Complex Adaptive Systems and the Threshold Effect: Views from the Natural and Social Sciences:
Papers from the AAAI Fall Symposium
(FS-09-03)

Emergence of Ultra-Conserved Protein Domains and Amino Acid Repeats: Adaptation, Competition and Thresholds

Mary M. Rorick and Gunter P. Wagner

Department of Ecology and Evolutionary Biology
Yale University
New Haven, CT 06520-8106
USA
molly.rorick@yale.edu

Abstract

Some proteins, such as homeodomain transcription factors, contain highly conserved regions of sequence that cannot be attributed to the constrains imposed by any single function. It has recently been suggested that multiple conserved functional domains overlap and together explain the high conservation of these regions. However, because these highly conserved domains are part of much larger proteins, we are still left with the question why so many functional domains cluster together. Here we have modeled an evolutionary mechanism that can produce this kind of clustering. Due to adaptive competition between different protein functions for control over amino acid residue identity, conserved functional domains get displaced from regions undergoing adaptive evolution. At first they undergo a steady random walk within the sequence for an indefinite amount of time; however, a threshold is reached when two functional domains happen to come into contact, at which point there is a dramatic shift in the adaptive dynamics such that the domains rapidly converge, lengthen, and evolve overlap-- stabilizing at a fully overlapped state.

We also studied the evolution of single amino acid tandem repeats (a.k.a. homopeptides), which are especially prevalent in transcription factors. Homopeptides that are encoded by nonhomogenous mixtures of synonymous codons cannot be explained by the neutral process of replication slippage. Our model provides two ways to explain the origin and maintenance of such repeats, and their overrepresentation in highly conserved proteins: competition between multiple functional domains for space within a sequence, or reuse of a sequence for many functions over time. Both processes depend on reaching certain critical thresholds, however they both deterministically cause the evolution of repeats once these thresholds are reached. Further, both of these processes are characteristic of multifunctional proteins such as homeodomain transcription factors.

We conclude that our model can explain two widely recognized features of transcription factor proteins: conserved domains and a tendency to accumulate homopeptides.

I. Introduction

The ultra-high conservation of some domain superfamilies requires an explanation. The simplest of these is that purifying selection to conserve 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 HoxA-11 sequence among shark, zebrafish, coelocanth, mouse and chicken-- and though this conservation has conventionally been attributed to the constraints imposed on its DNA binding function, there are actually only six residues in the homeodomain that contact DNA bases (Ledneva et al. 2001). Thus, an alternative hypothesis for the evolution and maintenance of highly conserved regions is that multiple conserved functional domains overlap, together imposing sufficiently strong sequence constraints (Roth et al. 2005). However, this hypothesis doesn't specify why so many conserved domains would restrict themselves to one small, constrained region of a much larger protein. In this paper we ask whether there is an evolutionary mechanism that could produce this kind of clustering.

Amino acid repeats (a.k.a simple sequence repeats (SSRs) or homopeptides) are homogenous stretches of a single amino acid type that are anywhere from a few to over 50 amino acids long, and they are a common feature of eukaryotic proteins (Faux et al. 2005). Repeats at the DNA level originate primarily through replication slippage. They are unstable, highly variable, preferentially found in regions of the proteome under weak purifying selection, and associated with deleterious phenotypes (Mularoni et al. 2007, Cummings & Zoghbi 2000, Karlin et al.

2002 Hancock et al. 2001, Levinson & Gutman 1987, Alba et al. 1999b, Alba et al. 2001). Mixed-codon repeats, in contrast, are slow evolving and predominantly located in highly conserved proteins, as well as over-represented in the most constrained proteins (Alba et al. 1999a, Hancock & Simon 2005, Hancock et al. 2001, Mularoni et al. 2007). 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), so this class demands a different explanation for its origin and maintenance. In this paper we propose two such mechanisms.

II. Model

A protein is represented as a sequence of R amino acid residues. There are 20 different amino acid types, designated 1 through 20. The protein can perform multiple distinct functions, each of which is enabled by a unique subset of amino acids that have the potential to contribute to a specific function. These subsets are termed "alphabets", and the amino acids within a given function's alphabet are called the "alphabet amino acids" for that function. In order for a protein to perform a function the protein must have a functional domain within its sequence. A functional domain is defined as a contiguous series of alphabet amino acids that is at least of minimum length m. The performance p of function i increases with b_i -- the number of amino acids in the functional domain of function 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) (Figure 1). According to this model 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.

We think this is a reasonable way to define and distinguish different protein functions because protein functional domains are determined primarily by amino acid composition at "local", consecutive amino acids (Ofran & 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.

A population with effective population size, N_e, is modeled as a single representative wildtype protein sequence that changes through time. We model mutation and selection by allowing a single viable mutation each time-step (each time-step thus represents the average waiting time between mutations). The fitness of the mutant, that has a functional domain of length h, is compared to the fitness of the wildtype, that 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 (1982):

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

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

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

Two types of functions are distinguished in the model. "Primary functions" are those which remain invariant through time and always important under all possible environmental conditions. For instance, transcription factors have several primary functions (such as DNA binding, nuclear localization and transcriptional activity) that are unaffected by 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.

All functional domains have a minimum length of m amino acids. Each of the primary functions is defined by its own unique alphabet (which is a list of numbers, because each amino acid type is represented by a unique number). The primary functions' alphabets are invariant and intersect to some degree. Functions 1 and 2 are examples of primary functions with alphabets that intersect at one amino acid type (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 each environmental change event, and that is constrained to never intersect with either primary function alphabet (to encourage competition between primary and secondary functions). Secondary functions were generally assigned larger alphabets than primary functions because, with every environmental change, they receive new, random alphabets that must have a high likelihood of allowing for functional domains.

The number of amino acids 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 wildtype sequence for the next time-step). The fixation probability is determined by equations 4-6. If the mutation is not fixed, the wildtype 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 generating a new random alphabet for a secondary function. If a viable functional domain exists, the new alphabet for the secondary function is accepted. The secondary function's performance and the protein's overall fitness are then reassessed according to the new secondary function alphabet.

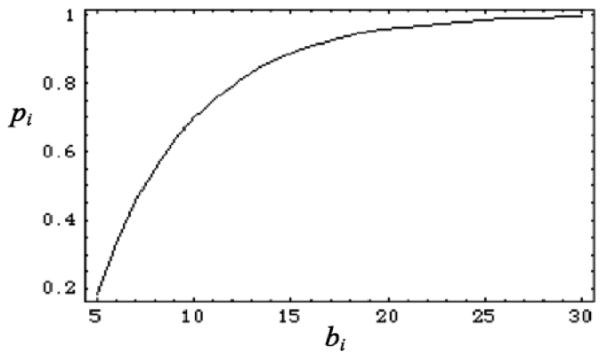

Figure 1: Performance p_i of a single function i depends on b_i, the length of a functional domain (equation 1). The function's performance is zero when the length of functional domain is less than the minimum length m=5. The addition of amino acids to a functional domain is characterized by diminishing returns. The function's performance increases as b_i increases, but this rate of increase of p_i decreases. Thus, a single amino acid within a functional domain is more valuable to performance when it is part of a small domain as opposed to a long domain.

III. Results

A. Phenomenology

Simulations of the model described above were performed under a variety of parameters and initial conditions. The location and lengths of all functional domains were monitored, as well as the amino acid composition of the sequence.

When two primary functional domains with partially intersecting alphabets evolve in the presence of a single competing secondary function, the ever-changing secondary functional domain appears to encroach on the space initially occupied by the two primary functional domains. Once the primary functional domains touch, their frequency of sticking together rather than separating again is 95%. After this "touch-event-threshold" is reached, the primary functional domains undergo a dramatic shift in their dynamics—instead of a slow bidirectional random walk, they move rapidly and directionally, converging with one another, and stabilizing at a fully overlapped state (Figure 2, Table I). Hence, given the presence of variable secondary function, 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.

In addition, if the parameters are set in such a way as to reach a critical threshold (that seems to be a function of selection strength, and thus, population size), regions of homogenous amino acid composition—so called "single amino acid tandem repeats" or "homopeptides"—consistently emerge in the simulations (Figure 3, Table I). We find this pattern intriguing because transcription factor proteins are know to be prone to the evolution of such repeats (Mularoni et al. 2007).

B. Analysis of Phenomenology

In order to understand the mechanisms that lead to the origin of conserved multifunctional domains and amino acid repeats in our simulations, we consider the dynamics of functional domains in various simple scenarios. First we consider a single primary functional domain and how it changes over time without interaction with other domains.

Analysis of the Evolution of a Single Primary Function. Mutation and selection will add and subtract amino acid residues from the domain at the edges, more or less one at a time (because loosing an "internal" amino acid breaks up the domain and causes a large reduction in fitness and mutations that increase the length by more than one amino acid are rare). Adding and losing amino acids at the edges of the domain leads to changes in the length of the domain, which will eventually fluctuate around an equilibrium length. Furthermore, if an amino acid is lost on one side, and another amino acid is gained on the opposite side of a domain, over evolutionary time the domain will essentially change location 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. We can predict this equilibrium length and rate of random walk based on our model.

We consider a discrete stochastic Markov process where the random variable, b, is the length of a domain at a time-step t. With this, we can calculate the expected equilibrium length and random walk rate. Results from simula-

tions with only primary functions are generally consistent with analytical predictions. This indicates that the simplified Markov model sufficiently captures the overall dynamics of the primary functional domain through evolutionary time—both its change in length and its movement along the sequence.

Analysis of Simulations with Primary and Secondary Functions. When primary and secondary functions compete for control over residue identity, we predict that secondary functions generally win because they will, on average, have lower performance than primary functions. To test this prediction, we measured the frequency at which a secondary function's performance increases at the expense of a primary function's performance in the simulations, and compared this to the frequency of the reverse occurrence. Consistent with our predictions, we find that the former occurs more frequently than the latter, indicating that secondary functions do tend to out-compete primary functions for control of residue identity.

Typical simulation results are consistent with our prediction that primary functional domains tend to evolve overlap once they touch (Figure 2, Table I). However, to be sure that this overlap evolution is actually driven by the encroachment of the flanking secondary functions, we must demonstrate that any alternative causes for the evolution of overlap are either unnecessary or insufficient. Indeed—given that the primary domains are provided with sufficient space to reach equilibrium length before they touch-- there are two possible forces driving the rapid overlap evolution in our simulations. First, overlap may provide a fitness advantage by reducing the total number of mutable sites, or by causing deleterious mutations to be more severe (since mutations can harm two functions at once), and thus, easier to select against. Overlap would therefore evolve because mutational degradation would be faster on the non-overlapped end compared with the overlapped end of a functional domain. While we cannot remove the effects of this mechanism, we can hold it constant while we test a second potential driving force for the evolution of overlap—i.e. the competitive degradation of primary functional domains by secondary functional domains preferentially on their non-overlapped ends. Only this latter mechanism features a role for secondary functions—so what we want to understand is whether it 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.

If secondary functions play a significant role in the evolution of overlap between primary functional domains, we would expect to observe significantly more proficient evolution of overlap between primary functional domains when secondary functional domains are present. Indeed, we find that the presence of secondary functions has many significant effects that are consistent with this prediction. Among other things, their presence increases the random walk rate of primary functional domains, the frequency at which they touch, the probability of "sticking", and the overlap-lengthening rate (Table I).

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

	Secondary functional domains absent	
	Before touch	After touch
Avg Length	22.97, 23.27	23.29, 23.62
Deleterious mut. rate	.4170, .4566	.3980, .4552
Beneficial mut. rate	.007813, .007964	.004987, .005094
Absolute value of avg mutation fitness effect	2.753, 2.814	2.671, 2.748
Avg mut. fitness effect	-2.807, -2.746	-2.744, -2.666
Avg repeat length *t*=400000	1.155, 1.193	
Freq. touch-events *t*=400000	.375	
Time 1st touch-event	240574, 459297	
Sticking frequency	.6667	
Random walk rate, before touch-event*	0.00541946, 0.00797856	
Random walk rate, after touch-event*	0.00420027, 0.00703362	
overlap growth rate*	0.00771769, 0.0126469	

	Secondary functional domains present	
	Before touch	After touch
Avg Length	19.40, 20.04	21.16, 21.74
Deleterious mut. rate	.4255, .4882	.3934, .4129
Beneficial mut. rate	.01550-.01746	.01488, .01566
Absolute value of avg mutation fitness effect	5.256, 5.433	4.628, 4.790
Avg mut. fitness effect	-4.774, -4.469	-4.277, -4.134
Avg repeat length t=400000	1.566, 1.697	
Freq. touch-events t=400000	1	
Time first touch-event	7146, 18270	
Sticking frequency	0.9545	
Random walk rate, before touch-event*	0.0120566, 0.022505	
Random walk rate, after touch-event*	0.0164134, 0.0232756	
overlap growth rate*	0.0338861, 0.0525291	

*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).

Mechanisms of Single Amino Acid Tandem Repeat Evolution. In running the simulations we also observed the evolution of single amino acid tandem repeats. Because the alphabets of the various functions contained at least four different amino acid types each, these repeats cannot be attributed to simple functional domain growth. Two causes for the evolution of these amino acid repeats are considered.

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 selec-

tion for overlap drives the evolution of amino acid repeats. A second possibility is that the origin of repeats depends on the secondary functions. Given the frequent emergence of new functions with new, random alphabets, a region of the sequence that happens to have a local over-representation of a single amino acid type is more likely to become a functional domain, and thus, undergo selection for further enrichment of the already over-represented amino acid type. Thus, given the design of our model under environmental change, any local over-representation of an amino acid might be self-enforcing.

We tested both of these mechanisms through simulations. The first mechanism can be isolated from the second by examining the evolution of several primary functional domains with partially intersecting alphabets evolving under crowded conditions within a sequence. The second hypothesized mechanism for the evolution of repeats can be isolated from the first mechanism by setting up a simulation in which there is only a single secondary function. We can then compare whether sequences resulting from either of these types of simulations contain significantly longer amino acid repeats than those of non-evolved sequences. We find that both of our hypothesized mechanisms are independently sufficient for the evolution of amino acid tandem repeats.

IV. Discussion

Here we model an evolutionary mechanism that can produce both 1) clustering of conserved functional domains into a relatively small, highly constrained multifunctional protein domain, 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 & Guigo 2004, Faux et al. 2005, Richard & Dujon 1997, Nakachi et al. 1997). Regulatory proteins are highly conserved, contain regions under adaptive evolution, and are the sites of innovation (Lynch et al. 2008)-- and these features are what drive the mechanisms described by our model.

An important future direction will be to analyze sequence data for the general patterns predicted by this model. For example, if amino acid repeats evolve by either of the mechanisms suggested here, we would expect repeat-containing regions to show evidence of competing functions. In one case we find this pattern (for HoxA13 (Crow et al. 2009)). The model also predicts that conserved domains will contain competing functions, and again, we have anecdotal evidence that this is the case (for HoxA11 (Lynch et al. 2008)). Of course, it would be useful to look for evidence of competing functions across many different classes of repeat-containing and conserved proteins.

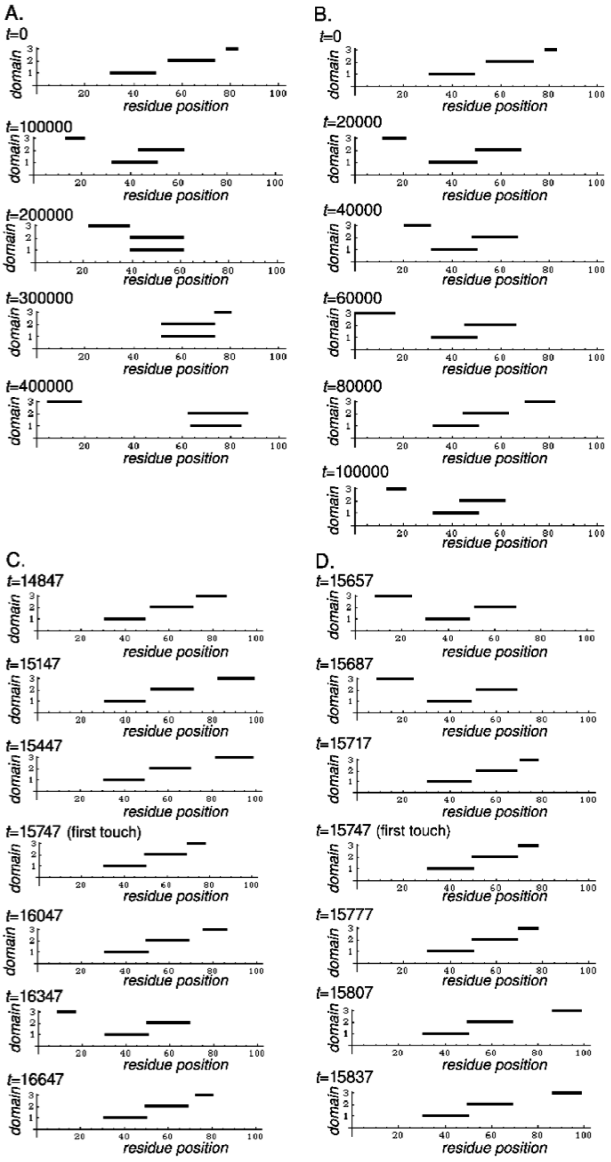

Figure 2: The location of functional domains along the sequence at various time-points *t* throughout a single typical simulation. Functions 1 and 2 are primary functional domains, function 3 is a secondary functional domain. "A" shows evenly spaced time-steps for the entire simulation of 400,000 time-steps. "B" shows evenly spaced time points for the first 100,000 time-steps of the simulation. "C" shows evenly spaced time-steps spanning a range about 1000 time-steps before and after the time-step when the primary domains first touch. D shows evenly spaced time-steps spanning a range about 100 time-steps before and after this time-step.

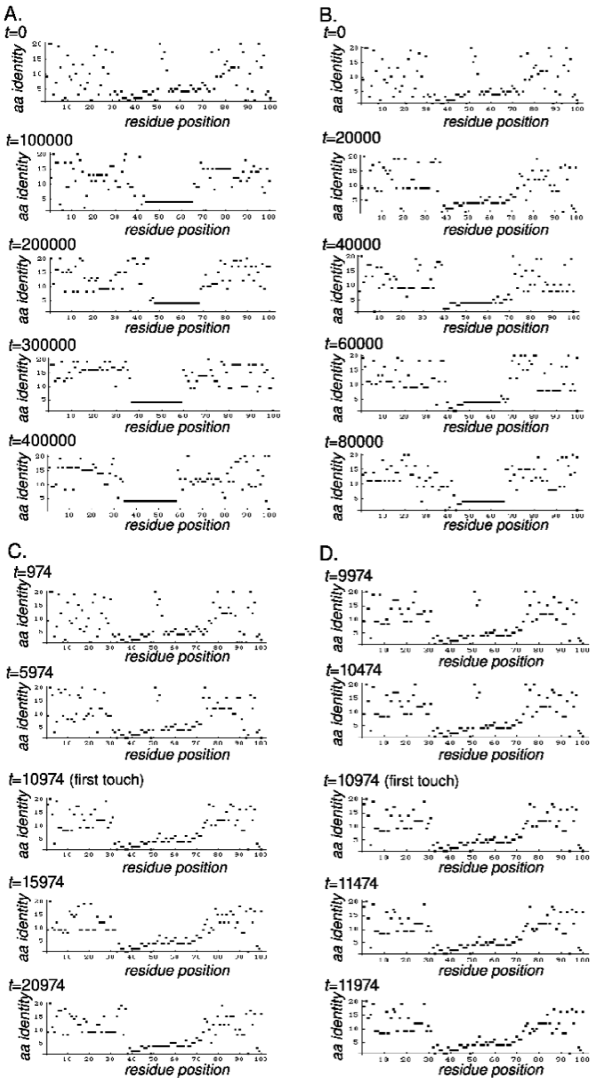

Figure 3: The residue identity along the sequence at various time-steps *t* throughout a single typical simulation. "A" shows evenly spaced time-steps for the entire simulation of 400,000 time-steps. "B" shows evenly spaced time-steps for the first 100,000 time-steps of the simulation. "C" shows evenly spaced time-steps spanning a range about 10000 time-steps before and after the time-step when the primary domains touch. "D" shows evenly spaced time-steps spanning a range about 1000 time-steps before and after this time-step.

Acknowledgements: 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 are not necessarily reflecting the views of the JTF.

References

Alba MM and Guigo R. 2004. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res* 14:549-554

Alba MM, Santibanez-Koref MF, Hancock JM. 1999a. Conservation of polyglutamine tract size between mouse and human depends on codon interruption. *Mol Biol Evol* 16: 1641-1644

Alba MM, Santibanez-Koref MF, Hancock JM. 1999b. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of slippage-like mutational process *J Mol Evol* 49 789-797

Alba MM, Santibanez-Koref MF, Hancock JM. 2001. The comparative genomics of polyglutamine repeats: extreme differences in the codon organization or repeat-encoding regions between mammals and Drosophila *J Mol Evol* 52: 249-259

Crow K, Amemiya CT, Roth J, Wagner GP. 2009. 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* Epub ahead of print.

Cummings CJ, Zoghbi HY. 2000. Trinucleotide repeats: Mechanisms and pathophysiology. *Annual Review of Genomics and Human Genetics* 1: 281-328

Faux NG, Bottomley SP, Lesk AM. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res* 15: 537-551

Hancock JM, Simon M. 2005. Simple sequence repeats in proteins and their significance for network evolution. *Gene* 345: 113-118

Hancock JM, Wothey EA, Santibanez-Koref MF. 2001. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Mol Biol Evol* 18 (2001) 1014-1023.

Hill WG. 1982. Rates of change in quantitative traits from fixation of new mutations. *PNAS USA* 79: 142-145

Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. 2002. Amino acid runs in eukaryotic proteomes and disease associations. *PNAS USA* 99: 333-338

Ledneva RK, Alexeevskii AV, Vasil SA, Spirin SA, Karyagina AS. 2001. Structural aspects of interaction of homeodomains with DNA. *Molecular Biology* 35(5): 647-659

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4: 203-221

Lynch VJ, Tanzer A, Wang Y, Leung FC, Gellersen B, Emera D, Wagner GP. 2008. Adaptive changes in the transcription factor HoxA-11 are essential for the evolution of pregnancy in mammals. *PNAS* 105(39): 14928-14933

Mularoni L, Veitia RA, Albà MM. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* 89: 316-325

Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S. 1997. Nucleotide compositional constraints on genomoes generate alanine-, glycine-, and poline-rich structures in transcription factors. *Mol Biol Evol* 14:1042-1049

Ofran Y, Rost B. 2003. Analysing six types of protein-protein interfaces. *J Mol Biol* 325: 377-387

Petes TD, Greenwell PW, Dominska M. 1997. Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. *Genetics* 146: 491-498

Richard GF, Dujon B. 1997. Trinucleotide repeats in yeast. *Res Microbiol.* 148:731-744

Roth JJ, Breitenback M, Wagner GP. 2005. *Journal of Experimental Zoology (Mol Dev Evol)* 304B: 468-475