pANT: A Method for the Pairwise Assessment of Nonfunctionalization Times of Processed Pseudogenes

Sarel J. Fleishman,[1] Tal Dagan,†[1] and Dan Graur†*

Departments of *Biochemistry and †Zoology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Ramat Aviv, Israel

We present a method for pairwise Assessment of Nonfunctionalization Times (pANT) in processed pseudogenes. Contrary to existing methods for estimating nonfunctionalization times, pANT utilizes previously calculated probabilities of nucleotide substitution as explicit rate measurements, rather than assume that the substitution rates are the same for all nucleotides. Thus, the method allows a more accurate computation of the time that has elapsed since the nonfunctionalization of a pseudogene. Whereas existing methods require the sequence of an orthologous functional gene, which is not always at hand, pANT only uses the pairwise alignment of the gene/pseudogene pair, thus expanding the range of problems that can be tackled. To estimate evolutionary times in nonfunctional sequences, pANT measures the differences in the pairwise alignment of a gene and its paralogous processed pseudogene, using only the first and second codon positions. It assumes that, because of functional constraints, these positions in the sequence of the functional homolog have not changed since the time of nonfunctionalization of the pseudogene. Hence, the sequence of the gene may be used as the ancestor of the pseudogene. We show that the method's reliance on a detailed substitution matrix, which is derived separately for each species, makes it more accurate than existing methods. We applied pANT to the case of the unitary α-1,3-galactosyltransferase human pseudogene and found that our estimate of the nonfunctionalization time was in agreement with that obtained by taxonomic and paleontological considerations pertaining to the divergence between platyrrhines (New World monkeys) and cattarhines (Old World monkeys).

Introduction

Several probabilistic models for nucleotide substitution in noncoding DNA have been suggested in the literature, of which the one- and the two-parameter models (Jukes and Cantor 1969; Kimura 1980) are the simplest. These models assume that the rate of substitution from any one nucleotide to any other is constant in time. The one-parameter model assumes that the substitution rates of all four nucleotides are the same, and the two-parameter model assumes that all transversions occur at the same rate, as do all transitions. Under these models it is possible to calculate the probability that a given nucleotide site was substituted by another within a given period of time. The one-parameter model has been used to describe nucleotide dynamics when functional constraints are absent (Li, Gojobori, and Nei 1981).

The assumption of equal probabilities of substitution in the one- and the two-parameter models is, however, unrealistic. For instance, under this supposition, the nucleotide frequencies at equilibrium (i.e., after a sufficiently long period of time) are expected to equal each other (Nei 1984). This is in contrast with the preference for nucleotides A and T generally observed in noncoding segments (Graur and Li 1999). Other methods that are less sensitive to deviations from equilibrium frequencies have been proposed (e.g., Tajima and Nei 1984).

Nei (1987) showed that Kimura's two-parameter model might be generalized, at least in theory, by incorporating all 12 types of possible substitutions among nucleotides. Nei (1987) demonstrated that such a formulation leads to a different frequency of nucleotides at equilibrium and therefore will reflect the pattern of nucleotide substitution more closely. Nei's formulation (1987) could not be tested in practice when it was proposed because it required "a reliable nucleotide-substitution matrix based on a large amount of substitution data," which was then unavailable. Today, the availability of a large set of pseudogene sequences makes possible the determination of such a substitution matrix.

Based on the assumption that pseudogenes are not subject to functional constraints, Gojobori, Li, and Graur (1982) devised a method for calculating the relative rates of nucleotide substitutions in pseudogenes in the course of evolution. This method was used to calculate relative substitution rates in a large number of human pseudogenes (Graur and Li 1999). Here, we incorporate these substitution-rates data into a method of evolutionary-distance measurement as proposed by Nei (1987). Thus, we derive a method for pairwise Assessment of Nonfunctionalization Times (pANT) by using the alignment between a functional gene and a paralogous processed pseudogene.

Existing methods for estimating nonfunctionalization times (e.g., Li, Gojobori, and Nei 1981 and Miyata and Yasunaga 1981) rely on knowledge of three sequences: (1) a functional gene, (2) a paralogous pseudogene, and (3) an orthologous functional gene from another species as an outgroup. In calculating the nonfunctionalization times, these methods assume that all three sequences evolve at the same rate as long as they are functional. This assumption was shown to be unrealistic, both because different functional constraints may be at work and because rates are known to vary with organismic lineage.

The model we present here assumes that, because of functional constraints, the gene has undergone a negligible number of substitutions in its nonsynonymous codon positions—i.e., the first and second positions—since the divergence between the gene and the pseudogene (Nei and Kumar 2000). This assumption means, in essence, that the sequence of the functional gene may be treated as the

[1] The first two authors contributed equally to this work.

Key words: molecular evolution, substitution rates, pseudogenes, nonfunctionalization, computational method, galactosyltransferase.

E-mail: tali@kimura.tau.ac.il.

Mol. Biol. Evol. 20(11):1876–1880. 2003
DOI: 10.1093/molbev/msg202
Molecular Biology and Evolution, Vol. 20, No. 11,

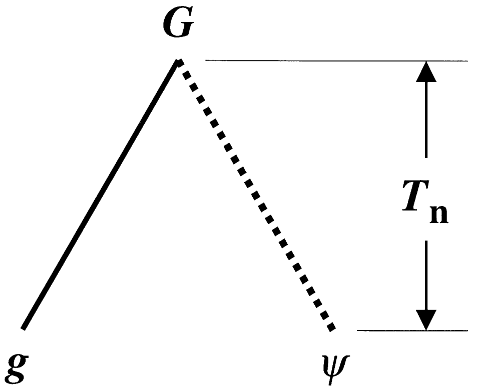

FIG. 1.—A tree for estimating the nonfunctionalization time of a processed pseudogene. G is the evolutionary parent of a contemporary gene (g) and its paralogous processed pseudogene (ψ). Because processed pseudogenes are "dead on arrival" (Graur and Li 1999), we assume that the nonfunctionalization time (T_n) is equal to the time of their creation through retroposition. The sequence of G cannot be inferred directly, but because of selective contraints it may be assumed that its nonsynonymous positions—i.e., its second and, to a lesser extent, its first codon positions—are identical to their positional homologs in g. In contrast, ψ has been under no selective constraint since its nonfunctionalization and has accumulated random mutations at a constant rate throughout the period T_n. The dashed line implies "nonfunctional."

pseudogene's forebear, and makes the use of an orthologous-gene sequence superfluous. Because pANT makes no use of a functional ortholog, we need make no assumption regarding the nucleotide-substitution rate in the orthologous species, thereby reducing inaccuracies.

pANT

Our approach assumes that the pattern of evolutionary change is the same at each site in the pseudogene sequence. Indeed, this assumption is not justified for coding genes, because of functional constraints and the different substitution patterns for the first, second, and third codon positions (Nei and Kumar 2000). However, in the case of pseudogenes, which are noncoding and devoid of any function, this assumption is justified.

We use the fact that processed pseudogenes and functional genes are subject to different nucleotide-substitution patterns. Let us assume the evolutionary scenario presented in figure 1. When G is the evolutionary progenitor of the contemporary gene g and the pseudogene ψ, we may assume that, because of functional constraints, the sequences of G and g in the nonsynonymous codon positions are the same. Because processed pseudogenes are "dead on arrival" (Graur and Li 1999), we assume that the nucleotide composition of ψ at the time of its nonfunctionalization was the same as that of G. Since its nonfunctionalization, ψ has been devoid of function and has accumulated substitutions at an accelerated rate compared to g.

In the following discussion, our derivation follows the general framework developed by Nei (1987). As in previous models, we assume that nucleotide substitution depends solely on the identity of the nucleotides being exchanged (A, T, C, or G), and that substitution occurs at a constant rate, according to the measured substitution probabilities (Graur and Li 1999). Let us denote as R the relative substitution-rate matrix (available at http://pant.tau.ac.il) obtained by Dr. Ron Ophir (cited in Graur and Li 1999) using 104 pseudogene sequences according to the method of Gojobori, Li, and Graur (1982). The pseudogene sequences were taken from various gene families and chromosomal locations, thus averaging the potential effects of substitution biases along the genome. Each element R^{ij} of this matrix corresponds to the probability that the nucleotide in column j would be replaced by the nucleotide in row i in the course of an arbitrary time unit. It is notable that the same analysis may be applied to any substitution-rate matrix obtained by the method of Gojobori, Li, and Graur (1982).

Under our assumptions, ψ diverged from g at the time of its nonfunctionalization. Let us mark by q_i^t the proportion of nucleotides in first and second codon positions of type i in sequence g at time t, which are identical in the sequence ψ. We mark by α_i the proportion of nucleotides, which, according to R, are other than i at time t, and are substituted for i at time $t+1$. The proportion of nucleotides i at time t that are substituted for other nucleotides at $t+1$ is marked by β_i. The proportion of identical nucleotides of type i at time $t+1$, q_i^{t+1}, can then be obtained:

$$q_i^{t+1} = (1-\beta_i)q_i^t + \alpha_i(1-q_i^t), \qquad (1)$$

where α_i is estimated as the average proportion of other nucleotides that are being substituted to i, $\alpha_i = \sum R^{ij}/3$, $j \neq i$ and $\beta_i = 1 - R^{ii}$.

Because $dq_i/dt = q_i^{t+1} - q_i^t$, then

$$dq_i/dt = \alpha_i - q_i(\alpha_i + \beta_i). \qquad (2)$$

Thus, our treatment considers eight different parameters, i.e., α and β for each of the four nucleotides. Solving equation (2) as an ordinary differential equation, and setting $q_i^{t=0} = 1$ (Boyce and DiPrima 1977), we obtain,

$$q_i = \frac{\alpha_i + \beta_i(e^{-(\alpha_i+\beta_i)t})}{\alpha_i + \beta_i}. \qquad (3)$$

Marking $(\alpha_i + \beta_i)t$ as d_i, defined as the evolutionary distance, and solving equation (3) for d_i, we obtain:

$$d_i = -\ln\left(\frac{q_i(\alpha_i + \beta_i) - \alpha_i}{\beta_i}\right). \qquad (4)$$

We thus obtain four different measurements for the distance d, one for each of the nucleotides. In the subsequent sections we estimate the distance d as the average of these four values:

$$d = \overline{d_i}. \qquad (5)$$

The pANT software is available online at http://pant.tau.ac.il.

Time Calibration

The evolutionary distance, d, obtained in equation (5) has arbitrary units of time. To calibrate these time units, we simulated the evolution of a pseudogene sequence according to the rates applicable to human pseudogenes (Graur and Li 1999) for different lengths of time, and compared these results to the distance obtained by pANT for these simulated sequences.

The time evolution of pseudogene sequences was simulated as spontaneous evolution of the mRNA of translation elongation factor 1α (GenBank accession number NM_001402). We used two parameters for the simulated evolution: (1) the substitution rate of human pseudogenes, i.e., 3.9×10^{-9} substitutions per year (Graur and Li 1999) and (2) the nucleotide-substitution matrix in Graur and Li (1999).

We designed the simulation so that each iteration corresponds to the spontaneous evolution occurring over the time span of 1 year. Thus, the number of iterations determines the nonfunctionalization time in years of the simulated pseudogene. In each iteration we determined whether each nucleotide i in the sequence undergoes mutation by sampling values from a uniform-probability distribution p_i $(0 - 1)$ and comparing these values to the yearly rate of substitution in pseudogenes. That is, if the condition $p_i < 3.9 \times 10^{-9}$ was satisfied, then the nucleotide i was a target for substitution.

Once a nucleotide was selected to undergo mutation, we sampled another value from a uniform-probability distribution, q_i $(0 - 1)$, to determine the new identity of the nucleotide at position i. The substituting nucleotide was determined by the sampled value in the cumulative probability function formed by the substitution matrix R.

The simulation times ranged from 0 to 400 Myr. We repeated the simulation 10 times for each time value (i.e., we produced 10 different pseudogenes for each nonfunctionalization time). For each resultant sequence we estimated its nonfunctionalization distance, d, using pANT. For comparison, nonfunctionalization times were also estimated using the method of Li, Gojobori, and Nei (1981).

We found good correlation between the evolution times being simulated and the pANT distances ($r^2 = 0.98$), and significant ($P \ll 0.05$) linear relation between the two variables. In fact, linearity is maintained for many time ranges. Using simple linear regression with the pANT distances as the predictor and the simulation evolutionary times as the predicted variables, we obtained a regression coefficient of 555 ± 1 Myr, which implies that each distance unit derived from pANT is equivalent to about 550 Myr (fig. 2). Considering the significant linearity and the good correlation between the pANT distance and the simulation times, we can use the regression equation from the simulation results as a reliable converter of the pANT distance units into units of time.

To test the quality of the estimation by the method of Li, Gojobori, and Nei (1981), we needed an outgroup sequence, for which the rat translation-elongation factor 1α (GenBank accession number NM_033539) was used. The estimates using the method of Li, Gojobori, and Nei (1987) exhibited a lesser degree of correlation with the simulated evolutionary times ($r^2 = 0.71$). Furthermore, the values obtained by using this method are relatively inconsistent, as they show large errors for simulation repeats using the same parameters (fig. 2). Finally, the method of Li, Gojobori, and Nei does not exhibit linearity with the evolutionary simulated times.

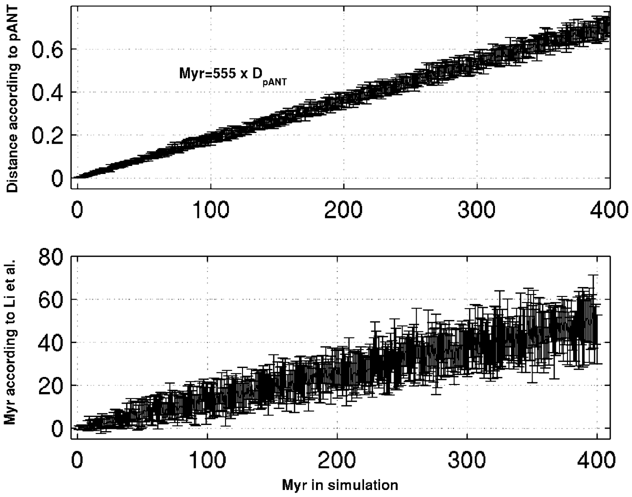

FIG. 2.—An error-bar plot of the simulation results. Each error bar shows the dispersal (i.e., mean ± SD) of the estimation results of 10 simulated pseudogenes that were simulated using the same parameter (i.e., simulated time). The upper plot shows the pANT estimation of the simulated pseudogenes' nonfunctionalization times. The slope of the linear regression gives an estimate of the actual time measured by one time unit in pANT, which is 555 ± 1 Myr. The lower plot shows the estimation of the nonfunctionalization times of the simulated pseudogenes using the method of Li, Gojobori, and Nei (1981). Note that this method underestimates the nonfunctionalization times by 80%–90% and results in large errors.

It is conceivable that biases in the substitution-rates matrix R may affect the nonfunctionalization-time estimates. The calculation of the R matrix in Graur and Li (1999) is based on 104 pseudogenes. To assess whether the choice of pseudogenes that are used for deriving the matrix R has a significant effect on our estimates, we recalculated the spontaneous-substitution matrix R for a group of 145 pseudogenes of the human ribosomal protein LA21 (Zhang, Harrison, and Gerstein 2002), which are distributed evenly among and along the chromosomes (the recalculated matrix R is available online at http://pant.tau.ac.il).

Using this recalculated substitution matrix, we conducted the evolutionary simulation of a pseudogene sequence. We then recalculated the linear-regression coefficient between the pANT results, using the substitution-rates matrix R taken from Graur and Li (1999) and the simulation results. This resulted in a linear coefficient of 544 ± 2 Myr, which is extremely close to the value we obtained by using the evolutionary simulations based on the matrix in Graur and Li (1999). This indicates that the pANT results are relatively insensitive to biases that may be present in the data sets.

Case Study

We applied pANT to studying the nonfunctionalization time of the α-1,3-galactosyltransferase (α1,3GT) unitary processed pseudogene, whose nonfunctionalization time was estimated using fossil data (Joziasse et al. 1991; Goodman et al. 1998; Koike et al. 2002). Most mammals express the cell-surface carbohydrate epitope galactose-α1,3-galactose (αGal). Catarrhines (Old World monkeys, apes, and humans), which lack the α1,3GT enzyme because of nonfunctionalization of the gene, are an exception. With the loss of αGal epitopes, the synthesis of which is dependent on an enzyme product of the α1,3GT gene, cattarhines produce anti-αGal antibodies that are responsible for the hyperacute rejection of organs transplanted from αGal-positive donors (Koike et al. 2002).

Based on the above phylogenetic information, the nonfunctionalization of the α1,3GT pseudogene was estimated to have occurred in the common ancestor of catarrhines, i.e., ∼40 MYA (Joziasse 1991; Goodman et al. 1998; Koike et al. 2002). In contrast, Galili and Swanson (1991) used the substitution patterns in genes and pseudogenes to estimate that the nonfunctionalization event occurred after the divergence of apes and monkeys, i.e., 17–25 MYA. Their estimate implies that the nonfunctionalization event occurred independently in three different lineages.

Because the human α1,3GT pseudogene is unitary—i.e., it has no functional homolog in human—we used a homologous sequence from marmoset (GenBank accession number AF384428). The sequences were aligned using ClustalW (Thompson, Higgins, and Gibson 1994), and gaps were eliminated from further consideration. Third codon positions were also excluded from the analysis. The nonfunctionalization-time estimate using the method of Li, Gojobori, and Nei (1981) requires an orthologous sequence, for which we used the α1,3GT from rat (GenBank accession number NM_145674).

The distance obtained using pANT was 0.083; multiplying this value by the factor derived from the evolutionary simulation (555 Myr) yields an estimate of 46 Myr, in agreement with the estimate based on paleontological evidence (Joziasse 1991; Koike 2002). Similar results were obtained by using the substitution-rates matrix R derived from the ribosomal LA21 pseudogenes (Zhang, Harrison, and Gerstein 2002). Using the method of Li, Gojobori, and Nei (1981) resulted in an estimate of 12 Myr, which is significantly lower than the fossil-based estimates, and would imply multiple independent losses of α1,3GT functionality.

Discussion

We used relative rates of nucleotide substitution, which were directly computed from pseudogene-sequence data, to calculate the time since nonfunctionalization of processed pseudogenes. Our formulation produces more accurate results than previous probabilistic models. The alternative Li, Gojobori, and Nei method (1981) makes two assumptions. The first is that the substitution rates are equal in the paralogous and orthologous genes. Present knowledge indicates that such assumptions are untenable even when dealing with closely related taxa such as different primate species. Another problem with the method of Li, Gojobori, and Nei (1981) is that, to estimate the nonfunctionalization time rather than the evolutionary distance, one must use an independent estimation of the speciation time between two species, and such estimates are known to be problematic (e.g., Shaul and Graur 2002). In fact, our simulations show that the method of Li, Gojobori, and Nei (1981) tends to underestimate nonfunctionalization times by about 80%–90% (fig. 2).

pANT's reliance on only two sequences allows its application even in cases in which other methods cannot be used. An additional factor that increases the accuracy of pANT with respect to other methods is its use of a substitution-rates table that was derived specifically for a particular species (Gojobori, Li, and Graur 1982), and reflects a particular pattern of nucleotide substitution. We can therefore use it in cases, such as bacterial pseudogenes, in which the mammalian pattern of substitution is most certainly inappropriate. Furthermore, the use of a simulation of evolutionary time in order to calibrate pANT's time units sidesteps the problem of using unreliable measures of speciation times, and opens the way to assessing nonfunctionalization times in species whose phylogenetic position is unknown.

Simulations that we conducted on a model sequence showed that the number of substitutions saturates after about 250 Myr, at a level of ∼70% sequence identity (data not shown). Thus, below 70% identity, pANT and other methods which rely on nucleotide distances should be considered unreliable. However, considering that pseudogenes are generally identified using a 60%–70% identity threshold (e.g., Zhang, Harrison, and Gerstein 2002), this limitation does not exclude a large number of potential pseudogene sequences.

The pANT estimate for the nonfunctionalization time of the α1,3GT was in agreement with estimates based on nonmolecular data. Indeed our estimate is somewhat larger than that reported by Goodman et al. (1998) in their assessment of the divergence time of platyrrhines and cattarhines (∼40 Myr). However, it was shown that estimates of divergence times using secondary calibration points should be regarded as indicative of the time *range* of the divergence rather than its exact value (Shaul and Graur 2002). In contrast, the 15-Myr estimate of Galili and Swanson (1991) and the 12-Myr estimate obtained by applying the method of Li, Gojobori, and Nei (1981) are implausible because they would imply multiple pseudogenization events in independent lineages. Given the findings of Koike et al. (2002) to the effect that rhesus, orangutan, and human share two identical molecular defects in their orthologous α1,3GT pseudogenes (a frameshifting deletion in pseudoexon 7 and a nonsense mutation in pseudoexon 9), a date of less than 40 Myr would require us to "believe as many as six impossible things before breakfast."

The various genome-sequencing projects have revealed many processed pseudogenes (and numerous others will most certainly be discovered), for which assessment of orthology or paralogy may be difficult. With pANT, we may estimate the nonfunctionalization times of these pseudogenes by using only the pairwise alignment of the pseudogene with any functional homolog.

Acknowledgments

We are grateful to the following people for critical review of the manuscript: Yuri Kagan, Rotem Sorek, Giddy Landan, Itay Mayrose, Shaul Shaul, Amit Huppert, Einat Hazkani-Covo, and Ron Ophir. We thank Ran Blekhman and Daphna Kritz for constructing the Web site.

Literature Cited

Boyce, W. E., and R. C. Diprima. 1977. Elementary differential equations. Wiley, New York.

Galili, U., and K. Swanson. 1991. Gene sequences suggest inactivation of α-1,3-galactosyltransferase in catarrhines after the divergence of apes from monkeys. Proc. Natl. Acad. Sci. USA **88**:7401–7404.

Gojobori, T., W. H. Li, and D. Graur. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:360–369.

Goodman, M., C. A. Porter, J. Czelusniak, S. L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C. P. Groves. 1998. Towards a phylogenetic classification of primates based on DNA evidence complemented by fossil data. Mol. Phyl. Evol. **9**:585–598.

Graur, D., and W. H. Li. 1999. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.

Joziasse, D. H., J. H. Shaper, E. W. Jabs, and N. L. Shaper. 1991. Characterization of an α-1,3-galactosyltransferase homologue on human chromosome 12 that is organized as a processed pseudogene. J. Biol. Chem. **266**:6991–6998.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–32 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Koike, C., J. J. Fung, D. A. Geller, R. Kannagi, T. Libert, P. Luppi, I. Nakashima, J. Profozich, W. Rudert, S. B. Sharma, T. E. Starzl, and M. Trucco. 2002. Molecular basis of evolutionary loss of the α-1,3-galactosyltransferase genes in higher primates. J. Biol. Chem. **277**:10114–10120.

Li, W. H., T. Gojobori, and M. Nei. 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292**:237–239.

Miyata, T., and T. Yasunaga. 1981. Rapidly evolving mouse alpha-globin-related pseudo gene and its evolutionary history. Proc. Natl. Acad. Sci. U.S.A. **78**:450–453.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, New York.

Ophir, R. 1999. The contribution of mutational factors and functional constraints to the molecular evolution of murids and humans. Doctoral thesis, Tel-Aviv University.

Shaul, S., and D. Graur. 2002. Playing chicken (*Gallus gallus*): methodological inconsistencies of molecular divergence date estimates due to secondary calibration points. Gene **300**:59–61.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Zhang, Z., P. Harrison, and M. Gerstein. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res. **12**:1466–1482.

Manolo Gouy, Associate Editor

Accepted June 16, 2003