Published in final edited form as:

J Mol Evol. 2010 January ; 70(1): 29–43. doi:10.1007/s00239-009-9305-7.

The Origin of Conserved Protein Domains and Amino Acid Repeats Via Adaptive Competition for Control Over Amino Acid Residues

Mary M. Rorick and Günter P. Wagner

Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520-8106, USA

Abstract

Some proteins, such as homeodomain transcription factors, contain highly conserved regions of sequence. It has recently been suggested that multiple functional domains overlap in the homeodomain, together explaining this high conservation. However, the question remains why so many functional domains cluster together in one relatively small and constrained region of the protein. Here we have modeled an evolutionary mechanism that can produce this kind of clustering: *conserved functional domains are displaced from the parts of the molecule that are undergoing adaptive evolution because novel functions generally out-compete conserved functions for control over the identity of amino acid residues.* We call this model COAA, for Competition Over Amino Acids.

We also studied the evolution of amino acid repeats (a.k.a. homopeptides), which are especially prevalent in transcription factors. Repeats that are encoded by non-homogenous mixtures of synonymous codons cannot be explained by replication slippage alone. Our model provides two explanations for their origin, maintenance, and over-representation in highly conserved proteins. *We demonstrate that either competition between multiple functional domains for space within a sequence, or reuse of a sequence for many functions over time, can cause the evolution of amino acid repeats.* Both of these processes are characteristic of multifunctional proteins such as homeodomain transcription factors. We conclude that the COAA model can explain two widely recognized features of transcription factor proteins: conserved domains and a tendency to accumulate homopeptides.

Keywords

homeodomain; conserved protein domain; transcription factor evolution; homopeptide; amino acid repeats

I. Introduction

The National Center for Biotechnology Information's Conserved Domain Database contains over 25,000 position-specific score matrices, each of which represents the conserved characteristics of an ancient protein domain. This massive universe of protein diversity can be reduced to variations on only a few thousand domain superfamilies, the members of which are compact structural units related by descent (Marchler-Bauer et al. 2003; Marchler-Bauer et al. 2007). The high conservation of some domain superfamilies requires an

explanation. The simplest of these is that purifying selection to maintain an essential function removes all random variations, but the level of conservation often far exceeds that which can be explained by any single function. For example, there is no variation within 58 consecutive residues of the HoxA11 sequence among chicken (*Gallus gallus*), zebrafish (*Danio rerio*), coelocanth (*Latimeria menadoensis*), mouse (*Mus musculus*), and shark (*Heterodontus francisci*). Though this conservation has been attributed to the domain's DNA-binding function (with potentially a large number of target sequences (Wagner and Pyle 2007)), there are actually only six residues in the homeodomain that contact DNA bases, and only four additional ones that contact the phosphate backbone (Ledneva et al. 2001). Moreover, although different homeodomain-containing proteins bind DNA in the same manner, their sequence conservation is paralog-specific (e.g., Lynch et al. 2006), since structural constraints do not generally impose significant sequence constraints (Babajide et al. 2001). Thus, an alternative hypothesis for the evolution and maintenance of highly conserved regions is that multiple functional domains overlap and together constrain the identity of the residues (Roth et al. 2005). However, this does not explain why so many conserved functional domains would restrict themselves to one small, highly constrained region of a much larger protein—and this is what we seek to address.

The model we propose is inspired by experimental evidence suggesting that homeodomains are sites of competition between functions. In mouse HoxA11 the conserved repressor function is restricted to 12 amino acids within the homeodomain, but in both zebrafish and chicken it is also located in a region N-terminal to the homeodomain (Roth et al. 2005). In this region there are also prominent signals of adaptive evolution at the stem lineage of placental mammals (Lynch et al. 2008)—that is, coinciding with the repressor domain's movement all the way into the homeodomain and suggesting that adaptation of novel domains in this region may have pushed the ancient repressor domain all the way into the conserved homeodomain. Because our model relies on competition between different functional domains over amino acid identity, we call it COAA, for Competition Over Amino Acids.

Although they are seemingly unrelated to highly conserved multifunctional domains, the COAA model also addresses the existence of conserved amino acid repeats. Also known as simple sequence repeats (SSRs) or homopeptides, amino acid repeats are homogenous stretches of a single amino acid that are anywhere from a few to over fifty amino acids long (Faux et al. 2005). For a protein that is 400 amino acids long and of typical composition, any run that is five amino acids or longer is statistically significant ($P=0.001$) (Karlin 1995). Repeats are a common feature of eukaryotic proteins, and they are longer and occur more frequently than can be explained by random chance (Green and Wang 1994; Karlin et al. 2002; Li et al. 2004; Faux et al. 2005; Hancock and Simon 2005). In humans for example, 18–20% of proteins contain one or more such runs (Alba and Guigo 2004; Karlin et al. 2002). An evolutionary explanation for their existence requires that we distinguish between two major classes of repeats for independent investigation, since they likely have distinct evolutionary origins (Hancock et al. 2001; Hancock and Simon 2005; Alba et al. 1999a).

The first class consists of the repeats that are referred to as transposable element (TE)-like sequences, microsatellites or trinucleotide repeats (Richard and Dujon 1997). These repeats are homogeneously encoded—meaning that they are perfect sequence repeats at the DNA level as well as the amino acid level. It is thought that these repeats originate primarily through a replication slippage mutational process that causes them to undergo evolutionary expansions (Hancock et al. 2001; Levinson and Gutman 1987; Alba et al. 1999b, 2001). Because they are unstable, these repeats are highly variable, preferentially found in regions of the proteome under weak purifying selection (Mularoni et al. 2007), and associated with

deleterious phenotypes (Mularoni et al. 2007; Hancock et al. 2001; Cummings and Zoghbi 2000; Karlin et al. 2002).

The second class of repeats is common in human and mouse proteomes, and is comprised of “mixed codon” repeats (Hancock and Simon 2005). Unlike homogenously encoded repeats, these are slow-evolving and predominantly located in highly conserved proteins (Alba et al. 1999a; Hancock and Simon 2005; Hancock et al. 2001; Mularoni et al. 2007), as well as over-represented in the most constrained proteins (Mularoni et al. 2007). Given that slippage is not likely to occur in the absence of trinucleotide repeats at the DNA level (Levinson and Gutman 1987; Petes et al. 1997; Alba et al. 1999a), another explanation is required for the origin and maintenance of this class of repeats. Mixed codon repeats are also of special interest because it seems they specifically characterize the proteomes of complex organisms (e.g., while they are common in mammals, they do not exist in yeast, fly, or prokaryotes (Hancock and Simon 2005)). Two evolutionary mechanisms underlying the COAA model can explain the origin and maintenance of these conserved, mixed codon repeats.

The proposed COAA model can produce both (1) clustering of conserved functional domains into a relatively small, highly constrained multifunctional segment of the protein; and (2) conserved amino acid repeats in the absence of a replication slippage mechanism. This model also provides an explanation for the fact that highly conserved domains and amino acid repeats are both found preferentially in developmental and regulatory proteins (Karlin et al. 2002; Alba and Guigo 2004; Faux et al. 2005; Richard and Dujon 1997; Nakachi et al. 1997). Regulatory proteins are highly conserved, contain regions under adaptive evolution, and are the sites of novel functional innovation (Lynch et al. 2008), and these features are what drive the mechanisms described by our model.

II. Model

A protein is represented as a sequence of R amino acid residues. There are 20 different types of amino acids, designated by numbers 1 through 20. The protein can perform multiple distinct functions, each of which is enabled by a unique subset of amino acids. These subsets are termed “alphabets”, and the amino acids within them are called the “alphabet amino acids” of a function. In order for a protein to perform a function the protein must have a functional domain within its sequence. In this model, a functional domain is defined as a contiguous series of alphabet amino acids that is at least of minimum length m . The performance of function i (p_i) increases with the number of residues in its functional domain (b_i), according to the following equation.:

$$p_i = 1 - e^{(-c(b_i - (m-1)))} \text{ if } b_i \geq m, \text{ and } = 0 \text{ otherwise,} \quad (1)$$

where c is an arbitrary constant that determines how strongly functional performance depends on the size of the functional domain (throughout this paper $c=0.2$) (Fig. 1A). The functional performance is zero if the longest contiguous segment of alphabet amino acids is less than m . As b_i increases, the performance measure increases and converges to a maximum of one for large functional domains. This is a typical “decreasing returns” model in that the addition of a residue to a small domain affects functional performance more than the addition of a residue to a large domain. This means there is negative epistasis between residues within domains.

We think this is a reasonable way to define and distinguish protein functions because protein interaction sites are largely determined by local amino acid composition (Ofra and Rost 2003). We assume that all functions are essential and equally important, so the fitness of the entire protein, w , is determined by:

$$w = \prod_{i=1}^F p_i, \quad (2)$$

for all F functions that the protein performs. Note that the fitness of a protein is zero if any of its functions have zero performance, and that because the above equation is multiplicative, there is no epistasis between functions.

A population with effective population size N_e is modeled as a single representative wild type protein sequence that changes through time. We model mutation and selection by allowing a single viable ($w_{mut} > 0$) point mutation each time-step (each time-step thus represents the average waiting time between mutations). The fitness of the mutant, which has a functional domain of length h , is compared to the fitness of the wild type, which has a functional domain of length g , to obtain a selection coefficient

$$s = (w_h - w_g) / w_g, \quad (3)$$

which, along with the effective population size, determines a fixation probability f for the mutant according to Hill's approximation (for a diploid population with additivity among genes and in which the effective population size is scored each generation) (1982):

$$\text{if } N_e s < -1, \quad f = 0, \quad (4)$$

$$\text{if } -1 < N_e s < 1, \quad f = 1 / (2 N_e) + s / 2, \quad (5)$$

$$\text{if } N_e s > 1, \quad f = s. \quad (6)$$

III. Simulations

In simulations of this model, we distinguish between two types of functions. “Primary functions” are those which remain invariant through time, under all possible environmental conditions. Selection for primary functions is thus a case of internal selection (Wagner and Schwenk 2000). For instance, transcription factors have several primary functions—such as DNA binding, nuclear localization and transcriptional activity—that are unaffected by their environmental or regulatory context. In contrast, “secondary functions” are dependent on environmental conditions that change regularly through time. Hence secondary functions have short average lifetimes.

To keep the model simple we assume that all functional domains have the same minimum length of m residues. Each of the primary functions is defined by its own unique alphabet (which is a list of numbers, because each amino acid is represented by a unique number). The primary functions' alphabets are invariable, and intersect to some degree. Function 1 and 2 are examples of primary functions with alphabets that intersect at one amino acid (i.e., 4):

Function 1 alphabet: {1,2,3,4}

Function 2 alphabet: {4,5,6,7}

Function 3 alphabet (initial): {8,9,10,11,12,13,14,15}.

A secondary function (e.g., Function 3 above) is defined by an alphabet that changes with environmental change events, and that is constrained to never intersect with primary function alphabets (to encourage competition between primary and secondary functions). Secondary functions were generally assigned larger alphabets than primary functions to ensure a high likelihood of viable functional domains ($b_i > m$) despite frequent environmental changes. The number of residues used to calculate function performance, b_i , is defined as the length of the longest functional domain for function i .

With each time-step, a viable ($w_{mut} > 0$) point mutation enters the population, but is not necessarily fixed (i.e., it does not necessarily become the wild type sequence for the next time-step). The fixation probability is determined by Eqs. 4–6. If the mutation is not fixed, the wild type sequence remains the same, and we proceed to the next time-step.

Environmental change events occur at intervals of r time-steps throughout the simulation. After a mutation is generated, and either fixed or not, the environmental change event occurs. This consists of randomly choosing one of the secondary functions (when there are multiple), and then generating a new random alphabet for this function. The new alphabet must consist of a certain number of unique amino acids that are not primary function alphabet amino acids. Once a suitable alphabet is generated, there is a search within the sequence for the longest functional domain. If a viable functional domain exists ($b_i > m$), the new alphabet for the secondary function is accepted, and the secondary function's performance and the protein's overall fitness are reassessed. In the rare case that the new alphabet does not permit a viable functional domain, and would thus result in a fitness of 0 (i.e., $b_i < m$), the alphabet selection process is repeated until viability is achieved (one exception is noted below).

Simulations were performed under a variety of parameters and initial conditions (see Appendix 2 in Supplementary material for details), and included cases with: a single primary function, two primary functions in the presence of a single secondary function, multiple primary and secondary functions, multiple primary functions, and a single secondary function. The locations and lengths of all functional domains were monitored, as well as the amino acid composition of the sequence.

IV. Theory: the evolution of a single primary function

Consider a single domain and how it changes over time without interaction with other domains. Mutation and selection will add and subtract amino acid residues from the domain at the edges, more or less one at a time. (This is because losing an “internal” residue breaks up the domain and causes a large reduction in fitness, and mutations that increase the length by more than one residue are rare.) Adding and subtracting residues at the edges of the domain cause changes in the length of the domain, which will eventually fluctuate around an equilibrium length. Furthermore, if a residue is lost on one side of the domain, and another is gained on the opposite side, the domain will undergo a random walk while remaining the same size on average. The evolution of a domain, as long as it is not interacting with other domains, can thus be analyzed as two complementary processes: changes in domain size and changes in domain location. Formalizing this into a discrete stochastic Markov model, we can predict the equilibrium length and random walk rate of a domain, and the probability of a “touch-event” between two domains (the event that two primary domains will run into each other). (See Appendix 1 in Supplementary material for details.)

V. Results

The evolution of a single primary function

The analytical model summarized above is in qualitative agreement with simulations of a single primary domain reaching equilibrium length and undergoing a random walk along the length of the peptide, however random walk rates observed in simulations are somewhat lower than analytically predicted rates (Figs. 2 and 3). The rate of random walk determines the probability that a touch-event between two primary domains will occur within a certain timeframe, and analytical predictions are also reasonably consistent with simulations in this regard. See Appendix 6 in Supplementary material for a detailed discussion.

Analysis of simulations with primary and secondary functions

General patterns—In simulations of two primary functions with a single competing secondary function, we find that the ever-changing secondary functional domain appears to encroach on the space initially occupied by the two primary functional domains (Fig. 4). Further, once the primary functional domains touch, their frequency of “sticking” (see Appendix 3 in Supplementary material) is 95%, which is significantly greater than 50% (a frequency 95% would only occur with a probability of 5×10^{-6} , given the null hypothesis of 50% and assuming a binomial distribution ($N=22$)). The primary functional domains then go on to stabilize at a fully overlapped state (Fig. 4, Table I). The above qualitative pattern is observed in simulations with many different alternative initial conditions and parameter choices (see Appendix 2 and 5 in Supplementary material). Hence, we observe the spontaneous origin of clusters of overlapping primary functional domains in the presence of variable secondary functional domains. We take this as a possible mode for the origin of highly conserved multifunctional domains in proteins.

Secondary functions out-compete primary functions for control of residue identity—An important consequence of our definition of functional performance (Eq. 1) is that the fitness contribution from adding an individual amino acid to a given domain diminishes with increasing domain length (Fig. 1A). Therefore, when neighboring functional domains compete for control over the identity of flanking amino acid residues, selection will favor the shorter domain because transferring the contribution of a residue from a long domain to a short domain offers a net fitness gain. Specifically, to the extent that novel functional domains are less well optimized (i.e., shorter) than ancient, conserved functional domains, we would expect them to encroach on flanking ancient domains until they are of equal length.

Therefore, when primary and secondary functions compete for control over residue identity, we predict that secondary functions generally win because they will generally have shorter domains, and thus, lower performance than primary functions (Fig. 1B). To test this prediction, we ran simulations with an equal number of primary and secondary functions, with equally sized, non-overlapping alphabets (see Appendix 2 in Supplementary material for details). We measured the frequency at which a secondary function’s performance increases at the expense of a primary function’s performance, and compared this to the frequency of the reverse occurrence. Consistent with our predictions, we found that the former occurs more frequently than the latter (Fig. 5), indicating that secondary functions do tend to out-compete primary functions for control over residue identity.

Primary functional domains evolve overlap—Another important consequence of Eqs. 1 and 2 is that, all else being equal, if a residue is a part of two functional domains, it contributes more to the protein’s fitness than if it were only a part of a single functional domain. This means that regions of overlap between functional domains enjoy added

protection from encroachment by competing functional domains. Therefore, we predict that primary functional domains will progressively increase their overlap with one another as a simulation progresses. Typical simulation results are consistent with this prediction: primary functional domains do tend to evolve overlap (Fig. 4, Table I). However, to be sure that this overlap evolution is actually driven by the encroachment of the flanking secondary functional domains, we must demonstrate that any alternative causes for the evolution of overlap are either unnecessary or insufficient. Indeed, there are several potential forces driving overlap evolution in our simulations.

The possible causes for the evolution of overlap—The first possibility is that the primary functional domains evolve overlap because they are simply still growing in length. We control for this by placing primary domains sufficiently far apart at the beginning of the simulation that they reach equilibrium length well before they touch. A second possibility is that overlap may provide a fitness advantage by reducing the total number of mutable sites, or by causing deleterious mutations to be more severe and easier to select against (since mutations can harm two functions at once). In this case, overlap might evolve because mutational degradation would be faster on the non-overlapped end compared with the overlapped end of a functional domain. (To allow a net increase in overlap, this slower degradation rate would have to more than compensate for the greater constraints imposed on growth at the overlapped end.) While we cannot remove the effects of this second mechanism, we can hold it constant (see below). The third possibility—our hypothesized cause for the evolution of overlap—is the competitive degradation of primary domains by secondary domains preferentially on their non-overlapped ends. (Like the second possibility, this requires that protection from encroachment more than compensate for the overlapped end's slower growth rate.) Only this third mechanism features a role for secondary functions in the evolution of overlap.

The above mechanisms are not mutually exclusive, and it is likely that the first two are sufficient to produce the evolution of overlap to some extent. What we want to understand is whether the third mechanism, which uniquely involves secondary functions, significantly enhances the evolution of overlap. If it does not, a theory for the origin of highly conserved regions would not need to invoke adaptive competition as a key component.

Testing possible mechanisms for the evolution of overlap between domains—

If secondary functions play a significant role in the evolution of overlap between primary functional domains, we should expect to see their presence enhance the evolution of overlap between primary functional domains. We find that secondary functional domains have many significant effects that are consistent with this prediction: (1) Their presence increases the average rate of random walk of a primary functional domain before and after overlap initiation (Fig. 6A and 6B, Table I). (2) The frequency at which two primary functional domains touch within the first 400,000 generations is increased from 37.5% (where $N=24$) to 100% (where $N=20$) ($P<.0001$, based on χ^2 contingency test) (Table I). (3) The average rate at which the region of overlap between the two primary functional domains lengthens is increased (Fig. 6C, Table I). (4) The frequency of primary functional domains “sticking” is increased (from 12 out of 18 to 21 out of 22 touch-events in independent simulations, $P<.018$ based on χ^2 contingency test). (5) While overlap does provide some fitness advantages even in the absence of secondary functional domains, the fitness advantages of overlap become greater in the presence of secondary functional domains. Specifically, in the absence of secondary functional domains, the average primary functional domain increases in length to some extent once overlap is achieved (from 23.12 to 23.45), but in the presence of secondary functional domains, this overlap-induced domain length increase is greater (from 19.72 to 21.45) (Table I). The overlap-induced reduction in the average mutation effect is also more significant in the presence of secondary functional domains. In their absence, the

average $N_e s$ goes from -2.78 to -2.71 , whereas in their presence, $N_e s$ goes from -4.62 to -4.21 with overlap. Overlap is also associated with a significant reduction in deleterious mutation rate only when secondary functional domains are present (Table I). (In the presence of secondary functional domains, overlap initiation reduces the deleterious mutation rate from .457 to .403 (with $N=15$ and 16, respectively, and $P<.001$ based on Student's t test), whereas in their absence, overlap only reduces the deleterious mutation rate from .437 to .427 (with $N=27$ and 15, respectively, and one-sided $P=.26$ based on Student's t test).) (6) The presence of secondary functional domains decreases the average length of primary functional domains, both before and after the primary domains touch (from 23.17 to 19.56 before they touch ($N=31$ for simulations with primary domains only and 15 for simulations with secondary domains) and from 23.44 to 21.58 after they touch ($N=18$ for simulations with primary domains only and 16 for simulations with secondary domains) $P<.0001$ for both, based on Student's t test), indicating that secondary functional domains increase the mutational decay rate of primary functional domains (Table I). This effect is less pronounced after overlap is achieved, presumably because overlap partially protects the primary functional domains from encroachment by the secondary functional domains. In summary, this analysis indicates that adaptive competition plays a key role in the emergence of highly conserved, multifunctional domains flanked by sequence regions undergoing adaptive change.

The Evolution of Amino Acid Repeats

In addition to the pattern described above, we observed another, completely unexpected pattern in our simulations. Given enough time-steps, regions of homogenous amino acid composition consistently emerged in the simulations even though our model does not allow for insertions, deletions, or duplications of sequences (Fig. 7, Table I). We find this pattern intriguing because transcription factor proteins are known to be prone to the evolution of amino acid repeats (Mularoni et al. 2007). We considered two potential mechanisms for the evolution of repeats in our simulations.

First, because overlapping regions have more limited alphabets than individual functions (overlapping regions effectively have alphabets equal to the intersection of the overlapping functions' alphabets), it is possible that selection for overlap drives the evolution of amino acid repeats. A second possibility is that the origin of repeats depends on the secondary functions. A sequence region of relative homogeneity will be prone to seeding novel functional domains, which means that, given repeated cycles of environmental change events, it will be prone to undergoing selection for further enrichment of the over-represented amino acid (since any function that uses a homogenous region as a domain must have an alphabet that includes the over-represented amino acid). Thus, local amino acid homogeneity may be self-reinforcing (Fig. 8).

We tested the sufficiency of both mechanisms (see Appendix 2 in Supplementary material). The first mechanism can be isolated from the second by simulating the evolution of sequences that contain several primary functional domains evolving under crowded conditions (in the absence of secondary functions) (Fig. 9). If we define average repeat length as the average length of any homogenous stretch of amino acid types of two amino acids or more, we find that there are significantly longer repeats ($\mu=5.84$, $\sigma=3.70$, $N=6$) in the resulting evolved sequences than in random, non-evolved sequences with no overlap that are comparable (i.e., that are of equal length but composed entirely of one function's alphabet amino acids) ($\mu=2.34$, $\sigma=0.17$, $N=100$, $P<0.001$, based on Student's t test).

The second mechanism can be isolated from the first mechanism by simulating the evolution of a sequence with a single secondary function (and to eliminate the risk of biasing alphabet selection to over-represented amino acids, in this case we allow a fitness of zero) (Fig. 10).

Even though there is only a single functional domain at any one time, because it repeatedly changes identity and location, we cannot rule out the possibility that previous domains have left “footprints” of amino acid homogeneity. Therefore, we are conservative here in our choice of null distribution: each non-evolved sequence in the distribution is composed entirely of alphabet amino acids. Despite this conservative null distribution, we find that the average length of repeats in the evolved sequences is greater than the average length of repeats in the null distribution (after 400000 time-steps of evolution with N_e of 10000, simulations have $\mu=10.86$, $\sigma=2.08$ ($N=4$), whereas non-evolved null-distribution sequences have $\mu=2.34$, $\sigma=0.17$ ($N=1000$), and $P<.0001$ based on Student's t test). This demonstrates that there is a kind of selection for “pre-adapted” sequences in that regions of higher-than-average amino acid homogeneity have an increased chance of functionality, and therefore, an increased chance to experience selection for further enrichment of the already over-represented amino acids. In summary, the evolution of overlap and the preferential enrichment of already over-represented amino acids via shifting directional selection are both independently sufficient mechanisms for the origin and maintenance of amino acid repeats in our model.

VI. Discussion

The evolution of highly conserved domains

Some proteins, such as homeodomain transcription factors, contain highly conserved regions of sequence. Multiple conserved functional domains may overlap and together explain the high conservation of such regions (Roth et al. 2005; Lynch et al. 2006). Here we present a model that can explain why so many conserved functions cluster together in one relatively small region: Novel functions out-compete conserved functions for control over the identity of amino acid residues (Fig. 1B). Simulations based on our model successfully exhibit the evolution of highly conserved domains surrounded by regions undergoing adaptive change, and further analysis confirms that the underlying dynamics of the simulations are indeed consistent with the hypothesized mechanism.

The evolution of amino acid tandem repeats

Through simulations based on our model we also observed and analyzed the evolution of amino acid tandem repeats. Although there is significant evidence that some of these repeats play specific functional roles (Mularoni et al. 2007; Fondon and Garner 2004; Alba et al. 2002; Han and Manley 1993; Faux et al. 2005), it has been generally assumed that they are neutral structures found in contexts of low purifying selection and a consequence of replication slippage (Lovell 2003; Hancock et al. 2001; Mularoni et al. 2007; Levinson and Gutman 1987; Alba et al. 1999b; Alba et al. 2001). However, homopeptides that are not encoded by perfect trinucleotide repeats must be at least maintained through selection for amino acid homogeneity. These include mixed codon repeats (Hancock et al. 2001; Alba et al. 1999a), “cryptic repeats” (imperfect amino acid tandem repeats that contain statistically significant overrepresentations of a single amino acid or a short amino acid motif) (Hancock and Simon 2005; Alba et al. 2002), and repeats comprised of non-identical, yet biochemically related families of amino acids (Karlin et al. 2002). Presumably, these kinds of repeats are either old enough to have diverged at their neutral sites or the result of nucleotide substitutions rather than trinucleotide duplication (slippage).

Other evidence that selection is at work in the evolution of repeats includes that transcription factors, other developmental and regulatory proteins, proteins that function as part of large complexes, protein kinases, signaling proteins and membrane transporters are all significantly overrepresented among proteins that contain repeats (Alba et al. 1999b; Alba et al. 2002; Li et al. 2004; Faux et al. 2005). Also, repeats are not evenly distributed between

UTRs and coding regions (Li et al. 2004), their positioning is biased toward specific reading frames (Hancock et al. 2001), and their amino acid composition depends on the particular type of protein in which they are found (Alba et al. 2002). Furthermore, compared to poorly conserved repeats (that are presumably under weak selection), repeats that are highly conserved in orthologous genes are encoded by more heterogeneous mixtures of synonymous codons (Alba and Guigo 2004; Alba et al. 1999a). In addition, while highly conserved repeats and highly variable repeats are both biased toward specific sets of amino acids, these biases differ markedly (Mularoni et al. 2007; Hancock and Simon 2005). Also, in some reconstructed evolutionary histories there is evidence of repeats evolving gradually via point mutations or recombination events rather than by trinucleotide duplication (Li et al. 2004; Crow et al. 2009). Fig. 11 shows an example of the gradual evolution of an alanine repeat by substitution in the zebrafish HoxA13a paralog (data from Crow et al. 2009). Lastly, Mularoni et al. (2007) finds a positive correlation between protein selective constraint and the number of repeats the protein contains (that is, the most constrained proteins have the greatest number of repeats), and these repeats turn out to be of the highly conserved variety. (The authors did not have an explanation for this unexpected observation, and stated that “selection may play a more important role than previously suspected in the preservation of repeats”.) All of the above supports a role for selection in the origin and maintenance of non-homogenously encoded repeats.

The COAA model provides two explanations for how positive selection may play a role in repeat evolution, and in turn, the prevalence of repeats in the most selectively constrained proteins that have many competing functional demands. We demonstrate that either competition between multiple functional domains for space within a sequence, or reuse of a sequence for many functions over time, can cause the evolution of such repeats. We would expect to find both of these mechanisms at work in the context of multifunctional, highly constrained proteins, such as homeodomain-containing transcription factors (e.g., Lynch et al. 2006) (see Appendix 7 in Supplementary material for a discussion about which mechanism might dominate in real proteins).

Furthermore, previous sequence data analysis on repeats supports both mechanisms for the evolution of repeats modeled here. First, supporting the correlation between repeat evolution and protein domain multifunctionality, the majority of repeat containing proteins are involved in processes that require the assembly of large, multiprotein or protein-nucleic acid complexes (Faux et al. 2005), which suggests that they have multiple interaction sites (i.e., multiple functional domains). Second, prokaryote proteins have far fewer repeats than eukaryotic homologues (Faux et al. 2005; Karlin et al. 2002; Li et al. 2004), which indicates that the evolution of repeats occurred after the evolution of the initial protein structure and perhaps concomitant with the evolution of novel functionalities that may have occurred to a greater extent in eukaryotic proteins.

Directions for future work

An important future direction will be to analyze sequence data for patterns predicted by the COAA model. For example, we expect that population size should negatively correlate with the rate of random walk (Fig. 3) and, in contrast, positively correlate with functional domain growth rate and size (Fig. 2). Thus, in very large populations we might expect functional domains to take a long time to come into contact, but to evolve overlap relatively quickly once this occurs. Similarly, based on our simulation results (data not shown), we expect that repeat evolution— at least via secondary function adaptation— should positively correlate with population size. We can test the above predictions by comparing proteins from species that sustain very different population sizes (e.g., similar to what has been done for genome architecture evolution (Lynch 2007)). This type of analysis may uncover a threshold population size that is necessary for the dynamics of this model to take hold.

We can also test the validity of our model by observing whether amino acid repeats and conserved domains show evidence of competition between functions. We already know this is the case for HoxA11, mentioned above (Lynch et al. 2008). However, it is possible to identify regions with competing functions across large datasets of proteins if we use efficient statistical methods. One such method, developed by Gu (2006), measures clade-specific, conserved (“Type-II”) functional divergence. It identifies “within-clade” amino acid conservation that is accompanied by “between-clade” amino acid divergence. When the stem lineage shares the amino acid identity of one of the daughter clades (as it often does), it can be concluded that competition between two alternative functions has occurred during the evolution of the divergent lineage. Using this test, our lab has found evidence for competing functions among the conserved and repeat-containing zebrafish HoxA13 paralogs at 23 alignment sites (Fig. 12) (Crow et al. 2009). While the HoxA11 and HoxA13 anecdotes suggest that competition for control over amino acids by different functions is a common mode of evolution among transcription factor proteins, it remains to be evaluated whether similar evidence exists across many different classes of repeat-containing and conserved proteins.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are tremendously grateful to JME Associate Editor, Mark Jensen, for his thorough and insightful review, and for the useful comments and suggestions of two anonymous reviewers. The experimental work in the Wagner lab is supported by a grant from the John Templeton Foundation (Grant number 12793). The views expressed in this paper do not necessarily reflect the views of the JTF.

References

- Alba MM, Guigo R. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.* 2004; 14:549–554. [PubMed: 15059995]
- Alba MM, Santibanez-Koref MF, Hancock JM. Conservation of polyglutamine tract size between mouse and human depends on codon interruption. *Mol Biol Evol.* 1999a; 16:1641–1644. [PubMed: 10555295]
- Alba MM, Santibanez-Koref MF, Hancock JM. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of slippage-like mutational process. *J Mol Evol.* 1999b; 49:789–797.
- Alba MM, Santibanez-Koref MF, Hancock JM. The comparative genomics of polyglutamine repeats: extreme differences in the codon organization or repeat-encoding regions between mammals and *Drosophila*. *J Mol Evol.* 2001; 52:249–259. [PubMed: 11428462]
- Alba MM, Laskowski RA, Hancock JM. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics.* 2002; 18(5):672–678. [PubMed: 12050063]
- Babajide A, Farber R, Hofacker IL, Inman J, Lapedes AS, Stadler PF. Exploring protein sequence space using knowledge-based potentials. *J Theor Biol.* 2001; 212:35–46. [PubMed: 11527443]
- Crow K, Amemiya CT, Roth J, Wagner GP. Hypermutability of HoxA13 and functional divergence from its paralog are associated with the origin of a novel developmental feature in zebrafish and related taxa (Cypriniformes). *Evolution.* 2009; 63(6):1574–1592. [PubMed: 19222565]
- Cummings CJ, Zoghbi HY. Trinucleotide repeats: Mechanisms and pathophysiology. *Annu Rev of Genom Hum Genet.* 2000; 1:281–328.
- Faux NG, Bottomley SP, Lesk AM. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res.* 2005; 15:537–551. [PubMed: 15805494]
- Fondon JW, Garner HR. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci Unit States Am.* 2004; 101(52):18058–18063.

- Green H, Wang N. Codon reiteration and the evolution of proteins. *Proc Natl Acad Sci Unit States Am.* 1994; 91:4298–4302.
- Gu X. A simple statistical method for estimating Type-II (Cluster-Specific) functional divergence of protein sequences. *Mol Biol Evol.* 2006; 23(10):1937–1945. [PubMed: 16864604]
- Han K, Manley JL. Functional domains of the *Drosophila* Engrailed protein. *EMBO J.* 1993; 12:2723–2733. [PubMed: 8334991]
- Hancock JM, Simon M. Simple sequence repeats in proteins and their significance for network evolution. *Gene.* 2005; 345:113–118. [PubMed: 15716087]
- Hancock JM, Wothe EA, Santibanez-Koref MF. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol Biol Evol.* 2001; 18:1014–1023. [PubMed: 11371590]
- Hill WG. Rates of change in quantitative traits from fixation of new mutations. *Proc Natl Acad Sci Unit States Am.* 1982; 79:142–145.
- Karlin S. *Curr Opin Struct Biol.* 1995; 5:360–371. [PubMed: 7583634]
- Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci Unit States Am.* 2002; 99:333–338.
- Ledneva RK, Alexeevskii AV, Vasil SA, Spirin SA, Karyagina AS. Structural aspects of interaction of homeodomains with DNA. *Mol Biol.* 2001; 35(5):647–659.
- Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol.* 1987; 4:203–221. [PubMed: 3328815]
- Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function and evolution. *Mol Biol Evol.* 2004; 21:991–1007. [PubMed: 14963101]
- Lovell SC. Are non-function, unfolded proteins (“junk proteins”) common in the genome? *FEBS Lett.* 2003; 554:237–239. [PubMed: 14623072]
- Lynch, M. *The Origins of Genome Architecture.* Sinauer Assocs., Inc; Sunderland, MA: 2007.
- Lynch VJ, Roth JJ, Wagner GP. Adaptive evolution of Hox-gene homeodomains after cluster duplications. *BMC Evol Biol.* 2006; 6(86)
- Lynch VJ, Tanzer A, Wang Y, Leung FC, Gellersen B, Emera D, Wagner GP. Adaptive changes in the transcription factor HoxA-11 are essential for the evolution of pregnancy in mammals. *Proc Natl Acad Sci Unit States Am.* 2008; 105(39):14928–14933.
- Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Shennan Lu, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* 2007; 35:D237–D240. [PubMed: 17135202]
- Mularoni L, Veitia RA, Albà MM. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics.* 2007; 89:316–325. [PubMed: 17196365]
- Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S. Nucleotide compositional constraints on genomes generate alanine-, glycine-, and poline-rich structures in transcription factors. *Mol Biol Evol.* 1997; 14:1042–1049. [PubMed: 9335144]
- Ofran Y, Rost B. Analysing six types of protein-protein interfaces. *J Mol Biol.* 2003; 325:377–387. [PubMed: 12488102]
- Petes TD, Greenwell PW, Dominska M. Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*. *Genetics.* 1997; 146:491–498. [PubMed: 9178000]
- Rice, S. *Evolutionary Theory: Mathematical and Conceptual Foundations.* Sinauer Associates; Massachusetts USA: 2004. p. 76-78.
- Richard GF, Dujon B. Trinucleotide repeats in yeast. *Res Microbiol.* 1997; 148:731–744. [PubMed: 9765857]
- Roth JJ, Breitenback M, Wagner GP. *J Exp Zool B Mol Dev Evol.* 2005; 304B:468–475. [PubMed: 16032701]
- Wagner GP, Pyle AM. Tinkering with transcription factor proteins: the role of transcription factor adaptation in developmental evolution. *Novartis Foundation Symposium.* 2007; 284:115–125.

Wagner, GP.; Schwenk, K. Book title: Evolutionary Biology. Vol. 31. Kluwer Academic/Plenum Publishers; New York, NY: 2000. Chapter title: Evolutionarily stable configurations: Functional integration and the evolution of phenotypic stability; p. 155-217.

Fig. 1.

Equilibrium Length Theoretical Predictions and Simulation Results **Fig. 2.**

Fig. 3.

Fig. 4.

Fig. 5.

Fig. 6.

Fig. 7.

*region with over-representation
of residue type 14* ↘

Initial Sequence: 15, 4, 7, 11, 4, 5, 14, 14, 18, 14, 3, 6, 4, 4, 17, 5, 2, 3, 15, 20

*Environmental
change #1*

{2,5,3,19}: No Domain Exists, longest contiguous stretch = 3

15, 4, 7, 11, 4, 5, 14, 14, 18, 14, 3, 6, 4, 4, 17, 5, 2, 3, 15, 20

*Environmental
change #2*

{18,6,5,4}: No Domain Exists, longest contiguous stretch = 3

15, 4, 7, 11, 4, 5, 14, 14, 18, 14, 3, 6, 4, 4, 17, 5, 2, 3, 15, 20

*Environmental
change #3*

{18,6,5,14}: Domain Exists, longest contiguous stretch = 5

15, 4, 7, 11, 4, 5, 14, 14, 18, 14, 3, 6, 4, 4, 17, 5, 2, 3, 15, 20

Fig. 8.

Fig. 9.

Fig. 10.

Fig. 11.

Fig. 12.

Table I

Simulation result statistics (95% C.I. of means)

	Without secondary functional domains		With secondary functional domains	
	Before touch	After touch	Before touch	After touch
Avg Length	23.12±.15	23.455±.165	19.72±.32	21.45±.29
Deleterious mutation rate	.4368±.0198	.4266±.0286	.45685±.03135	.40315±.00975
Beneficial mutation rate	.0078885±.0000755	.0050405±.0000535	.01648±.00098	.01527±.00039
Absolute value of avg mutation fitness effect	2.7835±.0305	2.7095±.0385	5.3445±.0885	4.709±.081
Avg mutation fitness effect	-2.7765±.0305	-2.705±.039	-4.6215±.1525	-4.2055±.0715
Avg length amino acid tandem repeat at t=400000	1.174±.019		1.6315±.0655	
Frequency of touch-event in 400000 time steps	.375		1	
Avg time point of first touch-event between primary functional domains	349935.5±109361.5		12708±5562	
Sticking frequency	.6667		0.9545	
Rate of random walk of primary function, before first touch-event *	0.00669901±.00127955		.0172808±.0052242	
Rate of random walk of primary function, after first touch-event *	0.005616945±.00141667		0.0198445±.0034311	
Rate of overlap growth of primary functions *	.010182295±.002464605		0.0432076±.0093215	

* Expressed as q in the least squares best-fit regression of form $q(t)^{0.5}$, fit to all time-steps of the simulation (which were started at the equilibrium length)