, H. Tang, and S. Finkel).

References


