PROCEEDINGS OF THE SIXTH INTERNATIONAL CONGRESS ON TWIN STUDIES

Rome: 28-31 August 1989

4.

https://doi.org/10.1017/S0001566000003640 Published online by Cambridge University Press

Acta Genet Med Gemellol 39: 419-425 (1990)

Sixth International Congress
on Twin Studies

A Stochastic Model of the Genetic Predisposition to Ageing: An Application to Twin Data

L. Gedda[1], G. Brenci[1], C. Rossi[2]

[1]*The Gregor Mendel Institute of Medical Research and Twin Studies, Rome, and*
[2]*Department of Mathematics, Second University of Rome, Italy*

Abstract. In previous papers a stochastic model of the ageing process has been proposed. Some genetic parameters (redundance, repair) have been used to explain the observed differential predisposition to the process and family heredity. Because the process is basically due to effective random mutations, any individual of the population would be predisposed differently to ageing according to the structure of his/her genome. In the present paper, the previous model is generalized to take into account an additional genetic parameter, namely, the stability against random mutations, defined as the probability that a random mutation in a codon would produce no mutation in the corresponding protein. Estimation problems connected with the model are approached on the basis of twin data in maximum likelihood estimation as well as in bayesian framework. Some comparisons between the two methods are reported.

Key words: Ageing, Twin data, Maximum likelihood estimation, Bayes' theorem

INTRODUCTION

In some previous papers [3,5] a stochastic model of the ageing process was proposed. On the basis of genetic parameters (redundance, repair), we proposed a model to take into account the differential predisposition to ageing and its inheritance. Relevant references to biomolecular studies on the subject were given at the time [5].

More recent suggestions allow to introduce a new genetic parameter to explain the differential predisposition to the ageing process. Indeed, when a random mutation occurs in a base, it can be effective or it cannot, according to the genetic code (Table 1).

Clearly, a random mutation of the third base of the codon CUC has no effect at all, since, whatever the third base is, the corresponding aminoacid will be leucine. Conversely, any random mutation of the third base of the codon AUG (methionine) will have an effect and produce a codon for isoleucine.

As the ageing process is basically due to effective random mutations, it seems evident that any given individual of the population will be predisposed in a different way to ageing according to the structure of its genome, ie, to its "stability" against random mutations. The stability of the genome for any given aminoacid can be defined, in a natural way, as the probability that a random mutation in the corresponding codon would have no effect at all. On this basis we can also define the stability of the genome for a protein as the mean stability of the chain.

Then, the model previously proposed [3] can be generalized to take into account this additional parameter. The problem of the inheritance of such parameter will not be considered here. The generalization of the model is trivial as we can introduce the stability of a protein (MS) to modify the parameter b according to:

$$b = b(1 - MS)$$

Then b is the intensity rate of the mutation process that takes into account the differential predisposition (MS) to ageing of any individual of a given population.

In the following, the probability distribution function of MS will be calculated and some inferential problems will be considered, related to MS, b or both.

We are proposing now, as in a previous paper [3], a mathematical model for the latent variable, which is commonly denoted by "frailty" [1]. The introduction of such an unobservable variable into biomedical models, particularly in survival models, is an attempt to take into account the observed heterogeneity among individuals, with respect to onset age of chronic and tumor diseases, responses to therapeutic protocols, and so on. A model of heterogeneity, based upon genetic factors, has been proposed elsewhere [4].

GENETIC STABILITY OF A PROTEIN

If we consider the genetic code (Table 1), we can define an index of stability for a protein in terms of the probability that a random mutation has no effect at all.

We can determine the quotient set, with respect to equal indices of stability, of the set of aminoacids, for which we can easily identify 6 different elements, ie:

$A_1 = \{leu, arg\}$
$A_2 = \{ser\}$
$A_3 = \{gly, ala, val, thr, pro\}$
$A_4 = \{ile\}$
$A_5 = \{lys, glu, asp, phe, asn, gln, tyr, cys, his\}$
$A_6 = \{met, trp\}$

https://doi.org/10.1017/S0001566000003640 Published online by Cambridge University Press

Table 1 - Genetic code

Leu	C U {U,C,A,G} / U U {A, G}	His	C A {U, C}
Arg	C G {U, C, A, G} / A G {A, G}	Tyr	U A {U, C}
Ser	U C {U, C, A, G} / A G {U, C}	Cys	U G {U, C}
Val	G U {U, C, A, G}	Lys	A C {A, G}
Pro	C C {U, C, A, G}	Glu	G A {A, G}
Thr	A C {U, C, A, G}	Asp	G A {U, C}
Ala	G C {U, C, A, G}	Gln	C A {A, G}
Gly	G G {U, C, A, G}	Asn	A A {U, C}
Ile	A U {U, C, A}	Trp	U G G
Phe	U U {U.C}	Met	A U G

We can then calculate the probability that we observe A_i if the codon bases are chosen at random. These probabilities are listed in Table 2 together with the observed relative frequencies of a sample of aminoacids (reported in Atlan [2]).

Table 2

Set	Probability	Frequency
A_1	17.6	11.8
A_2	8.6	8.1
A_3	32.8	32.7
A_4	5.2	3.8
A_5	32.8	40.4
A_6	3.4	3.1

We can compare the two distributions by means of regression analysis, and obtain $R^2 = 0.9433$, suggesting a good agreement between the two distributions.

If we define as an index of stability of class A_j the probability $P(A_j = A_j/\text{one random mutation}) = s_j$, we can calculate the following values:

$s_1 = 9/27$ $s_2 7/27$ $s_3 = 9/27$ $s_4 = 6/27$ $s_5 = 3/27$ $s_6 = 0$

On the basis of the above values we can then calculate the expectation of the index of stability for an aminoacid and the standard deviation. Using p and p′ as probability distributions for S, we obtain the following values:

$$p: \quad E(S) = 0.2374; \ \sigma(S) = 0.1080$$

$$p': \quad E(S) = 0.2228; \ \sigma(S) = 0.1101.$$

https://doi.org/10.1017/S0001566000003640 Published online by Cambridge University Press

By straightforward calculations we obtain the expectation of the mean stability of a protein containing N aminoacids chosen at random according to p or p′ and the standard deviation. We can define the mean stability of a protein as follows:

$$MS = \sum_{j=1}^{N} Si_j/N \quad i_j \epsilon \{1,2,3,4,5,6\}$$

$$P(i_j = k) = p_k \quad [\text{or } P(i_j = k) = p'_k]$$

Then we obtain $0 \leq MS \leq 1/3$, $E(MS) = E(S)$, $\sigma(MS) = \sigma(S)/\sqrt{N}$:

$$p: \quad E(MS) = 0.2374; \quad \sigma(MS) = 0.1080/\sqrt{N}.$$

$$p': \quad E(MS) = 0.2228; \quad \sigma(MS) = 0.1101/\sqrt{N}.$$

Considering, eg, $N = 250$ ($\sqrt{N} = 15.81$), we obtain:

$$p: \quad E(MS) = 0.2374; \quad \sigma(MS) = 0.0068$$

$$p': \quad E(MS) = 0.2228; \quad \sigma(MS) = 0.0070$$

This means that, for any protein, we can expect about 24% (22%) of the random mutations to be not effective since the beginning.

ESTIMATING THE GENETIC STABILITY OF A PROTEIN: MS

We can now estimate the effective mutation rate for each class A_j ($j = 1,2,\ldots,6$), neglecting the repair process [3]. If b is the mean mutation rate for aminoacid, the effective mutation rate for A_j can be easily calculated as follows:

$$b_j = b(1 - s_j)$$

The effective mutation process in a given protein is therefore a superposition of 6 mutation processes with intensities b_j ($j = 1,2,\ldots,6$) with weights q_j = proportion of A_j in the protein.

Then we observe a rate BE such that:

$$M = E(BE) = \sum_{i=1}^{6} b_i \cdot q_i = b(1 - MS)$$

https://doi.org/10.1017/S0001566000003640 Published online by Cambridge University Press

where the b_i are known and $\{q_i\}$ is an unknown probability distribution and our task is to estimate M. As N is usually very large ($\simeq 250$), we can use normal approximation for the random variable MS and so obtain a prior distribution, namely:

$$MS \sim N\ [E(MS),\ \sigma(MS)]$$

where, as an example, we can have:

$$E(MS) = 0.2374,\ \sigma(MS) = 0.0068 \quad [E(MS) = 0.2228,\ \sigma(MS) = 0.0070]$$

Then, we can calculate the likelihood function, namely, we can consider the Poisson approximation for the probability distribution function of the random variable BE with parameter given by M.

By straightforward calculation we obtain the posterior distribution for $MS = \theta$:

$$p(\theta|BE) = e^{-b(1-\theta)} \frac{[b(1-\theta)]^{BE}}{BE!}\ [2\pi\sigma^2(MS)]^{-1/2} \exp\left\{-\frac{1}{2}\left[\frac{\theta - E(MS)}{\sigma(MS)}\right]^2\right\}$$

We can then calculate the posterior mode to find a point estimate for θ (or we can calculate the posterior mean). Assuming that the likelihood approximation is accurate enough, we can also calculate the maximum likelihood estimate:

$$\hat{\theta} = 1 - BE/b; \quad \hat{M} = BE$$

ESTIMATING THE ENVIRONMENTAL MUTATION RATE: b

A protein for which MS is known can be conversely used for estimating b, if it is unknown ($b = \theta$). Taking as prior distribution for θ an exponential distribution with parameter α, we can easily calculate the posterior density for θ:

$$p(\theta/BE) = \alpha e^{-\alpha\theta} e^{-\theta(1-MS)} \frac{[\theta(1-MS)]^{BE}}{BE!}$$

which is a gamma function with respect to θ. We can then calculate the posterior mode θ^* (or mean) as well as the maximum likelihood for θ:

$$\theta^* = BE/(1 - MS + \alpha); \qquad \hat{\theta} = BE/(1 - MS).$$

We should note that actually $\alpha < MS$, so as to obtain feasible solutions for θ^* ($\theta^* \geq BE$).

https://doi.org/10.1017/S0001566000003640 Published online by Cambridge University Press

ESTIMATING THE MEAN STABILITY OF A PROTEIN: MS, AND THE ENVIRONMENTAL MUTATION RATE: b

The problem is much harder to solve if both parameters are unknown, namely, $MS = \theta_1$ and $b = \theta_2$. Assuming that the two unknown parameters are a priori independent and using the same prior density we have used before, we can easily calculate the posterior density function for the vector parameter θ:

$$p(\theta \mid BE) = f(BE \mid \theta) p_0(\theta_1) p_0(\theta_2) =$$

$$e^{-\theta_2(1-\theta_1)} \frac{[\theta_2(1-\theta_1)]^{BE}}{BE!} \cdot [2\pi\sigma^2(MS)]^{-1/2} \exp\left\{-\frac{1}{2}\left[\frac{\theta_1 - E(MS)}{\sigma(MS)}\right]^2\right\} \alpha e^{-\alpha\theta_2}$$

and we can calculate the posterior mode (or mean) to find point estimates.

It should be noted that, if we approach the problem by classical likelihood methods, we find two dependent likelihood equations, namely:

$$\begin{cases} -\theta_2 + BE/(1-\theta_1) = 0 \\ -(1-\theta_1) + BE/\theta_2 = 0 \end{cases} \qquad \rightarrow \theta_2(1-\theta_1) = BE.$$

Then the model cannot be identified in a classical framework, ie, the information given by prior distributions is essential, in some sense, to solve the estimation problem we are dealing with.

USING TWIN DATA

We can use data from monozygotic twins, living in different environments, to estimate the ratio between the two mutation rates, as follows. As the protein stability is the same for the two twins, we can consider it as a nuisance parameter, if it is unknown, so the parameters to be estimated are $b_1 = \theta_1$ and $b_2 = \theta_2$, while $MS = \tau$ is the nuisance parameter. We observe BE_1 and BE_2, then by straightforward calculations we can write:

$$\theta^* = BE_1/[\alpha_1 + (1-\tau^*)], \qquad \theta_2^* = EB_2/[\alpha_2 + (1-\tau^*)]$$

for the posterior modes as functions of the nuisance parameter. Then we get:

$$\theta_1^*/\theta_2^* = (BE_1/BE_2)[1 + (\alpha_1 - \alpha_2)/(\alpha_1 + 1 - \tau^*)].$$

https://doi.org/10.1017/S0001566000003640 Published online by Cambridge University Press

If $\alpha_1 \simeq \alpha_2$, we can write $\theta_1^*/\theta_2^* = BE_1/BE_2$. Otherwise we can integrate with respect to τ^*. Note that the estimate just obtained coincides with the maximum likelihood estimate of the unknown ratio.

The procedure can be easily generalized to take into account twins living in a different environment just for some periods of their lives.

REFERENCES

1. Aalen OO (1988): Heterogeneity in survival analysis. Stat Med 7:1121-37.
2. Atlan H (1972): Théorie de l'Information et Organisation Biologique. Paris: Hermann.
3. de Finetti B, Rossi C (1982): Mathematical models of mortality. In: Biological and Social Aspects of Mortality and the Length of Life. Liège: Ordina Editions, pp 315-29.
4. Gedda L, Brenci G (1969): Biology of the gene: The ergon/chronon system. Acta Genet Med Gemellol 18:329-79.
5. Rossi C (1972): Gene decay: Stochastic model of gene decay. Acta Genet Med Gemellol 21:191-196.

Correspondence: Professor Carla Rossi, Department of Mathematics, Second University of Rome, Via Fontanile di Carcaricola, 00133 Rome, Italy.