J Mol Evol (1988) 28:131–135

Journal of Molecular Evolution

© Springer-Verlag New York Inc. 1988

Evolution of Protein Inhibitors of Serine Proteinases: Positive Darwinian Selection or Compositional Effects?

Dan Graur[1] and Wen-Hsiung Li[2]

[1] Department of Zoology, George S. Wise Faculty of Life Science, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel
[2] Center for Demographic and Population Genetics, University of Texas Health Science Center, PO Box 20334, Houston, Texas 77225, USA

Summary. In at least two instances involving serine proteinase inhibitors it has been shown that functionally important sites evolve faster and exhibit more interspecific variability than functionally neutral sites. Because these phenomena are difficult to reconcile with the neutral theory of molecular evolution, it has been suggested that the accelerated rate of amino acid substitution at the reactive sites is brought about by positive Darwinian selection. We show that differences in the amino acid composition in the different regions of proteinase inhibitors can account for the differences in the rates of amino acid substitution. By using an index of protein mutability [D. Graur (1985) J Mol Evol 22:53–62], we show that the amino acid composition of the reactive center in the ovomucoids and *Spi-2* gene products is such that, regardless of function, they are expected to evolve more rapidly than any other polypeptide for which the rate of substitution is known. In addition, the reactive region in the *Spi-2* proteins is shown to be free of compositional constraint. Positive Darwinian selection need not be invoked at the present time in these cases.

Key words: Rate of amino acid substitutions — Amino acid composition — Serine proteinase inhibitors — Ovomucoids — *Spi-2* — Neutral theory — Positive Darwinian selection — Serpins — Kunitz-type inhibitors — Kazal-type inhibitors

Introduction

The neutral theory of molecular evolution (Kimura 1968, 1983) asserts that the majority of molecular changes in evolution are selectively neutral. Purifying selection weeds out deleterious mutations, and those few mutations that eventually become fixed in a population are selectively neutral or nearly so. To date, positive Darwinian selection has been implicated only in rare cases at the molecular level (Schreier et al. 1981; Shepard and Gutman 1981; Klein and Figueroa 1986). One major corollary derived from the neutral theory is that functionally less important proteins or parts of proteins evolve faster than the more important ones. Clearly, then, we expect that in orthologous proteins, residues at functionally important positions should be less varied than those at other positions. Recently, however, two, or possibly three exceptions to these rules have been observed, all in serine proteinase inhibitors (for a recent review of proteinase inhibitors, see Barrett and Salvesen 1986).

Hill and Hastie (1987) determined the nucleotide sequences of two serpins in rat and compared one of the sequences (1.8 kb, *Spi-2.1*) with its orthologs in mouse and human. They divided each sequence into three homologous regions: regions 1 and 3, flanking the reactive center in region 2, on the 5′ and the 3′ sides, respectively. They found out that the segment containing the reactive center evolved faster relative to the other two regions. The segment containing the reactive center may, in fact, be evolving at a rate higher than pseudogenes. This cannot be stated categorically, however, because region 2 is very short, which means that the rate estimate has a large standard error.

Laskowski et al. (1987a,b) sequenced the third domain of ovomucoids (Kazal-type inhibitors) from 112 avian species. They found out that the sites that are in contact with the the enzyme (the presumed active sites) show a greater interspecific, and sometimes intraspecific, heterogeneity than the other sites.

Offprint requests to: D. Graur

From their compilation we can see that there are, on average, 5.72 alternative amino acids occupying a contact site, as opposed to an average of 2.21 alternative amino acids at a site that is not in contact with the target molecule. In a sense, Laskowski et al.'s (1987a) findings are more troubling for the neutral theory of molecular evolution than those of Hill and Hastie (1987), because unlike the situation in the *Spi-2* gene products, there is no doubt about the functional specificity of the highly variable positions in the ovomucoids. Their function as sites of contact between the enzyme and the inhibitor had been studied directly by means of x-ray crystallography (Read and James 1986).

A third case of accelerated evolution at a functional site may be represented by the bovine pancreatic trypsin inhibitor (BPTI) and its close homolog, spleen inhibitor II (SI), both Kunitz-type inhibitors. These two proteinase inhibitors differ at 7 of their 58 amino acid residues, 3 of which are among the 12 residues in the enzyme–inhibitor contact site as determined by x-ray crystallography (Fioretti et al. 1985; Creighton and Charles 1987). BPTI and SI will not be dealt with here, because these two inhibitors are clearly paralogous and gene conversion may have been involved in their evolution.

The spatial distribution and the rate of amino acid substitution in these families of proteinase inhibitors do not seem to fit the paradigm of the neutral theory of molecular evolution. Hence, in each case positive Darwinian selection has been invoked (Laskowski et al. 1987b). In the following, we would like to suggest an alternative explanation that is consistent with the neutral theory.

Data

Amino acid data for the avian ovomucoids were taken from Laskowski et al. (1987a) and Kato et al. (1987). Nucleotide sequence data for the rat serpins, the mouse contrapsins, and the human α_1-antichymotrypsin were taken from Chandra et al. (1983), Hill et al. (1984), and Hill and Hastie (1987). Despite the fact that these genes are most probably orthologous, the human and mouse proteins have different proteinase specificities (contrapsins are inhibitors of trypsin proteinases and α_1-antichymotrypsin inhibits chymotrypsins). Hence, they will be referred to in the following by their locus designation, *Spi-2*. The specificity of the *Spi-2.1* gene product is not known.

Data Analysis

Graur (1985a) suggested that the propensity of an amino acid to remain conserved in the course of evolution depends not so much on its being featured in active sites, but on an intrinsic stability index, defined as the mean chemical distance (Grantham 1974) between the amino acid and its mutational derivatives produced by a single nucleotide substitution. The stability index of amino acid i, S_i, is calculated as $\Sigma\, p_{ij}d_{ij}$, where p_{ij} is the probability of amino acid i being substituted by amino acid j when a single nucleotide substitution occurs at random, and d_{ij} is Grantham's chemical distance between amino acids i and j. For example, methionine (Met) changes to arginine (Arg), isoleucine (Ile), leucine (Leu), lysine (Lys), threonine (Thr), and valine (Val) with relative probabilities of 1/9, 3/9, 2/9, 1/9, 1/9, and 1/9, respectively (Nei 1975). Grantham's chemical distances between Met, on the one hand, and Arg, Ile, Leu, etc., on the other, are 91, 10, 15, etc., respectively. Therefore, S_{Met} is 38.67. By using, in addition to Grantham's distances, other measures of chemical similarity (e.g., Miyata et al. 1979), it has been concluded that four amino acids (i.e., cysteine, tryptophan, tyrosine, and glycine) are intrinsically highly immutable during evolution, i.e., their stability indices are more than 2 standard deviations above the mean for the entire group of 20 amino acids. These four amino acids plus serine constitute the conservative group. Seven other amino acids (i.e., leucine, methionine, phenylalanine, glutamine, isoleucine, histidine, and threonine) are expected to be substituted frequently. Consequently, the amino acid composition of a protein could be used in an objective fashion to predict its rate of substitution.

By using the frequencies of amino acids in 60 mammalian genes, Graur (1985a) fitted in a stepwise fashion a multiple regression equation that maximized the correlation between the observed rate of substitution and the predicted rate from data on amino acid composition. Graur used a forward inclusion process (Nie et al. 1975, pp. 321–342) with increasing numbers of amino acids, starting with the amino acid that explained the largest amount of correlation, namely glycine. These multiple linear regression equations, called the empirical indices of mutability, were denoted as I_m, where m is the number of amino acids used in the multiple regression equation. Obviously, by increasing the number of amino acids, the fraction of the total variation in rates of amino acid substitution also increased. However, with the increase in the ability to explain the variation, the statistical confidence in the results decreased rapidly. Graur (1985a) showed that to obtain reliable predictions, m should not exceed 7. The empirical index of mutability for seven amino acids was found to be:

$$I_7 = 0.841 - 5.096f_{Gly} + 24.145f_{Asn} - 26.807f_{Tyr} - 7.398f_{Val} + 18.219f_{Phe} - 8.263f_{Asp} + 7.960f_{Ile}.$$

Ten empirical indices of mutability, with which one can predict the rate of amino acid substitution from the amino acid makeup of a protein, are listed in the Appendix of Graur (1985a).

Amino acid composition, it was claimed, may be more important than functional constraints in determining rates of evolution of proteins or parts of proteins (Graur 1985a). It is the aim of this paper to test whether differences in amino acid composition could account for the differences in the rates at which the different domains within proteinase inhibitors evolve.

The Spi-2 Gene Products

We shall first examine the *Spi-2* gene products. Region 1, the slowest evolving segment, contains more conservative amino acids and less highly mutable amino acids than region 2, the fastest evolving part. The differences are small, but, at least qualitatively, region 2 has the potential to evolve faster than region 1 due to compositional constraints. To quantify the effect of the amino acid composition, we calculated a mutability index, I_7, for region 1, and regions 2 and 3 pooled together. The reason for pooling regions 2 and 3 is that both regions are too short to provide reliable information on their own. (There are only 13 and 27 aligned amino acids in the three species for regions 2 and 3, respectively.) Because region 3 also evolves much faster than region 1, the two rapidly evolving regions were pooled together. It must be pointed out, however, that even by using the two pooled regions together we still overstretch the applicability of the method somewhat, which was originally intended to predict the rate of substitution of much larger polypeptides. Preliminary results (Graur and Ticher, unpublished) show that I_7 rapidly looses its predictive validity on polypeptides shorter than 60 amino acids long. Nevertheless, the calculations for segments 2 and 3, separately, yield essentially the same results as for the pooled date and do not alter the conclusions.

I_7 for region 1 is 0.3346, a value comparable to what has been obtained for many proteins (e.g., hemoglobin, insulin). I_7 for regions 2 and 3, on the other hand, is 3.7558. This result far exceeds the I_7 values that have been previously calculated for any other functional protein. Because the I_7 value is linearly correlated with the rate of amino acid substitution, it follows that, regardless of function, regions 2 and 3, which include the reactive center of the proteinase inhibitor, are by virtue of their primary structure alone expected to evolve about 11 times faster than region 1. Hill and Hastie's (1987) calculations of nonsynonymous substitution rates by using the method of Li et al. (1985) show that region 2 evolves about 3–10 times faster than region 1. Because K_a, the number of nonsynonymous substitutions per site, could not always be determined for region 2, the above estimates are minimal values. Thus, the predictions derived from the amino acid composition agree with the empirical observations. In absolute terms, the rate of substitution in regions 2 and 3 is expected to exceed that of even the fastest evolving proteins, such as interferon γ and the fibrinopeptides. By using an alternative alignment (Yoon et al. 1987), the value of I_7 in region 1 comes out as 1.261, a little less than that for relaxin, one of the fastest evolving proteins. The alignment does not seem to make much difference in the results.

We were also concerned with the pattern of amino acid variability in the three regions of the *Spi-2* gene products, and whether or not any form of compositional constraint is evident in region 2. The test is based on the well-established fact that chemically similar amino acids are more interchangeable with each other than dissimilar ones (Clarke 1970; Jukes and King 1971; Gojobori et al. 1982; Graur 1985a,b). Following Dickerson and Geis (1983) we classify the 20 amino acids into five categories: acidic (aspartic acid and glutamic acid), basic (lysine, arginine, histamine), external-neutral (aspargine, glutamine), ambivalent (proline, threonine, serine, cysteine, alanine, glycine, tyrosine, tryptophan), and hydrophobic or internal-neutral (phenylalanine, leucine, isoleucine, methionine, valine). We, then, distinguish within each of the three regions between invariable sites, i.e., sites that are occupied by the same amino acid in the four proteins, and variable sites. If a variable site contains only amino acids from one of the categories defined above, we call this site a conservative site. Conservative sites indicate that a compositional requirement does exist, but that it is not very specific. For instance, a hydrophobic amino acid may be required at a particular site, but it does not matter which of the five hydrophobic acids is used. The pattern of amino acid variability in *Spi-2* proteins (Table 1) is quite illuminating. We see that all the sites in region 2 are variable, and none is constrained by composition. In contrast, both regions 1 and 3 show that even variable sites are in many instances constrained in amino acid composition. Interestingly, regions 1 and 3 differ from each other in the proportion of variable sites, but among the variable sites an equal percentage is constrained in composition.

The test for compositional constraint requires a valid alignment. In working with serpins it is generally easy to align regions 1 and 3, but very hard to align region 2 (M. Laskowski, personal communication). The Hill and Hastie (1987) alignment has not yet been tested, and it is possible that some or all of the postulated gaps in the alignment are incorrect. Nevertheless, by using an alternative align-

Table 1. Pattern of amino acid variability in regions 1, 2, and 3 of *Spi-2* proteins

Region	No. of aligned positions	No. of variable positions	No. of conservative positions
1	174	81 (46.6%)	30 (37.0%)
2	13	13 (100%)	0 (0%)
3	27	21 (77.8%)	7 (33.3%)

Table 2. Mean stability indices calculated as in Graur (1985a) for four categories of sites in avian ovomucoid third domains

Type of site	Number of sites	Mean stability index	SE	Expected range of indices
Invariable	14	122.40	11.80	38.67–168.14
Slightly variable	13	84.57	2.73	45.19–159.36
Variable	13	72.68	3.53	49.60–147.68
Hypervariable	11	72.67	3.38	53.70–129.69

ment (Yoon et al. 1987), the results come out the same.

Avian Ovomucoids

In respect to the third domain of avian ovomucoids it is not possible to perform the same comparison as in the case of the *Spi-2* proteins, because only 16 positions are identified as being in contact with at least one proteinase (Laskowski et al. 1987b). The interspecific data are so monumentally massive, however, that we can still check the importance of compositional versus functional constraints in determining the extent of amino acid variability at a site. We divided the 51 aligned sites within the third domain of avian ovomucoids into invariable (only 1 amino acid at a site), slightly variable (2 alternative amino acids), variable (3–4 alternatives), and hypervariable sites (5–9 alternatives). This division was decided upon so as not to have less than 10 sites in each category. For each of the categories the mean stability index (Table 2 of Graur 1985a) of the amino acids that occupy the site was computed. The results are presented in Table 2. Clearly, invariable sites are mostly occupied by highly conservative amino acids, whereas variable sites are not. The argument by Laskowski et al. (1987a), that the unvaried sites are structurally important but not functionally, is in agreement with the contention by Graur (1985a) that the mutability of an amino acid is determined not so much by its position, but by the consequences of its being replaced on the overall chemical and spatial structure of the protein.

Were the amino acids distributed at random with respect to the degree of variability at a site, the mean stability index would have been 83.48 for all four categories in Table 2. We see that the invariable sites show an index that is significantly higher than the expectation. A statistically significant deviation in the opposite direction is seen in the variable and hypervariable categories.

Discussion

There are two possible scenarios that can account for the evolution of inhibitors of proteinases. The one advanced by Hill and Hastie (1987) and by Laskowski et al. (1987a,b) envisions a substantial number of advantageous mutations occurring in the reactive centers of the inhibitors. These mutations had been frequently selected for by natural selection. Because the rate of substitution for advantageous mutations is expected to be much higher than that for neutral mutations, the problem seems solved, although the selective agent remains unknown (Brown 1987). Positive Darwinian selection, however, has an additional consequence, and that is a low degree of polymorphism within populations. The reason is that advantageous mutations tend to be fixed very rapidly within populations, so that they do not generate extensive transient polymorphism (Nei and Graur 1984). This prediction is probably not met in the present case. At least in the avian ovomucoid third domains, intraspecific polymorphism seems to be quite common (Laskowski et al. 1987b). Regrettably, at the present time, there are no systematic studies into the question of intrapopulational polymorphism in these genes.

We propose a different explanation for the phenomena. The data analysis indicates that the differences in the rates of amino acid substitution and consequently in the degree of interspecific variability are to a certain extent attributable to differences in the amino acid composition. We see that the amino acid composition in region 2 of the *Spi-2* gene product is extraordinary in the sense that most mutations are not expected to cause major disturbances in polarity, molecular volume, and electric charge. Consequently, region 2 is expected to mutate relatively free of constraints. The amino acid composition in region 1, on the other hand, resembles that of many ordinary proteins, and its rate of substitution is expected to be average. In ovomucoids we see the same picture. Highly variable positions are occupied by highly mutable amino acids, whereas 10 out of the 14 invariable sites are occupied by highly immutable amino acids. This proportion is typical of many proteins. For example, in a comparison between 13 cysteine proteinases belonging to the papain superfamily (data from Barrett et al. 1984), we find out that 21 out of the 33 invariable sites are occupied by highly immutable amino acids.

No significant difference in amino acid composition has been found between highly variable positions at contact sites and highly variable positions elsewhere. This means that the rate of amino acid substitution at a given site is independent of its being a functional site or not not.

Because no compositional constraints are evident at the enzyme–inhibitor contact region in inhibitors of serine proteinases, we propose that most observed substitutions are selectively neutral. The difference in the rates of substitution between the various regions of the protein results from the particular amino acid composition in these areas. This conclusion agrees with findings in other proteins (Graur and Ticher, unpublished). There may indeed have been several advantageous mutations in the segment containing the reactive part of the inhibitors, and these would further increase the rate of substitution in this region. Nevertheless, even without advantageous mutations the rate of substitution is expected to be quite high. We will have to invoke positive Darwinian selection only if the rate of substitution in region 2 of the *Spi-2* gene product is indeed larger than that in pseudogenes, and that has not been established with much confidence yet.

Our scenario has an additional corollary, which is that many of the serine proteinase inhibitors either do not have inhibition as their function (M. Laskowski, personal communication), or that inhibition does not depend on the composition of the sites that are in contact with the proteinase.

Acknowledgments. We thank Dr. Michael Laskowski, Jr. for carefully reviewing this note and for sending us a prepublication manuscript. This study was supported in part by a grant from the Foundation for Basic Research of Tel Aviv University and by NIH grant GM 30998.

References

Barrett AJ, Salvesen G (eds) (1986) Proteinase inhibitors. Elsevier, Amsterdam

Barrett AJ, Nicklin MJH, Rawlings ND (1984) The papain super family of cysteine proteinases and their protein inhibitors. In: Elödi P (ed) Proteinase action. Akadémiai Kiadó, Budapest, pp 203–217

Brown AL (1987) Positively Darwinian molecules? Nature 326: 12–13

Chandra T, Stackhouse R, Kidd VJ, Robson KJH, Woo SLC (1983) Sequence homology between human α_1-antichymotrypsin, α_1-antitrypsin and antithrombin III. Biochemistry 22: 5055–5060

Clarke B (1970) Selective constraints on amino-acid substitution during the evolution of proteins. Nature 228:159–160

Creighton TE, Charles IG (1987) Sequences of the genes and polypeptide precursors for two bovine protease inhibitors. J Mol Biol 194:11–22

Dickerson RE, Geis I (1983) Hemoglobin: structure, function, evolution and pathology. Benjamin/Cummings, Menlo Park CA

Fioretti E, Iacopino G, Angeletti M, Barra D, Bossa F, Ascoli F (1985) Primary structure and antiproteolytic activity of a Kunitz-type inhibitor from bovine spleen. J Biol Chem 260: 11451–11455

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

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Graur D (1985a) Amino acid composition and the evolutionary rates of protein coding genes. J Mol Evol 22:53–62

Graur D (1985b) Pattern of nucleotide substitution and the extent of purifying selection in retroviruses. J Mol Evol 21: 221–231

Hill RE, Hastie ND (1987) Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature 326:96–99

Hill RE, Shaw PH, Boyd PA, Baumann H, Hastie ND (1984) Plasma protease inhibitors in mouse and man: divergence within the reactive centre. Nature 311:175–177

Jukes TH, King JL (1971) Deleterious mutations and neutral substitutions. Nature 231:114–115

Kato I, Schrode J, Kohr WJ, Laskowski M (1987) Chicken ovomucoid: determination of its amino acid sequence, determination of the trypsin reactive site, and preparation of all three of its domains. Biochemistry 26:193–201

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Klein J, Figueroa F (1986) Evolution of the major histocompatibility complex. CRC Crit Rev Immunol 6:295–386

Laskowski M, Kato I, Ardelt W, Cook J, Denton A, Empie MW, Kohr WJ, Park SJ, Parks K, Schatzley BL, Schoenberger OL, Tashiro M, Vichot G, Whatley HE, Wieczorek A, Wieczorek M (1987a) Ovomucoid third domains from 100 avian species: isolation, sequences, and hypervariability of enzyme-inhibitor contact residues. Biochemistry 26:202–221

Laskowski M, Kato I, Kohr WJ, Park SJ, Tashiro M, Whatley HE (1987b) Positive Darwinian selection in evolution of protein inhibitors of serine proteinases. Cold Spring Harbor Symp Quant Biol 52 (in press)

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Miyata T, Miyazawa S, Yasunaga T (1979) Two types of amino acid substitution in protein evolution. J Mol Evol 12:219–236

Nei M (1975) Molecular population genetics and evolution. North Holland, Amsterdam

Nei M, Graur D (1984) Extent of protein polymorphism and the neutral mutation theory. Evol Biol 17:73–118

Nie NH, Hull CH, Jenkins JG, Steinbrenner K, Bent DH (1975) SPSS. McGraw-Hill, New York

Read RJ, James MNG (1986) Introduction to the protein inhibitors: x-ray crystallography. In: Barrett AJ, Salvesen G (eds) Proteinase inhibitors. Elsevier, Amsterdam, pp 301–336

Schreier PH, Bothwell ALM, Mueller-Hill B, Baltimore D (1981) Multiple differences between the nucleic acid sequence of the $IgG2^a$ and $IgG2a^b$ alleles of the mouse. Proc Natl Acad Sci USA 78:4495–4499

Shepard HW, Gutman GA (1981) Allelic forms of rat *k* chain genes: evidence for strong selection at the level of nucleotide sequence. Proc Natl Acad Sci USA 78:7064–7068

Yoon J-B, Towle HC, Seelig S (1987) Growth hormone induces two mRNA species of the serine protease inhibitor gene family in rat liver. J Biol Chem 262:4284–4289

Received October 22, 1987/Revised March 24, 1988