SI Appendix

Thresholds for the Delineation of the Gene Age Classes

To determine the thresholds for the number of hits required to assign a protein to a particular age class all BLAST hits from a genome to the RefSeq database were analyzed in the following way. A distribution of the number of hits to different organisms was constructed for each category; the location of the rightmost peak of the distribution (rounded to an integer) was considered to be the "effective number of genomes" in a taxonomic category (e.g. see Figure A1).

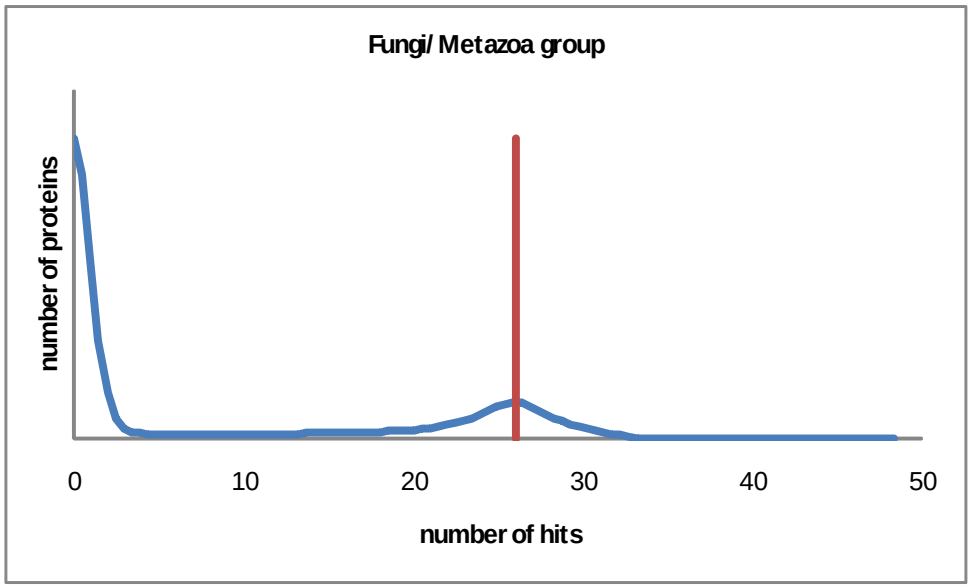

Figure A1. Distribution of the number of hits from Aspfu proteins to the Fungi/Metazoa group organisms in RefSeq database. Red bar shows the "effective number of genomes" in the Fungi/Metazoa group (26) from the Aspfu perspective.

The threshold for the number of hits required to assign a protein to the given age class was determined as a fraction of the effective number of genomes (Table A1).

Table A1.

Taxonomic Class	Effective No. of Genomes	Threshold No. of Hits
Homsa - Primates	2	1
Homsa - Mammalia	7	3
Homsa - Chordata	4	2
Homsa - Metazoa	13	6
Homsa - Fungi/Metazoa group	38	19
Homsa - Eukaryota	10	5
Homsa – cellular organisms	20	10
Drome - Drosophila	1	1
Drome - Diptera	3	1
Drome - Insecta	3	1
Drome - Metazoa	18	9
Drome - Fungi/Metazoa group	36	18
Drome - Eukaryota	10	5
Drome – cellular organisms	20	10
Aspfu - Aspergillus	3	1
Aspfu - Pezizomycotina	10	5
Aspfu - Ascomycota	12	6
Aspfu - Fungi	4	2
Aspfu - Fungi/Metazoa group	26	13
Aspfu - Eukaryota	10	5
Aspfu – cellular organisms	20	10

SI Text

Steady-State Model of Gene Gain and Loss

Consider a genome of a constant size under a steady-state process of gene acquisition and loss. Let $l(t,a,x)$ be the number of genes of the age a having the loss rate x in a genome at time t. Then

$$\partial l/\partial t + \partial l/\partial a = -xl, \tag{1}$$

The density function $l(t,a,x)$ should satisfy the boundary and initial conditions:

$$l(t,0,x) = Cg(x), \tag{2}$$

$$l(0,a,x) = \varphi(a,x). \tag{3}$$

The condition (2) means that new genes (of age zero and the loss rate x) appear with the rate that does not depend on time, and the total number of newly acquired genes is equal to some constant *C*.

We suppose also that the continuity condition is satisfied (see equation (7) below).

Proposition. The solution of model (1)-(3) is given by the formulas

$$l(t,a,x) = Ce^{-ax}g(x) \text{ at } t > a$$

and

$$l(t,a,x) = e^{-tx}\varphi(a-t,x) \text{ at } t < a.$$

Proof. The general solution of (1) is an arbitrary function of the first independent integrals of the system of ordinary differential equations

$$\frac{dl}{xl} = dt = da.$$

These integrals are $I_1 = t - a$, $I_2 = 2\ln l + (t+a)x$.

So, the general solution of equation (1) is of the form

$$l(t,a,x) = e^{-x(t+a)/2} F(t-a,x) \tag{4}$$

where $F(t,x)$ is an arbitrary (differentiable) function.

Then $l(t,0,x) = e^{-tx/2} F(t,x) = Cg(x)$,

and we should take $F(t,x) = Ce^{tx/2} g(x)$ at $t > 0$

to satisfy the boundary condition (2).

Next, $l(0,a,x) = e^{-ax/2} F(-a,x) = \varphi(a,x)$

and we should take $F(-a,x) = e^{ax} \varphi(a,x)$ at $a > 0$

to satisfy the initial condition (3).

Hence, the solution of system (1)-(3) is

$$l(t,a,x) = e^{-(t+a)x/2} F(t-a,x) = Ce^{-ax} g(x) \text{ at } t > a \tag{5}$$

and

$$l(t,a,x) = e^{-(t+a)x/2} F(t-a,x) = e^{-tx} \varphi(a-t,x) \text{ at } t < a \tag{6}.$$

To provide the continuity of the solution, we suppose that

$$\varphi(0,x) = Cg(x). \tag{7}$$

Corollaries

1. The density $l(t,a,x)$ does not depend on t at $t > a$.

2. The total size of a genome at $t > a$ is $L = \int_0^\infty \int_0^\infty l(t,a,x)\,da\,dx = C \int_0^\infty \frac{g(x)}{x} dx$.

3. The distribution of genes in a genome over loss rate

$$f(x) = \int_0^\infty l(t,a,x)\,da / L = \frac{g(x)}{x} / \int_0^\infty \frac{g(x)}{x} dx \text{ at } t > a$$

and the distribution of genes being lost at any given moment $t > a$ is

$$xf(x) = g(x)$$

(i.e. genes being lost have the same distribution of loss rate as genes being gained; effectively a condition of an equilibrium).

4. The distribution of genes in a genome over age

$$p(a) = \int_0^\infty l(t,a,x)dx / L = \int e^{-ax} g(x)dx / \int_0^\infty \frac{g(x)}{x} dx \text{ at } t > a$$

is proportional to the Laplace transform of $g(x)$.

It follows that, if the age distribution of genes in a genome $p(a)$ is known, then we can reconstruct the distribution $g(x)$ in the gene universe, and conversely. In practice, however, the data on $p(a)$ obtained in the present study are too crude to attempt a meaningful reconstruction of $g(x)$.

Example.

If $g(x)$ (the distribution of the loss rate among the genes of zero age) is a gamma-distribution with the shape parameter α>1, then $f(x)$ (the distribution of the loss rate across the whole genome) is also a gamma-distribution with the shape parameter α-1 (Figure A2) and the distribution of gene ages is a Pareto distribution with a power-law tail with an exponent of –α (Figure A3). Distributions of the loss rate among the genes belonging to different age classes could be quite distinct in this case (Figure A4).

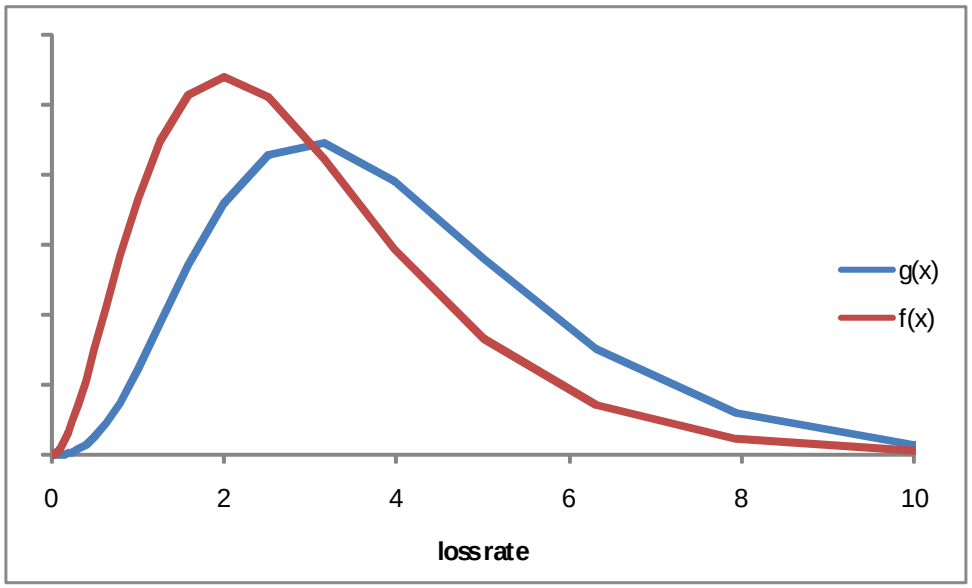

Figure A2. Distribution of the loss rate for the genes of zero age ($g(x)$; a gamma-distribution with the shape parameter α=4) and the overall distribution of the gene loss rate in the genome.

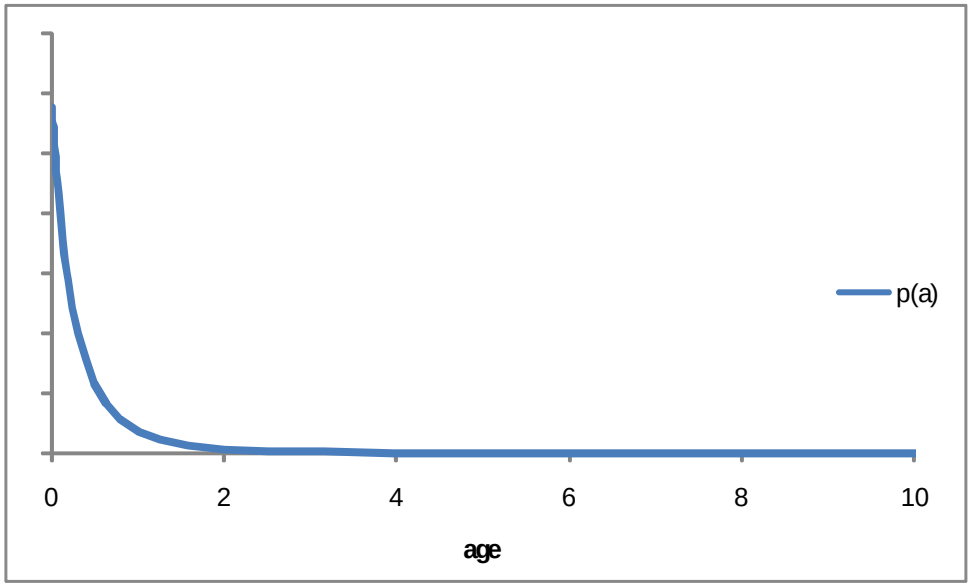

Figure A3. Distribution of the gene ages for the above model parameters.

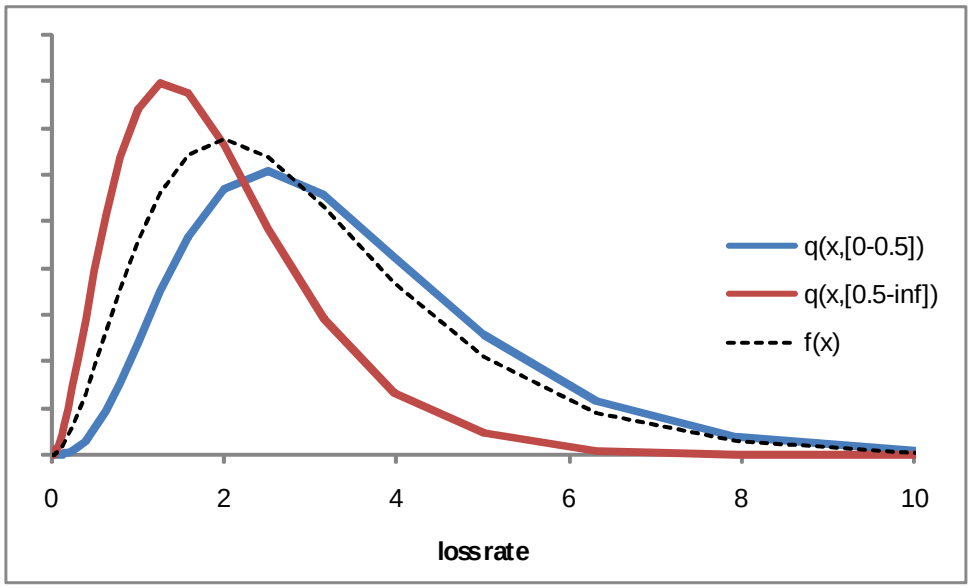

Figure A4. Distribution of the loss rate for the two age classes of genes: ages 0 to 0.5 and ages 0.5 to infinity.