

Repeat-Modulated Population Genetic Effects in Fungal Proteins

F. N. Braun,¹ D. A. Liberles^{1,2}

¹ Department of Biochemistry and Biophysics, and Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden
² Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway

Received: 16 October 2003 / Accepted: 23 January 2004

Abstract. A number of fungal lineages, notably *N. crassa*, have evolved a novel mechanism of processing genomic duplication events known as repeat-induced point (RIP) mutation. This mechanism appears, on the one hand, to act as a conservative genomic safeguard, by introducing stop codons into duplicated nucleotide sequences, thereby preempting consequences such as dosage effects. However, it also typically performs further nonsynonymous (i.e., amino acid-changing) nucleotide substitutions, the significance of which is unclear. We explore here the possibility that RIP-mutated genes which evade silencing may have some microevolutionary impact on functional sequences. Our approach focuses on structurally important hydrophobic/polar (HP) amino-acid substitutions effected by RIP. We exploit a simple generic protein folding model to predict the associated emergence of increased protein-structural stability and variance within a large population.

Key words: Fungi — Protein evolution — Gene duplication — Population genetic variability

Introduction

The recent publication of a draft whole genome sequence for the filamentous fungus *Neurospora crassa* (Galagan et al. 2003) has confirmed considerable in-

cidence of the novel genomic mechanism known as repeat-induced point mutation (RIP). Following a gene duplication event, RIP is able to detect and respond to sequence similarity by indiscriminately subjecting both the original and its copy to rounds of heavy G-to-A and C-to-T nucleotide substitution (Watters et al. 1999). The rounds may continue until sequence similarity falls below a threshold of roughly 80%.

RIP has been observed in filamentous fungi only. However, related homology-dependent mechanisms occur in other fungi, in plants, and even in metazoa. While the biochemical nature of these processes remains obscure, it appears feasible that they have derived respectively from some shared protomechanism which existed in the last common ancestor (Meyer 1996). Current understanding posits RIP as a genome defense mechanism capable of forestalling duplication-mediated proliferation of deleterious new and/or selfish genes, dosage effects, and dominant negative effects. This follows from the observation that RIP generates amber and ochre stop codons (TAG, TAA) with a high probability, such that both of the pair are likely to be silenced, i.e., rendered functionally inactive. The probability P_s that a gene survives any given round functionally intact is nonvanishing nevertheless, and the latter scenario is the focus of the present paper.

To roughly estimate P_s , consider the table of amino acid (aa)-changing nucleotide substitutions effected under RIP (Fig. 1). A RIP pass performs either multiple G-to-A or multiple C-to-T but not both simultaneously (Watters et al. 1999). Moreover, the shaded scenarios in the table are generally observed with sig-

nificantly higher frequency. Watters et al. have estimated the high mutation probability of CpA sites during a C-to-T pass at around $r \simeq 0.3$. Let us derive from r the probability $1 - P_s$ that a random N -residue aa sequence is silenced during a C-to-T pass:

$$\begin{aligned}
 1 - P_s &= \sum_{x=1}^N P(x \text{ stop precursors in aa seq}) \\
 &\quad \times P(\text{at least one of } x \text{ converted}) \\
 &= \sum_{x=1}^N \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x} \times (1 - [1-r]^x)
 \end{aligned} \tag{1}$$

Here p is the probability that a given site along the sequence is a stop precursor. The only stop precursor in a CpA-to-TpA pass is Gln, from the table (Fig. 1). If we construct codons by randomly stringing together nucleotides, we have for the Gln precursor $p = 1/32$ since it is encoded by 2 of the total 4^3 possible codons.

RIP is thought to detect duplicates only above a threshold length of ~ 400 bp. Figure 2 shows that, at least close to this threshold, the survival probability P_s estimated from Eq. (1) is sufficiently high to motivate an appraisal of the possible consequences of rip-mutated genes which escape silencing. In particular, the evident ‘‘scrambling’’ nature of RIP might influence population genetic variability. The objective in the following is to develop semiquantitatively the theoretical basis for such an effect.

Theory

The section below introduces an iterative population dynamics framework appropriate to genes affected by RIP. We proceed from the observation that RIP is able to alter the relative composition of hydrophobic (H) vs polar (P) residues along the translated aa sequence; C-to-T passes lead mostly to polar-to-hydrophobic aa substitutions, while G-to-A lead mostly to hydrophobic-to-polar substitutions. It is convenient then to follow a matrix formulation of HP compositional dynamics, similar to recent investigations of intragenic recombination (Cui et al. 2002; Xia and Levitt 2002). This will serve as a basic platform from which to examine in a general way how, on an evolutionary time scale, duplication-triggered RIP shapes population distributions in the HP interpretation of genotype space. The approach lends itself, moreover, to analogous calculation over the phenotype space of the gene, where phenotype is dictated by folding.

Quantitative accuracy is curtailed by several simplifying assumptions. These partly reflect gaps in the

available experimental data with respect to key parameters of our model but also help to keep the analysis straightforward and illustrative. Perhaps most importantly, we assume that each duplication event triggers only a single pass of RIP, as opposed to the multiple passes observed experimentally. Another important assumption, controlled via the parameter α below, is that of equal frequency of C-to-T passes relative to G-to-A passes. This is not obvious from experimental data, and it will affect the robustness of the conclusions we reach concerning stability and packing of tertiary structure.

Repeat Modulated Sequence Microevolution

In Fig. 1 we have annotated the HP context of each nonsynonymous transition. For example, Ser-to-Leu constitutes P-to-H. The coarse folded structure of proteins generically comprises an H-rich core shielded from the aqueous environment by P residues exposed at the surface, so the capacity of RIP to alter H:P composition is very significant in this tertiary-structural respect.

Chan and Bornberg-Bauer (2002) have recently reviewed various simple exact model perspectives on the evolution of sequence HP composition. This evolution is driven by some spontaneous rate of HP flipping, originating from the error rate in nucleotide replication, i.e., the molecular clock of Zuckerkandl and Pauling (1965). About 30% of all the possible nucleotide substitutions consitute HP flips at the codon level. For a gene comprising ~ 400 bp, if we assume a selectively neutral clock rate of order 10^{-9} per year per encoded aa site and make the Jukes–Cantor (1969) approximation (all nucleotide switches equally likely), the neutral rate of HP flipping events per gene is $\mu_{\text{hp}} \sim 0.01 \text{ Myr}^{-1}$.

From a microevolutionary perspective, μ_{hp} drives the emergence of differences in HP composition of this gene between members of a population. Taking as our HP compositional space variable the number n of hydrophobic residues comprising the gene, a large population evolves with time t as

$$\Gamma \phi_n(t + \delta t) = \phi_n(t) + \mu_{\text{hp}} M_{nn'} \phi_{n'}(t) \tag{2}$$

Here ϕ_n denotes the frequency of composition n within the population and Γ is a normalization factor guaranteeing $\sum_n \phi_n(t + \delta t) = 1$.

The matrix $M_{nn'}$ defines the relative probabilities of sequence mutations $n' \rightarrow n$. It is convenient to assume equal weighting of H-to-P flips vs P-to-H flips (in fact, codon table inspection suggests that under the Jukes–Cantor scheme, H-to-P flips are more frequent than P-to-H flips by a factor ~ 1.2). With this

3rd pos ⁿ		A	C	G	T	
1st & 2nd pos ⁿ	AC			Thr		P
	r					
	AT	Ile	Ile	Met	Ile	H
	r					
	CC			Pro		H
	r					
	CT			Leu		H
	r					
	GC			Ala		H
	r					
	GT			Val		H
	r					
	TC			Ser		P
	r					
	TT	Leu	Phe	Leu	Phe	H

3rd pos ⁿ		A	C	G	T	
1st & 2nd pos ⁿ	CA	Gln	His	Gln	His	P
	r					
	TA	■	Tyr	■	Tyr	P
	r					
	CC			Pro		H
	r					
	TC			Ser		P
	r					
	CG			Arg		P
	r					
	TG	■	Cys	Trp	Cys	P
	r					
	CT			Leu		H
	r					
	TT	Leu	Phe	Leu	Phe	H

(b) non-synonymous G \xrightarrow{r} A substitutions

3rd pos ⁿ		A	C	G	T	
1st & 2nd pos ⁿ	AG	Arg	Ser	Arg	Ser	P
	r					
	AA	Lys	Asn	Lys	Asn	P
	r					
	CG			Arg		P
	r					
	CA	Gln	His	Gln	His	P
	r					
	GG			Gly		H
	r					
	GA	Glu	Asp	Glu	Asp	P
	r					
	TG	■	Cys	Trp	Cys	P
	r					
	TA	■	Tyr	■	Tyr	P

3rd pos ⁿ		A	C	G	T	
1st & 2nd pos ⁿ	GA	Glu	Asp	Glu	Asp	P
	r					
	AA	Lys	Asn	Lys	Asn	P
	r					
	GC			Ala		H
	r					
	AC			Thr		P
	r					
	GG			Gly		H
	r					
	AG	Arg	Ser	Arg	Ser	P
	r					
	GT			Val		H
	r					
	AT	Ile	Ile	Met	Ile	H

3rd pos ⁿ		A	C	G	T	
1st & 2nd pos ⁿ	AG	Met	Ile			H → H
	r					
	TG	Trp	■			

Fig. 1. Table of all possible nonsynonymous (amino acid-changing) codon substitutions realizable in a single RIP pass. Each pass is observed experimentally to comprise either (a) C-to-T or (b) G-to-A, but not both simultaneously. The shaded substitutions occur with higher probability, reflecting an observed bias toward mutation of CpA/TpG sites. This bias applies also across codons. Hence the stripe-shaded substitutions have a high probability if the preceding codon has a T in the third position. Mutation to the black squares denoting stop codons effectively silences translation of a functional protein. The focus of the present discussion is on the columns at the right, showing that RIP also effects structurally important substitutions between hydrophobic (H) and polar (P) residues.

assumption we have straightforwardly for an N-residue sequence (Braun, 2004)

$$M_{nn'} = \left(\frac{n'}{N}\right)\delta_{n,n'-1} + \left(1 - \frac{n'}{N}\right)\delta_{n,n'+1} - \delta_{n,n'} \quad (3)$$

To study RIP mutation, we generalize Eq. (2) to

$$\Gamma\phi_n(t+1) = \phi_n(t) + \mu_{\text{hp}}M_{nn'}\phi_{n'}(t) + 2\zeta P_s R_{nn'}\phi_{n'}(t) \quad (4)$$

where ζ is the rate of gene duplication, and $R_{nn'}$ is a RIP-mutational matrix analogous to the clock matrix $M_{nn'}$. Here we have implicitly introduced the simplifying assumption that only a single pass of RIP occurs with each duplication event. The probability that one of the duplicate pair survives is then $\simeq 2P_s$. It is reasonable to neglect the scenario in which both survive.

In deriving $R_{nn'}$, we remark first with reference to Fig. 1 that in the HP perspective a C-to-T pass becomes tantamount to a P-to-H pass, while conversely, a G-to-A pass interprets as a H-to-P pass. To see this, note that there are two high-probability (i.e., shaded) P-to-H substitutions in a C-to-T pass, but no H-to-P. Conversely, during a G-to-A pass, there are eight possible H-to-P, but never P-to-H. Let us introduce a weighting factor α which takes some fixed value between 0 and 1, expressing the relative frequencies of

occurrence of the two types of pass. As mentioned above, we do not know this parameter experimentally. In the Calculation section we set $\alpha=0.5$, which is tantamount to assuming that C-to-T vs G-to-A passes occur with equal frequency.

The two P-to-H precursors are ACA (Thr) and TCA (Ser). Thus in a random sequence, the probability that a given polar residue is also a P-to-H precursor is $2/33$ (a total of 33 codons encode polar residues), yielding for the probability that any given polar codon flips to hydrophobic during a C-to-T pass

$$\hat{r} = 2r/33 \simeq 0.02$$

The matrix elements $R_{nn'}$ now follow for $n' < n$ as the α -weighted probability of $n - n'$ P-to-H flips from $N - n'$ polar residues.

$$R_{nn'} = \frac{(N - n')!}{(N - n)!(n - n')!} \alpha \hat{r}^{n-n'} (1 - \hat{r})^{N-n} \quad (n' < n) \quad (5)$$

For convenience, we take \hat{r} also as the H-to-P flipping probability during a G-to-A pass. More generally, we could of course implement a separate probability. However, we are not aware of any experimental value for r for this case, on which to base an estimate.

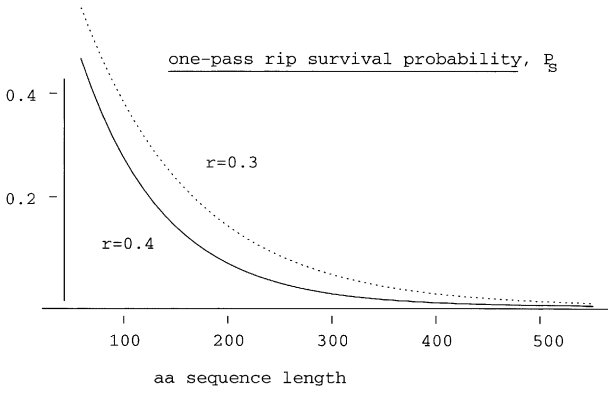


Fig. 2. With each RIP pass acting on a protein-coding gene, there is a high probability that a stop codon will be introduced, which usually can be assumed to effectively silence the gene. However, the probability P_s that the gene avoids this fate is nonnegligible near the threshold ~ 400 bp, above which RIP is observed to operate. In the simulated trend calculated here from Eq. (1), we generate a random nucleotide sequence and mutate CpA sites with probability r .

For the $n' > n$ side (G-to-A pass), we then have

$$R_{nn'} = \frac{n!}{n!(n'-n)!} (1-\alpha)\hat{r}^{n'-n}(1-\hat{r})^n \quad (n' > n) \quad (6)$$

i.e., the $(1-\alpha)$ -weighted probability of $n'-n$ flips from n' hydrophobic residues.

Note here the useful property

$$R_{nn'} = \frac{1-\alpha}{\alpha} R_{(N-n)(N-n')} \quad (n' > n) \quad (7)$$

Finally, along the diagonal $n' = n$,

$$R_{nn'} = \alpha(1-\hat{r})^{N-n} + (1-\alpha)(1-\hat{r})^n - 1 \quad (n' = n) \quad (8)$$

Coupling to Tertiary-Structural Traits

Thus far our evolutionary dynamics focuses exclusively on “genotype,” interpreted in the sense of HP composition of the encoded aa sequence. In order to develop a phenotype perspective, we introduce here a mapping between HP composition and a crude illustrative interpretation of tertiary-structural “phenotype traits”: (i) residue packing fraction and (ii) thermodynamic stability.

The generics of folding with respect to HP composition are captured in a heteropolymer collapse approach initiated by Dill (1985). Following in this spirit, as recently implemented by Braun (2004) in a sequence-evolutionary context, we consider a conformational free energy comprising three terms,

$$F = F_1(\text{hydrophobicity}) + F_2(\text{hydration}) + F_3(\text{polymeric elasticity}) \quad (9)$$

The first term expresses an effective attraction between the n hydrophobic residues, governed by a Flory parameter χ ,

$$F_1/NkT = -\rho\chi\left(\frac{n}{N}\right)^2 \quad (10)$$

where ρ is the residue packing fraction.

The hydration term accounts for the distributional entropy of solvent molecules trapped within the conformation, after the Flory–Huggins fashion,

$$F_2/NkT = \frac{1-\rho}{\rho} \ln(1-\rho) \sim \rho/2 + \rho^2/6 + O(\rho^3) \quad (11)$$

The third term expresses the cost of stretching the random coil conformation entropically favored by polymeric connectivity, $R^2 \sim Na^2$, where R is the mean end-to-end distance and a is a characteristic residue dimension,

$$F_3/NkT = \frac{3}{2} \left(\frac{R}{Na} \right)^2 \quad (12)$$

Writing $\rho = 3Na^3/4\pi R^3$, Eq. (9) expands as

$$F[\rho]/NkT \simeq A\rho + B\rho^2 + C\rho^{-2/3} \quad (13)$$

with

$$A = 1/2 - \chi\left(\frac{n}{N}\right)^2; \quad B = 1/6; \quad C = \frac{3}{2} \left(\frac{3}{4\pi} \right)^{2/3} N^{-4/3} \quad (14)$$

Minimization of $F[\rho]$ with respect to ρ determines equilibrium structure and thermodynamic observables. We define the first of our two phenotype traits as the equilibrium packing fraction $\bar{\rho}$. Familiar scaling concepts (Bryngelson and Billings 1997) lead to qualitatively distinct regimes for this trait. If $A = 0$ in Eq. (13), the protein is in the so-called random coil state, scaling as $\bar{\rho} \sim N^{-1/2}$. If $A > 0$, the protein swells to a lower packing fraction, scaling as $\bar{\rho} \sim N^{-4/5}$. Let us label this swollen state “unfolded.” In the converse situation $A < 0$, the protein is “folded” in the crude sense that it is compact, i.e., $R^3 \sim Na^3$. The protein folds according to Eqs. (13) and (14) above a threshold HP composition,

$$\frac{n}{N} > \frac{1}{\sqrt{2\chi}} \quad \text{folded (compact)} \quad (15)$$

Steric interaction between the residues imposes some maximally compact packing fraction ρ_{\max} ,

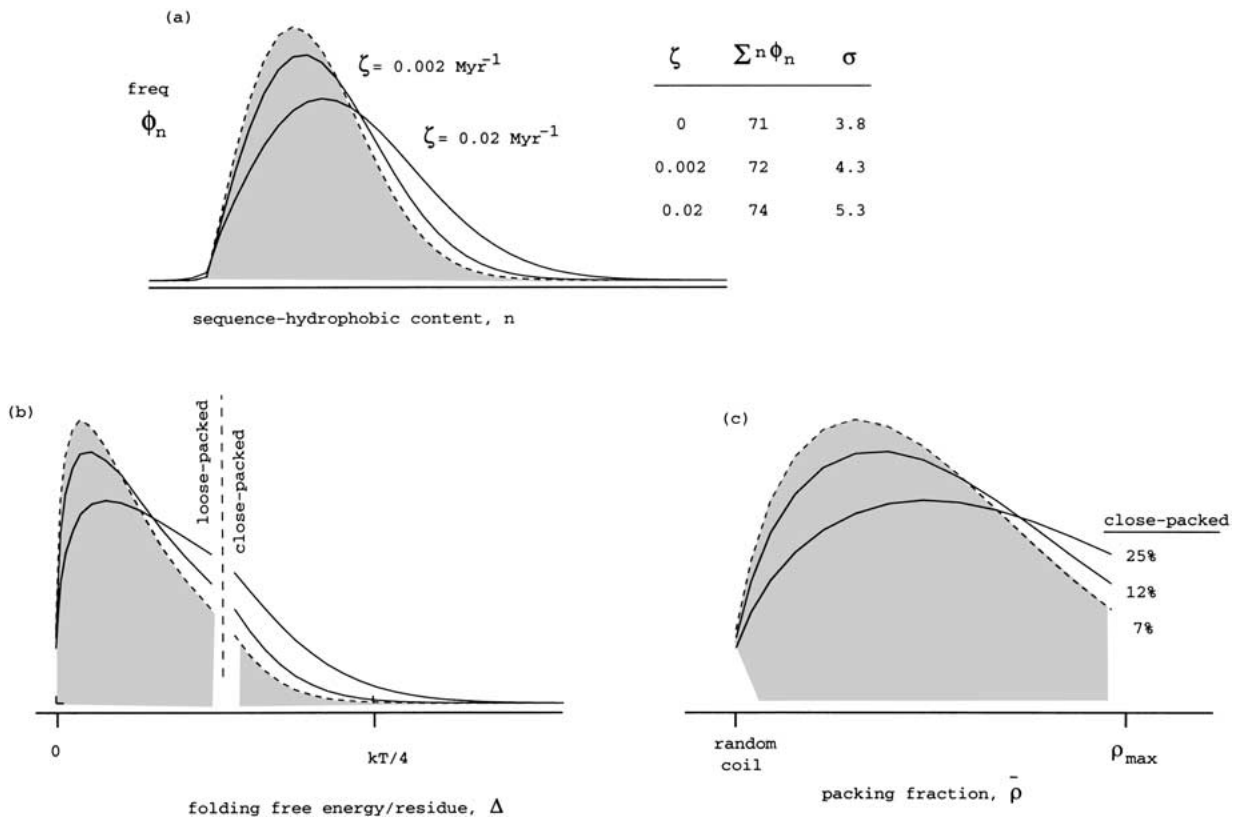


Fig. 3. Steady states of the RIP-modulated evolutionary dynamics, Eq. (4), for a 400-bp protein-encoding gene evolutionarily constrained by compact thermodynamic stability of the folded phenotype. The profiles in **a** depict relative HP compositional frequencies of this gene within a large population. RIP modulation is triggered by the rate of duplication ζ . The RIP survival probability P_s (Fig. 2) drives an increase both in mean hydrophobic content

$\langle n \rangle = \sum_n n \phi_n$ and in variance σ^2 . Profiles **b** and **c** show the corresponding population distributions over the phenotype space of tertiary-structural folding stability Δ and packing density $\bar{\rho}$. The phenotype is close-packed above the stability threshold marked in **b**. The percentages to the right in **c**, giving the close-packed proportion of the respective distributions, show an increase with duplication rate.

Thus we may further distinguish a “close-packed” folded conformation with $\bar{\rho} = \rho_{\max}$ vs a “loose-packed” folded conformation, $\bar{\rho} < \rho_{\max}$. We adopt $\rho_{\max} = 0.64$, a well-known result for randomly packed spheres.

For our stability trait, we take the folding free energy per residue, defined with respect to the random coil,

$$\Delta = \frac{1}{N} \{F[\text{random coil}] - F[\bar{\rho}]\} \quad (16)$$

Calculation

To complete the objective of the theory, we wish now to contrast the population genetic structure of an archetypal RIP-modulated gene at different rates of duplication. For organisms such as *N. crassa* where whole-genome data are available, an averaged rate of gene duplication, the parameter ζ of our evolutionary dynamics Eq. (4), can in principle be inferred from the age distribution of ancient duplicates. The age of a pair follows from the extent to which their respec-

tive sequences have neutrally diverged. For our model purposes, we consider the range $\zeta \sim 0.002 - 0.02 \text{ Myr}^{-1}$ estimated by Lynch and Conery (2000) from several genomes, including the yeast *S. cerevisiae*.

In Fig. 3 we show the resulting population steady states of our model, as applied to a 400-bp protein-encoding gene. The dynamics is constrained by the heteropolymer folding condition, Eq. (15). That is, nonfolding mutants are removed from the population in each iterated step. The hydrophobic Flory parameter χ governing folding is set to a physically plausible $\chi = 2$ (Braun 2004).

The shaded profiles in Fig. 3 represent the bare clock-driven steady state, in the absence of duplication events. When duplication is introduced, RIP mutation broadens and shifts the bare profile to the right (solid lines). In the case of “genotype”, Fig. 3a, this interprets as a shift to a higher mean hydrophobic content $\langle n \rangle$ and variance σ^2 . We see from the tabulated values that a repeat-modulated 400-bp sequence is predicted to have on average one to three extra hydrophobic residues. Figures 3b and c, the “phenotype trait” profiles, respond accordingly; in-

creasing hyperphobic content of the sequence underpins an increase in mean stability and compactness of the tertiary structure.

Digressing slightly, it is worth noting that experimentally, globular proteins generally exhibit marginal stability, in agreement with our Fig. 3b. Marginal stability emerges in the present heteropolymer idealization because the HP flipping probability drives the genotype toward the binomial mean, and this happens to correspond to a marginally stable phenotype. In this respect, the heteropolymer approach falls in line with an argument of Taverna and Goldstein (2002), who attribute marginal stability to designability in sequence space when the major determinant of selection is folding. From these models, a neutrally evolving population tends to wander toward some maximally designable folding phenotype, the binomial mean of the simple HP dynamics.

Conclusion

The model presented here suggests, on the one hand, that in an otherwise neutrally evolving population the acquisition of RIP confers an increase in protein primary- and tertiary-structural variability. This is a reasonably intuitive corollary of our starting ansatz, namely, that protein-encoding genes which are “scrambled” by RIP mutation may nevertheless, with some small probability, continue to translate into functional proteins. Less obviously, our detailed attention to the HP context of RIP substitutions, coupled to a very basic heteropolymer view of tertiary-structural folding, leads to the assessment that RIP also tends to generate more compact and thermodynamically stable tertiary structure.

It should be stressed that the latter point rests to an extent on the as yet rather arbitrary assumption of equal incidence of the two types of pass. We argued that the C-to-T type performs mainly P-to-H flips along a sequence, which, in a heteropolymer phenotype, constitutes a stabilizing influence. In contrast, a G-to-A pass performs H-to-P flips, exerting a destabilizing influence.

It is important also to qualify the single-pass nature of the analysis. Galagan et al. (2003) have carried out simulations which suggest that, under further rounds, the survival probability P_s is reduced by up to an order of magnitude by the time the duplicate pair has been mutated to below 80% similarity. We

anticipate a correspondingly reduced population genetic impact.

Figure 2 illustrates that even in a single pass, P_s falls off rapidly anyway for genes significantly larger than the 400-bp threshold above which RIP operates. However, insofar as 400 bp constitutes a typical length for globular protein genes, one way of attempting to loosely corroborate our predictions would be to compare the folding stabilities of fungal globular proteins with those of nonfungal orthologous counterparts.

Acknowledgments. Funding for this work was provided by the Swedish Foundation for Strategic Research. We also acknowledge the reviewers of this paper for some valuable comments.

References

- Braun FN (2004) Sequence variability of proteins evolutionarily constrained by solution-thermodynamic function. *Phys Rev E* 69:011903:1–8
- Bryngelson JD, Billings EM (1997) From interatomic interactions to protein structure. In: Flyvbjerg H. et al (eds) *Physics of biological systems: From molecules to species*.
- Chan HS, Bornberg-Bauer E (2002) Principles of protein evolution: A perspective from simple exact models. *Appl Bioinform* 1:121–144
- Cui Y, Wong WH, Bornberg-Bauer E, Chan HS (2002) Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc Natl Acad Sci USA* 99(2):809–814
- Dill KA (1985) Theory of the folding and stability of globular proteins. *Biochemistry* 24:1501–1509
- Galagan JE, et al. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422:859–868
- Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) *Mammalian protein metabolism*, Academic Press, New York, pp 21–123
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155
- Meyer P (1996) Repeat-induced gene silencing: Common mechanisms in plants and fungi. *Biol Chem Hoppe-Seyler* 377:87–95
- Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? *Proteins* 46:105–109
- Watters MK, Randall TA, Margolin BS, Selker EU, Stadler DR (1999) Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sized in *neurospora*. *Genetics* 153:705–714
- Xia Y, Levitt M (2002) Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc Natl Acad Sci USA* 99(16):10382–10387
- Zuckerandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) *Evolving genes and proteins*. Academic Press, New York, pp 97–166