

# Retention of enzyme gene duplicates by subfunctionalization

F.N. Braun<sup>a,\*</sup>, D.A. Liberles<sup>a,b</sup>

<sup>a</sup> *Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden*

<sup>b</sup> *Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway*

Received 28 January 2003; received in revised form 13 June 2003; accepted 16 June 2003

## Abstract

Duplication–degeneration–complementation (DDC) describes a process by which evolving duplicates of a pleiotropic ancestral gene divide up the multiple functions of the ancestor between them (i.e. subfunctionalize), and this ultimately frustrates the rate of pseudogene formation. Focusing explicitly on enzyme-like pleiotropic function, we model DDC driven by sequence divergence between duplicates. The model incorporates an idealized sequence–function mapping in which enzyme–substrate binding affinity is related to hydrophobic versus polar (HP) amino-acid composition of tertiary structure about the binding pocket. In this sense, a transparent coupling between physical–chemical function of an enzyme and sequence evolution is presented.

© 2003 Elsevier B.V. All rights reserved.

*Keywords:* Gene duplication; Sequence–structure–function correlation; Enzyme specificity; Protein evolution

## 1. Introduction

Gene duplication has long been regarded as one of the principal engines powering the evolution of protein function. Duplicates mutating collectively arguably have greater freedom to explore sequence space which would otherwise be deleterious, and this may occasionally foster neofunctionalization, the spontaneous evolution of one of the duplicates to a novel function, while the original function continues to be carried by a paralog [1]. Neofunctionalization (and the retention of duplicates in general) is however clearly a rare outcome of duplication events, and in the main it will be preempted by several possible genomic mechanisms which contribute to effectively silencing redundant duplicates, such that they no longer translate into functional proteins. Large chunks of eukaryotic genomes comprise dead pseudogene debris testifying to this birth/silencing cycle [2].

Existing data suggest that the incidence of duplicates which are retained is higher than expected via immediate neofunctionalization alone. Force et al. [3] have pointed out that if the duplicated gene is initially pleiotropic, i.e. is responsible for more than one function, then a further possible retention mechanism, ‘duplication–degeneration–

complementation’ (DDC) in their terminology, may help to explain the discrepancy. This describes a situation in which duplicates subfunctionalize, in the sense of dividing up ancestral functions between them, eliminating redundancies and hence inhibiting the rate of pseudogene formation. It is further suggested that subfunctionalization is in many cases a transition state to neofunctionalization.

Here we wish to envisage DDC as a process driven by mutations in the sequence–encoding regions, adopting enzyme physical chemistry as an explicit template for illustration. Deoxyribonucleoside kinase [4] serves as a classic enzymatic example of this type of evolution, where other examples are also known [5].

Our starting point in the section below is a simple idealized mapping between enzyme sequence and function. We use this in Section 3 to construct a model sequence–fitness landscape describing an enzyme having pleiotropic function. In Section 4, referring to this landscape, we discuss neutral sequence evolution of a population responding to fixation of a duplication event. We show in this approach how DDC effects an increase in the mean lifetime of a duplicate.

## 2. A sequence–function mapping for enzymes

A natural choice of quantitative measure determining enzyme function is the rate  $v$  in solution of substrate to product

\* Corresponding author. Tel.: +46-8-55584043; fax: +46-8-55584199.  
E-mail address: chefnb@sbc.su.se (F.N. Braun).

conversion. In the limit of low substrate concentration [S], this is given according to the classical Michaelis–Menten theory by [6]

$$v = k_{\text{cat}} K[E][S], \quad (1)$$

where [E] is the enzyme concentration,  $K$  is the association constant governing reversible formation of enzyme–substrate complexes in the solution, and the catalytic constant  $k_{\text{cat}}$  is the first-order rate of chemical conversion of these to enzyme–product complexes.

The idea in the following is to explicitly relate the association constant  $K$  to enzyme sequence. To this effect, we adopt a highly idealized lattice-like view of the enzyme’s binding pocket, or ‘loop’.

Let us say that there are  $N$  amino-acid residues in the substrate binding loop,  $n$  of which are hydrophobic (H) and  $N - n$  of which are polar (P). We characterise the unliganded loop thermodynamically according to a Gibbs free energy

$$G_0 = ng_{\text{H}} - TS_0 \quad (\text{unliganded state}). \quad (2)$$

The first term expresses the hydrophobic effect, with  $g_{\text{H}}$  a free energy cost per H residue exposed to the polar environment of the solvent. In this unliganded state, we assume that all  $N$  residue sites of the loop are solvent-exposed. The entropy term derives from the different possible ways of arranging the  $N$  residues on the lattice

$$S_0 = k \ln \left[ \frac{N!}{n!(N-n)!} \right], \quad (3)$$

where  $k$  is Boltzmann’s constant.

Next we assume that the presence of a ligand in the loop has the effect of burying  $\alpha$  of the loop sites, such that they are shielded from the solvent. Hydrophobicity drives the H residues into these buried sites, but at a cost in distributional entropy. For  $n_{\text{b}}$  buried H residues, with  $n_{\text{e}} = n - n_{\text{b}}$  remaining exposed, we have

$$S(\alpha) = k \ln \left[ \frac{\alpha!}{n_{\text{b}}!(\alpha - n_{\text{b}})!} \right] + k \ln \left[ \frac{(N - \alpha)!}{n_{\text{e}}!(N - \alpha - n_{\text{e}})!} \right]. \quad (4)$$

Let us also introduce an elastic stress  $g_{\text{stress}}$  per ligand-buried site, such that for given  $n_{\text{e}}$  the total free energy of the bound enzyme–ligand complex is

$$G(\alpha) = n_{\text{e}}g_{\text{H}} - TS(\alpha) + g_{\text{stress}}\alpha \quad (\text{liganded state}). \quad (5)$$

The effective binding free energy of the enzyme–ligand complex follows by minimising with respect to  $n_{\text{e}}$ ,

$$\Delta G_{\text{bind}}(\alpha) = G^{(\text{min})}(\alpha) - G_0, \quad (6)$$

and from this we obtain according to the standard definition

$$K = K_{\text{ref}} \exp \left[ \frac{-\Delta G_{\text{bind}}}{kT} \right], \quad (7)$$

where  $K_{\text{ref}} = 1M^{-1}$  sets the standard units (inverse molar concentration).

This implies an explicit dependence of  $K$  on hydrophobic composition of the loop, i.e.  $K \sim f(n/N)$ , which defines our sequence-function mapping.

In a simple extension, we introduce the ratio  $v_{\text{A}}/v_{\text{B}}$  as a measure of pleiotropic function. This defines the specificity of an enzyme in the presence of two competing substrates A and B, i.e. the degree to which it discriminates A versus B. Neglecting any difference between the respective catalytic constants  $k_{\text{cat}}$ , we have

$$\frac{v_{\text{A}}}{v_{\text{B}}} = \frac{K_{\text{A}}}{K_{\text{B}}} \times \frac{[\text{A}]}{[\text{B}]} = \exp \left[ \frac{-\Delta \Delta G}{kT} \right] \times \frac{[\text{A}]}{[\text{B}]}, \quad (8)$$

where

$$\Delta \Delta G(\alpha_{\text{A}}, \alpha_{\text{B}}) = \Delta G_{\text{bind}}(\alpha_{\text{A}}) - \Delta G_{\text{bind}}(\alpha_{\text{B}}). \quad (9)$$

### 3. Sequence-fitness landscape

In Fig. 1, we use the above mapping to construct a primitive pleiotropic sequence-fitness landscape. The mutation-accessible sequence space of the enzyme is just the one-dimensional range of possible loop HP compositions  $n = 0, \dots, N$ . For given  $n$ , fitness is either 1 (viable) or 0 (non-viable) according to whether various arbitrarily specified functional criteria are met.

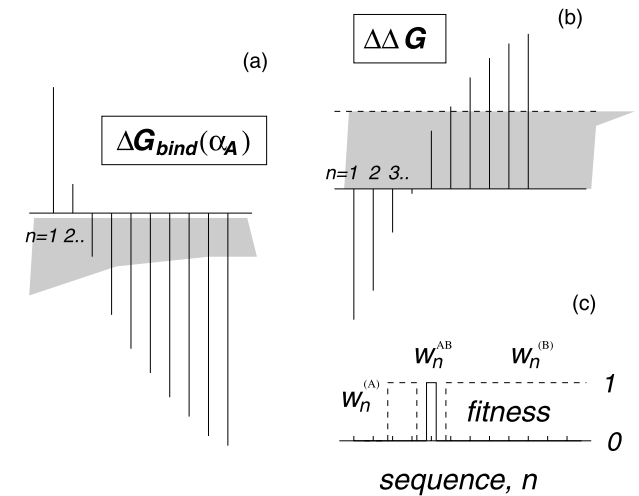


Fig. 1. Calculation of a pleiotropic fitness landscape in the 1D sequence space  $n$  of our approach, for model parameters  $N = 10$ ,  $\alpha_{\text{A}} = 3$ ,  $\alpha_{\text{B}} = 4$ ,  $g_{\text{stress}} = 2kT$ ,  $g_{\text{H}} = 4kT$ ; (a) absolute binding strength  $\Delta G_{\text{bind}}(\alpha_{\text{A}})$  for enzyme–substrate A complexes, with respect to a functional criterion  $\Delta G_{\text{bind}} < 0$  (shaded); (b) relative binding strength  $\Delta \Delta G = \Delta G_{\text{bind}}(\alpha_{\text{A}}) - \Delta G_{\text{bind}}(\alpha_{\text{B}})$ , with respect to a broad specificity pleiotropic criterion  $\{0 < \Delta \Delta G < kT\}$  (shaded); (c) strict pleiotropic fitness landscape  $w_n^{AB}$  (solid), and partial fitness landscapes  $w_n^{(A)}$ ,  $w_n^{(B)}$  (dashed). A partially fit enzyme is effective in binding either A or B, but not both simultaneously.

The solid contour  $w_n^{AB}$  of Fig. 1c follows by inspection of a pleiotropic criterion on  $\Delta\Delta G$  depicted graphically in Fig. 1b. An enzyme having  $w_n^{AB} = 1$  has full pleiotropic fitness, i.e. it is effective in binding both substrates, A and B.

The dashed partial fitness contours of Fig. 1c are significant in the DDC context we wish to implement. An enzyme having partial fitness  $w_n^{(B)} = 1$  discriminates excessively in favour of substrate B, such that it is ineffective in binding A. Conversely an enzyme having partial fitness  $w_n^{(A)} = 1$  is ineffective in binding B.

In constructing Fig. 1c, we have also to observe in general some absolute criterion on binding strength  $\Delta G_{\text{bind}}$ , corresponding to upper/lower constraints on absolute rate  $\nu$ . Inspection of an example of such an absolute criterion, depicted in Fig. 1a, results in a truncation of  $w_n^{(A)}$ .

#### 4. Sequence evolution

For eukaryotic genomes, the lifetime of a gene duplicate is thought to be typically a few million years [7]. Thus, the rate of silencing, say  $\lambda$  per gene, is comparable with the per-site fixation rates of neutral amino-acid mutations within a population [8].

With this in mind, let us consider a duplicate pair evolving neutrally across our pleiotropic fitness landscape. We define  $\theta$  as the per-gene rate at which neutral HP flipping mutations go to fixation within a population. Provided  $\theta$  is higher than  $\lambda$  (in fact, silencing occurs mechanistically through mutation to introduce a stop codon, frameshift, knockout transcription, etc. and must account for a subset of mutations [9]), then for each amino-acid substitution, an attempted duplicate silencing event occurs with probability  $\lambda/\theta$ . Silencing is subject to redundancy of one of the pair. That is, if neither is redundant, silencing is frustrated. This occurs when the pair form a subfunctionalized DDC complement  $w_n^{(A)} = 1$ ,  $w_n^{(B)} = 1$ . Thus, whereas an unfrustrated rate of silencing would result in a mean duplicate lifetime  $\lambda^{-1}$ , DDC-derived frustration leads in general to some longer *effective* mean lifetime  $\tau_{\text{eff}} > \lambda^{-1}$ .

We can write a simple relation linking  $\tau_{\text{eff}}$  to DDC explicitly,

$$\frac{\lambda}{\theta} \sum_{s=1}^{\theta\tau_{\text{eff}}} [1 - P_{\text{DDC}}(s)] = 1, \quad (10)$$

where  $P_{\text{DDC}}(s)$  is the probability that DDC has occurred as a function of the number of HP flipping substitutions  $s$  fixed in a population subsequent to a duplication event.

For the landscape of Fig. 1c, it is easy to see  $P_{\text{DDC}}(1) = 0$  and  $P_{\text{DDC}}(2) = 1/3$ . In a first approximation we can write

$P_{\text{DDC}}(s) \simeq P_{\text{DDC}}(2)$  for  $s > 2$ , yielding

$$\theta\tau_{\text{eff}} \simeq 1 + \frac{3}{2} \left( \frac{\theta}{\lambda - 1} \right). \quad (11)$$

Thus, in this approximation, if we assume for example an HP flipping rate of once every million years, i.e.  $\theta^{-1} \sim 10^6$  years, versus a silencing rate  $\lambda^{-1} \sim 5 \times 10^6$  years, then the mean lifetime of a pleiotropic enzyme duplicate evolving over our landscape is  $\tau_{\text{eff}} \sim 7$  million years.

A proper calculation of  $\tau_{\text{eff}}$  might proceed by averaging over repeated birth/silencing simulations, i.e. we initiate the simulation with a duplication event, and then cycle stochastically through the following: (i) one of the pair is randomly selected and an HP flipping substitution is randomly assigned along its sequence. The substitution is accepted provided the duplicate pairing remains functionally viable (i.e. either they maintain a subfunctionalized complement, or one of the pair maintains full pleiotropy). (ii) With probability  $\lambda/\theta$ , an attempt is made to silence one of the pair during the cycle. The attempt is successful and the simulation finishes if one of the pair is functionally redundant.

#### 5. Conclusion

In summary, we have presented a detailed model illustration of how DDC might lead to enhanced retention of enzyme duplicates. The mechanism driving DDC in this approach is neutral, constrained by a simplified physical chemistry of enzyme-ligand binding. Ultimately, related evolutionary simulations [10] incorporating a degree of positive selection at some proportion of amino-acid sites (reviewed in [11]) may be more appropriate for certain gene families [12]. Beyond enzymes, further examples of gene families retaining large numbers of duplicates include immune system genes [13] and olfactory receptors [14], the largest gene family in the human genome.

#### Acknowledgements

Funding for this work was provided by the Swedish Foundation for Strategic Research. FN Braun thanks CBU in Bergen for their hospitality during the final stages of preparation of the manuscript.

#### References

- [1] Ohno S. Evolution by gene duplication. Berlin: Springer; 1970.
- [2] Harrison PM, Echols N, Gerstein M. Nucl Acids Res 2001;29: 818.
- [3] Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. Genetics 1999;151:1531.
- [4] Knecht W, Sandrini MPB, Johansson K, Eklund H, Munch-Petersen B, Piskur J. EMBO J 2002;21:1873.

- [5] Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA. *Genome Biol* 2001;2(8):research0028.1.
- [6] Ferscht A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. New York: WH Freeman; 1999.
- [7] Lynch M, Conery JS. *Science* 2000;290:1151.
- [8] Kimura M. *Nature* 1968;217:624.
- [9] Hardison RC, Roshkin KM, Yang S, Diekans M, Kent WJ, Weber R. *Genome Res* 2003;13:13.
- [10] Siltberg J, Liberles DA. *J Evol Biol* 2002;15:588.
- [11] Liberles, DA, Wayne ML. *Genome Biol* 2002;3(6):reviews1018.1.
- [12] Gilad Y, Segre D, Skorecki K, Nachman MW, Lancet D, Sharon D, et al. *Nat Gen* 2000;26:221.
- [13] Nei M, Gu X, Sitnikova T. *Proc Natl Acad USA* 1997;94:9799.
- [14] Glusman G, Yanai I, Rubin I, Lancet D. *Genome Res* 2001;11:685.