J Mol Evol (1990) 31:81–91

Journal of
Molecular Evolution

Codon Usage Changes and Sequence Dissimilarity between Human and Rat

Dominique Mouchiroud and Christian Gautier

Laboratoire de Biométrie, Génétique et Biologie des Populations (CNRS U.R.A. 243), Université Claude Bernard, 69622 Villeurbanne, France

Summary. This paper reports on the relationship between the number of silent differences and the codon usage changes in the lineages leading to human and rat. Examination of 102 pairs of homologous genes gives rise to four main conclusions: (1) We have previously demonstrated the existence of a codon usage change (called the minor shift) between human and rat; this was confirmed here with a larger sample. For genes with extreme C+G frequencies, the C+G level in the third codon position is less extreme in rat than in human. (2) Protein similarity and percentage of positive differences are the two main factors that discriminate homologous genes when characterized by differences between rat and human. By definition, positive differences result from silent changes between A or T and C or G with a direction implying a C+G content variation in the same direction as the overall gene variation. (3) For genes showing both codon usage change and low protein similarity, a majority of amino acid replacements contributes to C+G level variation in positions I and II in the same direction as the variation in position III. This is thus a new example of protein evolution due to constraints acting at the DNA level. (4) In heavy isochores (high C+G content) no direct correlation exists between codon usage change (measured by the dissymmetry of differences) and silent dissimilarity. In light isochores the opposite situation is observed: modification of codon usage is associated with a high synonymous dissimilarity. This result shows that, in some cases, modification of constraints acting at the DNA level could accelerate divergence between genomes.

Offprint requests to: D. Mouchiroud

Key words: Codon usage — Evolutionary rate — Isochores — Silent dissimilarity

Introduction

Genomes of warm-blooded vertebrates are characterized by a mosaic of DNA segments (>300 kb), the isochores, which belong to a small number of classes showing different C+G levels and a fairly homogeneous base composition (Bernardi et al. 1985; Ikemura and Aota 1988). Although the isochore class depends upon the overall C+G frequency in the segment, a strong correlation with the C+G frequency in position III of codons (CG_{III}) allows one to predict the isochore class of a gene from its coding part. The limits have been defined from the shape of the distribution of human CG_{III} and from some experimentally localized genes (Bernardi et al. 1985; Mouchiroud et al. 1987). Only three isochore classes have been considered here:

CG_{III}(human) $\leq$ 0.50: L1 (the lightest isochore)
0.50 < CG_{III}(human) < 0.77: L2 + H1 + H2 (intermediate isochore classes)
0.77 $\leq$ CG_{III}(human): H3 (the heaviest isochore)

These limits are somewhat arbitrary, due to the small number of presently localized genes. But slight modifications of these limits do not change the conclusions of this study (to verify the stability of the main results, 0.55 has been used as the L1 limit; all conclusions remained valid).

In murids, the mean C+G composition of genes is the same as in other mammals, but the variance is significantly lower (Bernardi et al. 1988; Mouchiroud et al. 1988). This difference (called the minor shift) implies that in many genes, the C+G con-

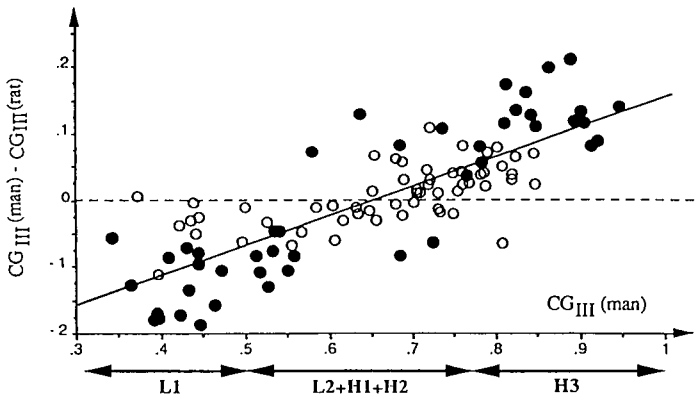

Fig. 1. C+G differences for third codon positions of homologous human and rat genes (human values minus rat values; ordinate) are plotted against the C+G level of third codon positions in humans (abscissa). Limits of the extreme isochore classes are shown. The minor shift is shown by the positive correlation, and the solid line is the regression line (using the least-squares method). Black points indicate genes for which the difference is significantly dissymmetric.

tent in the third codon position is quite different in murids compared to other mammals, particularly for genes with extreme C+G frequencies. In Fig. 1, we confirmed on a greater sample the existence of the minor shift between human and rat. A test for the dissymmetry of differences (see Mouchiroud and Gautier 1988) shows that 44% of genes are significantly affected by this codon usage change. If only L1 and H3 isochores are considered, this percentage increases to 59%. So some genes with extreme C+G content do not have coding usage change and this paper represents the first analysis of this variability.

Codon usage changes within a given gene family are due to a bias in the substitution process, which depends upon evolutionary tree branches. This bias allows sequences with different codon frequencies to be obtained from a common ancestor. Such nonstationarity in the evolutionary process was shown by Lanave et al. (1984) and Gautier (1987) in mammalian mitochondrial sequences. This paper addresses the relationship between this nonstationary process and the frequency of silent differences (SDF) between homologous genes. This simple parameter is one component of genetic distance between species. Processes that imply an increase in SDF could contribute to the divergence between genomes and so may play a role in speciation mechanisms. This feature justifies its study per se; moreover, it is clearly linked to a much more complicated parameter, which is the silent substitution rate. It has been claimed (Wu and Li 1985) that this rate is greater in the rodent lineage than in other mammals. A discussion of how minor shift and substitution rate are related is presented at the end of this paper.

Materials and Methods

To avoid any effect due to divergence time, we have limited this study to two species: a murid (*Rattus norvegicus*) and another mammal (*Homo sapiens*). These two species are those for which the number of homologous genes available is the largest. Complete coding sequences have been extracted from Genbank release 57 (Bilofsky et al. 1986) using the ACNUC retrieval system (Gouy et al. 1985). The list of the 102 genes used is given in the Appendix. However, due to the existence of multigenic families, the association of homologous sequences from two species is complex. The strategy used here separates three cases that are identified in the Appendix:

1) Only one sequence is known in rat and human for the studied gene and no reference exists regarding the possibility of multigenic organization: the two sequences are supposed to be homologous.
2) Several copies of the gene are known in at least one of the two species, these copies are very similar and no reference to multigenic organization is known: these copies are supposed to result from sequencing errors or polymorphism and so one of them is chosen at random. The great similarity between these copies (always more than 99%) cannot, however, modify the analysis. The only case known is noted by a + in the Appendix.
3) The gene is known to belong to a multigenic family. If several copies are known, the chosen pairing is the one that maximizes similarity between sequences. These cases, noted in the Appendix by an *, are not numerous (6 among 102), and their contribution to the present results is discussed below.

Each pair of genes was aligned using the algorithm of Smith et al. (1981). Dissimilarity between the two sequences was then quantified by the frequency of silent differences in position III of codons (SDF) and the percentage of nonsilent differences (NSD). More precisely:

$$\text{SDF} = \frac{\text{number of position III differences between synonymous codons}}{\text{number of synonymous codons}}$$

$$\text{NSD} = \frac{\text{number of amino acid differences}}{\text{number of amino acids}}$$

To analyze the link between variations of CG_{III} and codon usage changes, a dissymmetry index (D) has been defined as $D = \frac{|N_r - N_h|}{N_r + N_h}$ where N_r (resp. N_h) is the number of silent position III with C or G (resp. A or U) in rat and A or U (resp. C or G) in human. The range of variation of D is between 0 ($N_r = N_h$) and 1 when all differences that imply a modification of CG_{III} are in the same direction. Thus, if the ratio between the two directions is 2, then D equals 0.33, and if this ratio is 3, then D equals 0.5. If the substitution processes that have taken place in the two lineages are identical, then each of the $N_r + N_h$ differences has the same probability of falling into one of the two categories. In this case the conditional distribution of N_r, when $N_r + N_h$ is known, is a binomial law with a probability of 0.5. This is true whatever the processes are and particularly even if they are non-

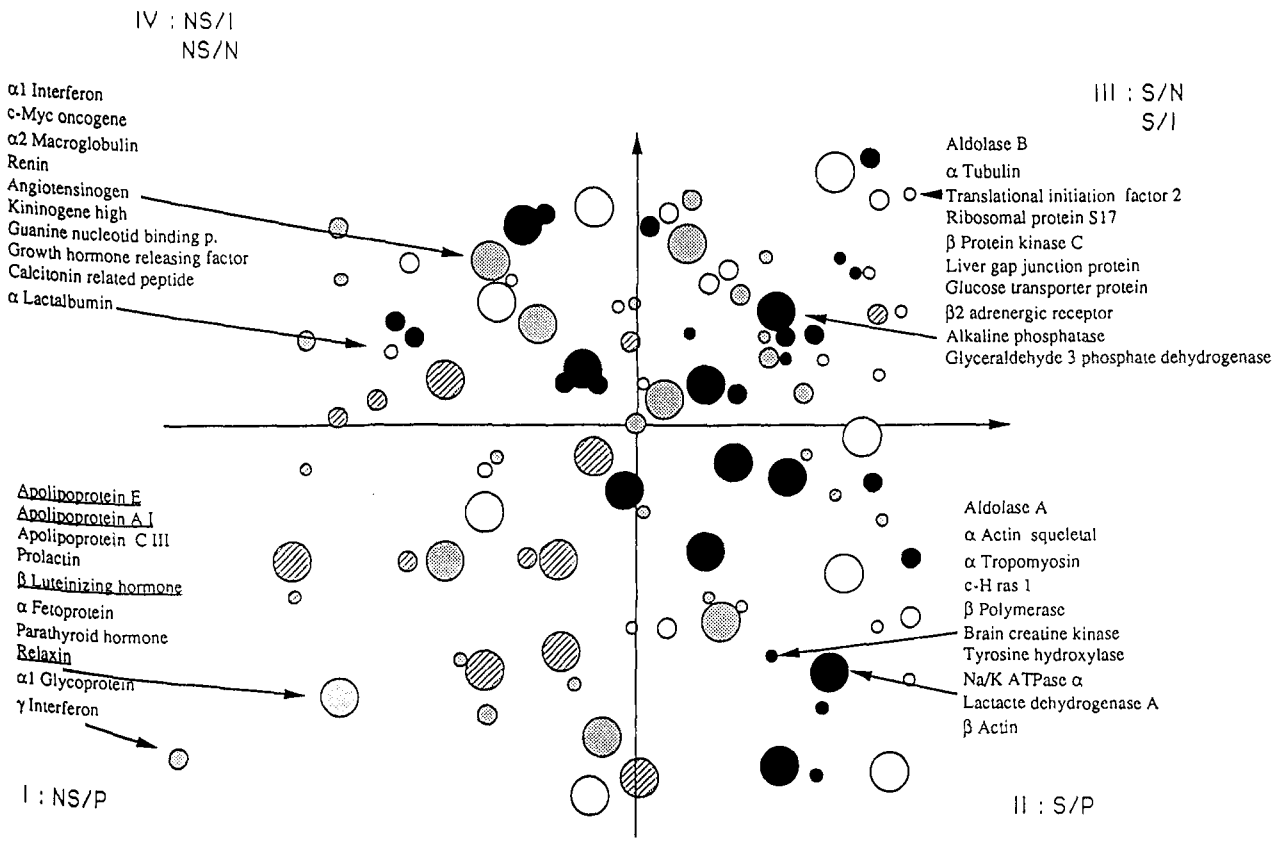

Fig. 2. Homologous genes are plotted using the values from the two first factors of the correspondence analysis, and circle sizes are a function of the synonymous rate values (small for SDF <0.27; middle for SDF between 0.27 and 0.31; large for SDF >0.31). Color pattern represents the nature of the proteins encoded: ■ enzyme; ▨ plasma protein; ▩ hormones growth factors; and □ structural proteins, other proteins. For each part of the factorial plane, the names of some genes have been mentioned. Genes with arrows are described more precisely in Table 1.

stationary. So process identity can be tested either directly using a normal approximation of D (Chessel and Gautier 1977) or by a χ^2 test with one degree of freedom (Mouchiroud and Gautier 1988). For simplicity this last test has been used here. It is important to note that D and SDF are not, a priori, functionally linked. This would not be the case if the difference of C+G in position III (ΔCG_{III}) had been used instead of D. Indeed ΔCG_{III} cannot be high if the number of changes is low.

If D has been specifically designed to handle codon usage changes linked to the variation in CG_{III} level, it cannot take into account all relationships between the substitution process and the minor shift. A more complete description could be achieved by combining the directionality of CG_{III} changes, the total amount of variation, and the silent/nonsilent ratio. To do that, differences have been distributed following two classifications. The first one separates silent from nonsilent differences, and the second one classifies differences into positive, negative, and indifferent ones. If differences imply a modification of the CG_{III} content, they are said to be positive if this variation is in the same direction as the overall variation in position III, and negative in the opposite case. Indifferent differences are those that do not imply a modification in the CG_{III} content. Crossing these two classifications leads to the definition of six numbers that describe the nature of the differences between the two sequences. To summarize this description, a correspondence analysis has been used (Benzecri 1973; Grantham et al. 1980). This method draws genes as points on a plane so that those that have similar frequencies for the six categories are neighbors. The two axes of the plane are the two main factors responsible for variability among the genes. SDF and protein classes have been graphically symbolized by size and darkness of points on the correspondence analysis plane (Fig. 2) in order to study their relationship within the two classifications.

Results

The main factor contributing to the variability that is evident after correspondence analysis is strongly correlated with nonsilent dissimilarity ($r = 0.97$). This is a somewhat trivial remark resulting from the great variability of protein homology among genes (see for example Li et al. 1987). The second main factor is correlated both with ΔCG_{III} ($r = 0.95$) and the percentage of positive differences ($r = 0.91$). This is not surprising and results directly from the existence of the minor shift. This modification of the isochore organization in murids implies codon usage change only in genes with extreme C+G content. The resulting variability between genes is detected by the correspondence analysis; however, the scattering of genes on the whole factorial plane was not expected (Fig. 2). Genes falling into each of the four parts of the factorial plane are observed (I: high codon usage change with low protein similarity; II: high codon usage change with high protein similar-

Table 1. Difference profiles of genes with all possible combinations of ΔCG_{III}, silent (SDF), and nonsilent (NSD) difference percentages

	Gene[a]	% differences in third codon position[b]						ΔCG_{III}[c]	NSD[d]	SDF[e]
		Sp	Si	Sn	NSp	NSi	NSn			
I	1. Relaxin	27	5	11	35	14	8	17.30	0.486	0.358
	2. γ-interferon	20	1	4	42	16	17	18.71	0.606	0.295
II	3. Lactate dehydrogenase	62	11	19	4	3	1	15.90	0.057	0.344
	4. Brain creatine kinase	63	11	12	2	7	5	11.50	0.068	0.222
III	5. Alkaline phosphatase	44	13	31	5	2	5	3.81	0.099	0.317
	6. Translation initiation factor	45	13	42	0	0	0	0.63	0.009	0.185
IV	7. Angiotensinogen	25	12	20	15	11	17	1.26	0.361	0.375
	8. α-lactalbumin	30	0	20	16	11	23	0.71	0.309	0.224

[a] Homologous genes located in the four parts of the plane (Fig. 2): part I, ΔCG_{III} and NSD high; part II, ΔCG_{III} high and NSD low; part III, ΔCG_{III} and NSD low; part IV, ΔCG_{III} low and NSD high. For each part, genes can have very different SDF values

[b] Percentage of the six types of differences in the third codon position. S = silent, NS = nonsilent, p = positive, i = indifferent, b = negative

[c] Absolute value of the difference between C+G content in position III of a codon in human and rat

[d] Silent difference percentage (SDF) is the number of synonymous codon differences in the third position per synonymous codons

[e] Nonsilent difference percentage (NSD) is the percentage of amino acid differences

ity; III: low codon usage change with high protein similarity; and IV: low codon usage change with low protein similarity). Moreover, as shown by point sizes in Fig. 2, gene positions on the factorial plane are independent of SDF. Existence of genes belonging to each of the eight possibilities (see Table 1) shows that no simple relation could exist between nonsilent difference percentages (NSD), codon usage change (ΔCG_{III}), and SDF.

To determine if the nature of encoded proteins plays a role in the positioning of the genes in the factorial plane, four protein categories have been defined: hormone, plasma protein, enzyme, and others. As seen in the Appendix the last group is in fact less homogeneous even if a majority of its sequences are structural ones. Figure 2 shows that these groups have different protein similarity, and this is confirmed by testing for equality of mean NSD in each group (enzyme, 0.15; hormone, 0.23; plasma protein, 0.26; and others, 0.13; $P < 0.01$). On the contrary no significant difference exists between SDF of each protein category (enzyme, 0.30; hormone, 0.29; plasma protein, 0.30; others, 0.27). So no argument appears to favor a particular role for protein nature in the silent substitution process in the rat and human lineages. However, we note that, surprisingly, no enzyme gene shows a codon usage change and a weak protein similarity simultaneously (part I of Fig. 2).

In the murid lineage, genes have been submitted to an evolutionary process that is not directly linked to the specific information carried by these genes. This suggests the possibility of a conflicting situation between this evolutionary process and natural selection acting to maintain the gene functions. Such an example has already been discussed in mammals for mitochondrial genomes (Gautier 1987). A first step in that study is to determine if the process implies modifications in the encoded proteins. Mouchiroud et al. (1988) have already cited examples of such modification in the murid lineage, but no systematic analysis has yet been done.

To determine if changes in nonsilent codon positions are also affected by the same substitution process as are silent positions, a χ^2 test was developed (Mouchiroud et al. 1988). Amino acid changes between homologous genes are classified according to C+G modification in the first and second codon positions (0 G/C, 1 G/C, or 2 G/C). Positive and negative amino acid differences are defined in a way similar to those for silent differences. The amino acid differences associated with a C+G change in the two first positions in the same direction as in third codon position are positive changes. If the direction is the opposite one, the amino acid differences are said to be negative ones. The number of positive amino acid changes is compared to the number of negative changes by a χ^2 test.

This test was made for the 28 homologous genes that have a clear codon usage change ($\Delta CG_{III} > 0.10$). These genes are located in parts I and II of Fig. 2. Among them, 21 show changes in the first and second positions oriented in the same direction as those in the third codon position, and eight of them lead to a significant value of the test. The repartition of these eight genes according to the percentage of NSD is the following: 3/9 for NSD >25%; 5/7 for NSD comprising between 10% and 25%; and 0/5 for NSD <10%. In this last case, the number of amino acid changes is not sufficient to do a test. A relationship between C+G content and amino acid composition of proteins has been reported (Bernardi and Bernardi 1986; Hanai and Wada 1988): in an organism, proteins coded by genes that

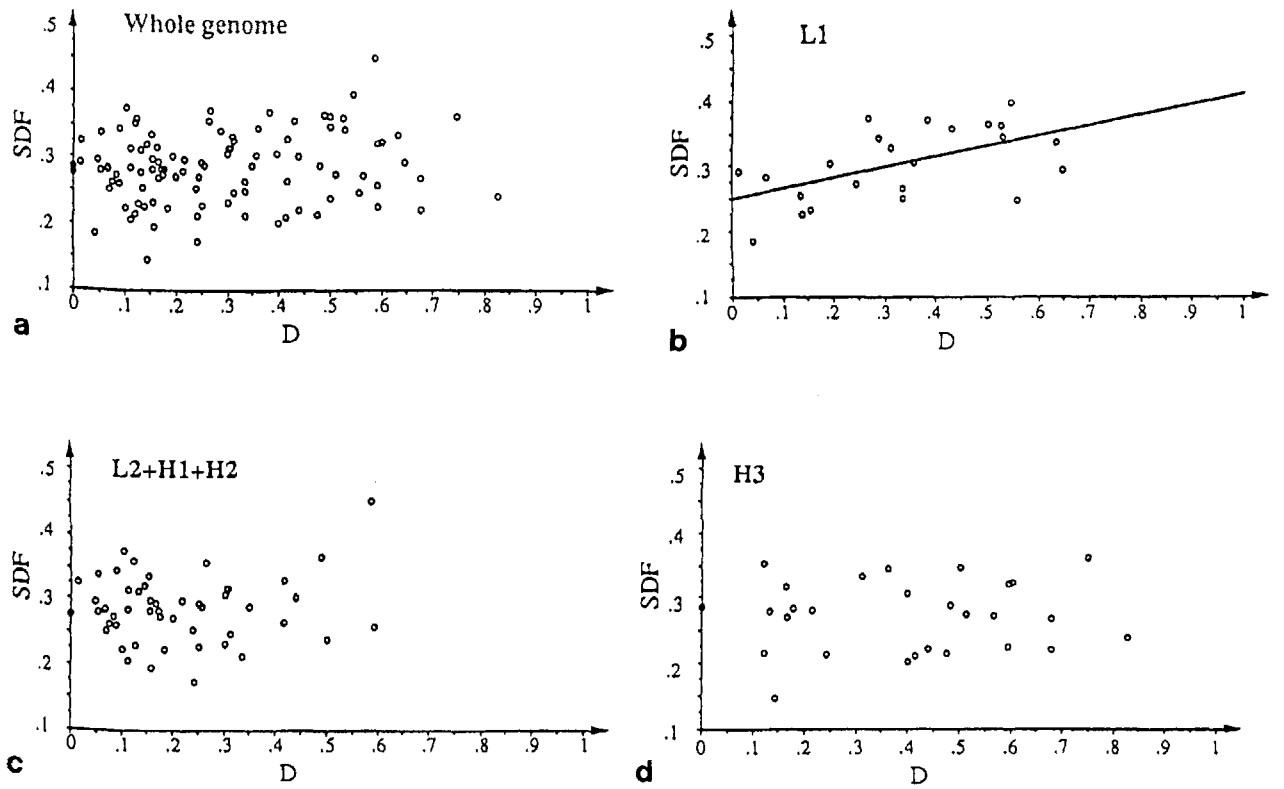

Fig. 3. Synonymous difference frequencies (SDF) are plotted against the dissymmetry index D (see Materials and Methods). **a** When all genes are considered simultaneously, no relation appears between D and SDF. **b–c** When the genes belonging to the same isochore are considered separately, a linear relation appears only in lightest isochore L1.

belong to different isochore classes have generally different amino acid compositions. This result has been linked to thermal stability of the protein (Argos et al. 1979). The present paper extends these results to the evolution of the same gene in the lineage of the two species. As already suggested, the preceding results directly demonstrate that protein evolution could be constrained by processes acting at the DNA level, independent of the function of this protein. A modification of base composition in a genomic region can orient changes in all codon positions of a gene and thus can lead to amino acid substitutions.

Discussion

Protein Nature and Minor Shift

Correspondence analysis revealed four typical evolutionary patterns. Each of them corresponds to different relationships between protein and DNA evolution.

In part I, some hormones or plasma proteins (Fig. 2) show codon and amino acid changes simultaneously. We may assume that either these genes are subject to fewer or weaker selective constraints because their function does not require a very specific protein structure or that they must adapt to a quickly evolving environment. For example, genes encoding the α-fetoprotein and apolipoprotein have already been discussed in the literature. For the first gene, Minghetti et al. (1985) have shown that changes involve amino acids with different chemical properties and give rise to loss of structural features such as disulfide bonds. The authors' conclusion is that the majority of changes are indifferent to natural selection. For the apolipoprotein family, the structural requirement of lipid binding is not stringent and a high rate of evolution may occur with considerable change in primary structure without impairment of function (Luo and Li 1986). We note that all genes belonging to the apolipoprotein family are located in part I. So they have undergone codon usage changes although they are not linked on the same chromosome. No enzyme shows a similar evolutionary feature.

However, it must be noted that local changes in base composition are not always linked to a modification of amino acid composition. For genes in part II, the oriented changes are limited to silent codon position. These genes seem to be submitted to high selective constraints because their function requires very specific protein structures. This is in agreement with the nature of the genes concerned. These genes encode mainly enzymes or structural proteins. Many studies have shown the strong con-

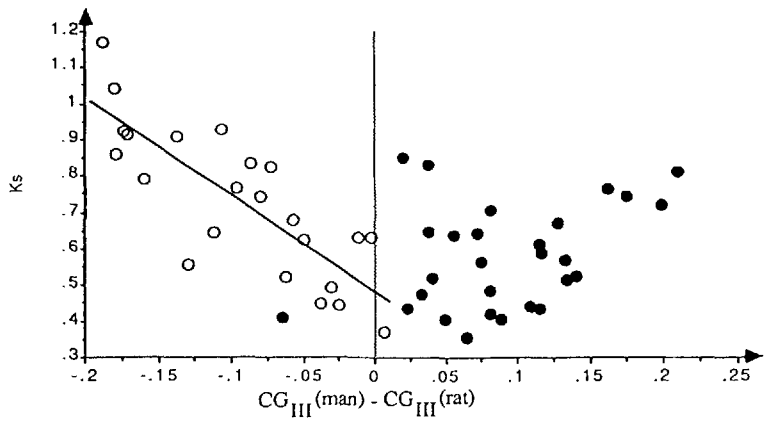

Fig. 4. Only genes belonging to H3 (black points) and to L1 (white points) are represented. The synonymous substitution rate (Ks) is plotted against the C+G differences in third codon position between human and rat. Ks is the measure of silent substitution rate proposed by Li et al. (1985). The result is the same as that obtained with the percentage of dissimilarity (SDF). Only genes in L1 show a linear correlation. Due to the minor shift, the abscissa segregates black and white points. Only one gene, met1 (metallothionein) belonging to H3, shows a greater CG_{III} in rat than in human. This gene is known to be multigenic (see Appendix).

servation of amino acid composition of some gene families such as actin (Erba et al. 1986), kinase (Buskin et al. 1985), and aldolase (Sakakibara et al. 1985), for example. For this last gene family, it is interesting to note that two genes encoding, respectively, aldolase A and B are in different parts of the factorial plane. The first gene shows codon usage change (part II) and the second one does not (part III). The aldolase A gene belongs to H2 (CG_{III} human = 0.73), and the aldolase B gene belongs to L2 (CG_{III} human = 0.59). So it seems that after divergence from a common ancestor, the genes have been scattered in different isochores.

For genes in part III, functional constraints are the same as for the genes in part II, but no codon usage change is observed. The genes concerned also encode enzymes and structural proteins. The fact that homologous genes have the same codon usage indicates that either they have an intermediate CG_{III} level or they are strongly conserved despite an extreme CG_{III} level (Fig. 1). Some of these genes are housekeeping genes (necessary for the life of cells) such as tubulin, ribosomal proteins, or some enzymes. But this is not a rule, as some housekeeping genes show codon usage change (kinase, actin, part II).

For the last gene category (part IV), the variability of the base composition in the three codon positions is the same as the variability for genes of part I, but changes are not oriented. These genes are located in DNA regions with an intermediate C+G content. This evolutionary pattern is followed by some genes belonging to each protein category. We note that α-interferon belongs to part IV and γ-interferon to part I. So in the interferon gene family, the evolutionary pattern is not the same between genes.

Base Composition and Evolutionary Rate

In the introduction we have emphasized the biological significance of silent dissimilarity; however, this is linked to another evolutionary parameter: the silent substitution rate. In the present sample, SDF and Ks [one of the more classical measurements for silent substitution rate (Li et al. 1985)] are strongly correlated ($r = 0.80$) on a clearly linear scatter diagram (not shown here). The main variability factor with regard to the regression line seems to be due to the different methods used in counting silent differences (see Materials and Methods). So our results could be reformulated using substitution rate instead of dissimilarity. Figure 4 shows that the relationship between dissymmetry of differences and silent dissimilarity actually leads to a relationship between codon usage change and Ks. This allows for a discussion of some previously published results.

1) Ticher and Graur (1989) have shown a strong negative correlation between the C+G level in the third codon position and Ks. This correlation results from individual correlations of each base frequency (0.70 for A_{III} and -0.65 for C_{III}). These correlations seem to be partly due to the sample used, as their values are weaker on the larger sample used here (0.31 for A_{III} and -0.29 for C_{III}). Moreover, the increase of Ks when a codon usage change occurs in L1 and the absence of such a relation in H3 seems to be sufficient to explain the significant correlation between C+G content and Ks ($r = 0.31$, $P < 0.01$). So it appears that the complex mutational process proposed by Ticher and Graur is not necessary to explain their results.

2) Wu and Li (1985) have shown an increase of Ks in rodent lineage. The results presented here are only relative to two species and so do not allow a discussion of the global variation of Ks. However, the increase of Ks for genes belonging to L1 that have been submitted to a codon usage change suggests that the minor shift could be one of the factors that participated in the Ks increase. A new examination of this increase, taking into account the isochore class of each gene family, is necessary to estimate the exact contribution of the modification of isochore organization in murids.

However, analysis of results presented requires a discussion of the biological significance of substitution rate estimations and particularly of Ks. Lanave et al. (1984) have emphasized the importance of a stationary substitution process. As already discussed in the literature [see in particular Preparata and Saccone (1987) and Blaisdell (1985)], the estimators are based on a Markov process for which several constraints have been placed on the transition matrix. This approach has several drawbacks:

1) The characteristics imposed upon the transition matrix determine more or less completely the limiting distribution (or stationary distribution) of the process. Particularly, a symmetric transition matrix (as used by Ks) implies a uniform limiting distribution. Very few mammalian genes have equal base frequencies and the existence of the isochore organization rules out the possibility that these deviations from uniform distribution could result from the random variability of the process.

2) A Markov chain is said to be homogeneous if the transition matrix is a constant. If, in addition, the process has reached its limiting distribution, it is said to be stationary. Lanave et al. (1984) have been the first to clearly state that the stationary hypothesis is necessary for all substitution rate estimations. The existence of the minor shift implies that the process is not even homogeneous. We note that the recent approach of Blaisdell (1985), which hypothesizes the existence of two homogeneous processes in each of the two lineages, may be a valuable one.

3) We think that the Markovian nature of the process has been insufficiently discussed in the literature. This hypothesis states that the knowledge of base frequencies at a given time t is sufficient to infer the distribution at any time $t + h$ ($h > 0$). In other words, all the necessary information is reduced to base frequencies in the gene of the ancestor species. This seems to be largely unrealistic, mainly in the case of silent substitutions. In fact, at a given time, a species is a set of individuals grouped into populations, each of which has its own polymorphism. The Markovian hypothesis is equivalent to the negation of the contribution of this strong and complex organization to evolution.

So interpreting the relationship between dissimilarity and codon usage change relative to substitution rate implies studies on the robustness of estimation procedures. Such analysis does not seem to have been done yet, and mathematical analysis of the direction of the bias under some alternative hypothesis is particularly lacking. Saccone et al. (1989) in a recent paper suggest that the acceleration in rodent lineage results only from such bias and that the silent substitution rate is essentially constant among genes validating the molecular clock hypothesis in this context (Zuckerkandl and Pauling 1965). The work of these authors is similar in part to the present one, yet it relies on a smaller number of genes (20 comparisons between human and rodents against 102 here). The test used to determine if the process is stationary is the one previously developed by the same group (Lanave et al. 1984). This test gives results similar to our dissymmetry test, allowing easy comparison of the two approaches. The POMC (Proopiomelanocortin) gene is the only gene for which the two tests lead to different conclusions: this gene is stationary according to Saccone et al. but is significantly dissymmetric according to the χ^2 test. The main difference in the results presented here is that Saccone et al. do not recognize the difference between L1 and H3 and we do. This difference can be explained by the nature of the sample used by these authors, which contains very few H3 genes (two stationary ones and two or four, following the table, nonstationary ones). If caution must be taken to associate substitution rate to dissimilarity or to some estimation like Ks when codon usage change occurs, other data support the hypothesis of acceleration of the silent substitution process in rodents (Catzeflis et al. 1987). Moreover, no clear reason seems to support the fact that the bias would be greater in L1 than in H3.

If new, more realistic estimation procedures are needed for substitution rate, dissimilarity seems to be a simpler parameter with a clear biological meaning. Its relationship to isochore organization shows that evolution of properties affecting the whole genome, independent of particular gene function, could increase the genetic distance between species. The mechanism that links dissimilarity to codon usage change has different effects in the two extreme isochore classes and this seems to be incompatible with a simple model, regardless of which frame is chosen (selectionism or neutralism). Wolfe et al. (1989) have proposed a model to explain the apparition of isochores in which the essential role is played by mutational pressure. Directional bias in mutation is due to the existence of two nucleotide pools corresponding to early and late replication stages. The first one is supposed to be rich in C+G and is used to replicate high isochore classes. The other one corresponding to heterochromatin is supposed to be rich in A+T. The relationships between isochore classes, chromosomal banding, and replication timing (see a review in Bernardi 1989) led Wolfe et al. to the conclusion that early replicating regions have been submitted to mutational bias that has increased their C+G content and given rise to high isochores, and that conversely late replicating regions have been submitted to the opposite bias. The most important point in this hypothesis is not the existence of the two nucleotide pools but the fact that they preceded the creation of isochores. So they do not result from biochemical adaptation to the replication needs, as

seems to be the case for tRNA in multicellular organisms relative to translation (Garel et al. 1970). Lack of adaptation, under this model, leads to two possible interpretations of the minor shift: either a new modification of the nucleotide pools in the murid (or rodent) lineage has decreased the difference between their C+G content, or a new constraint has appeared in addition to previous mutational pressure. In both cases the complexity of the model needs to be increased to take into account the dissymmetry between L1 and H3. Opposed to this point of view has been the idea that isochores have a functional role (Bernardi et al. 1988; Bernardi 1989). Selectionist models are much more difficult to translate into mathematical language (this is not a reason to consider that they are more complex ones). But it must also be recognized that present knowledge is not sufficient to relate the minor shift to some murid characteristics whether they be physiological (activity for example) or ecological (adaptive strategy, life history parameters).

To conclude we note that the increase of available data will allow comparisons among three species. Such comparisons could be tested if genes that have been submitted to codon usage change have some specific characteristics. One particularly important point is to detect if the greater dissimilarity of genes belonging to L1 is also present within these same genes between species having similar codon usage.

Acknowledgments. We thank Giorgio Bernardi, Cecilia Saccone, and Cecilia Lanave for discussion.

References

Argos P, Rossmann MG, Grau UM, Zuber H, Frank G (1979) Thermal stability and protein structure. Biochemistry 18:5698–5703

Benzecri JP (1973) L'analyse des correspondances. In: L'analyse des données, tome 2. Dunod, Paris

Bernardi G (1989) The isochore organization of the human genome. Annu Rev Genet 23:637–661

Bernardi G, Bernardi G (1986) Compositional constraints and genome composition. J Mol Evol 24:1–11

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, Swindell LD, Tung CS (1986) The Genbank genetic sequence data base. Nucleic Acids Res 14:1–4

Blaisdell BE (1985) A method of estimating from two aligned present-day DNA sequences their ancestral composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site. J Mol Evol 22:69–81

Buskin JN, Jaynes JB, Chamberlain JS, Hauschka SD (1985) The mouse muscle creatine kinase cDNA and deduced amino acid sequences: comparison to evolutionarily related enzymes. J Mol Evol 22:334–341

Catzeflis FM, Sheldon FH, Ahlquist JE, Sibley CG (1987) DNA–DNA hybridization evidence of the rapid rate of muroid rodent DNA evolution. Mol Biol Evol 4:242–253

Chessel D, Gautier C (1977) Des statistiques non paramétriques pour l'analyse des données binaires. Rev Stat Appl 25:57–73

Erba HP, Gunning P, Kedes L (1986) Nucleotide sequence of the human γ cytoskeletal actin mRNA: anomalous evolution of vertebrate nonmuscle actin genes. Nucleic Acids Res 14: 5275–5294

Garel JP, Mandel P, Chavancy G, Daillie J (1970) Functional adaptation of tRNAs to fibroin biosynthesis in the silkgland of *Bombyx mori*. FEBS Lett 7:327–329

Gautier C (1987) Codon usage changes during evolution: animal mitochondria example. C R Acad Sci (Paris) 304:123–128

Gouy M, Gautier C, Attimonelli M, Lanave C, di Paola G (1985) ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. CABIOS 1:167–172

Grantham R, Gautier C, Gouy M (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res 8:1893–1912

Hanai R, Wada A (1988) The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. J Mol Evol 27:321–325

Ikemura T, Aota S (1988) Global variation in C+G content along vertebrate genome DNA. Possible correlation with chromosome band structures. J Mol Biol 203:1–13

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93

Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li WH, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330–342

Luo CC, Li WH (1986) Structure and evolution of the apolipoprotein multigene family. J Mol Biol 187:325–340

Minghetti HP, Simon WL, Dugaiczyk A (1985) The rate of molecular evolution of α-fetoprotein approaches that of pseudogenes. Mol Biol Evol 2:347–358

Mouchiroud D, Gautier C (1988) High codon usage change in mammalian genes. Mol Biol Evol 5:192–194

Mouchiroud D, Fichant G, Bernardi G (1987) Compositional compartmentalization and gene composition in the genome of vertebrates. J Mol Evol 26:198–204

Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in humans and murids. J. Mol Evol 27:311–320

Preparata G, Saccone C (1987) A simple quantitative model of the molecular clock. J Mol Evol 26:7–15

Saccone C, Pesole G, Preparata G (1989) DNA microenvironments and the molecular clock. J Mol Evol 29:407–411

Sakakibara M, Mukai T, Yatsuki H, Hori K (1985) Human aldolase isozyme gene: the structure of multispecies aldolase B mRNAs. Nucleic Acids Res 13:5055–5069

Smith TF, Waterman MS, Fitch WM (1981) Comparative biosequence metrics. J Mol Evol 18:36–46

Ticher A, Graur D (1989) Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. J Mol Evol 28:286–298

Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ

among regions of the mammalian genome. Nature 337:283–285

Wu CI, Li WH (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166

Received July 21, 1989/Revised February 1, 1990

Appendix. List of homologous genes between human and rat

Gene[a]	No. of sites[b]	Human		Rat	
		Name[c]	References[d]	Name	References
Serum albumin	1824	ALBAF1	Gene 32:255–261 (1984)	ALBM	PNAS 78:243–246 (1981)
Alpha-2 macroglobulin	4410	A2M	PNAS 82:2282–2286 (1985)	A2M	JBC 262:446–454 (1987)
Actin beta cytoplasmic	1128	ACCYBA	MCB 5:2720–2732 (1985)	ACCYB	NAR 11:1759–1771 (1983)
Actin alpha skeletal muscle	1134	ACTASK	MCB 3:787–795 (1983)	ACSKA	Nature 298:857–859 (1982)
Alcohol dehydrogenase alpha class I	1128	ADH1CA	BIOCH. 25:2465–2470 (1986)	ADH	Gene 48:287–291 (1986)
Alpha-1 acid glycoprotein	600	AGP1A	Gene 44:127–131 (1986)	AGPA1H	JBC 260:4397–4403 (1985)
Aldolase A	1095	ALDA	BBRC 131:413–420 (1985)	ALDA	JBC 261:3347–3354 (1986)
Aldolase B	1095	ALDB	PNAS 81:2738–2742 (1984)	ALOBG2	JMB 181:153–160 (1985)
Alkaline phosphatase L/B/K	1575	ALPL02	JBC 263:12002–12010 (1988)	PHOA	PNAS 85:319–323 (1988)
*Alpha amylase pancreatic (amy 2)	1512	AMY201	Gene 60:57–64 (1987)	AMLS	Nature 287:117–122 (1980)
Alpha-fetoprotein	1566	ALBAF4	Gene 32:255–261 (1984)	AFPM	PNAS 78:3521–3525 (1981)
Alpha-lactalbumin	426	LACTA	BJ 242:735–742 (1987)	ALAC	Nature 308:377–380 (1984)
Angiotensinogen	1428	ANG	BIOCH. 23:3603–3609 (1984)	ANG2	JBC 259:8063–8065 (1984)
Apolipoprotein A-I	774	APOAI1	JBC 263:6857–6864 (1988)	APOA01	JBC 261:13268–13277 (1986)
Apolipoprotein C-III	288	APOAI1	JBC 263:6857–6864 (1988)	APOA02	JMB 187:325–340 (1986)
Apolipoprotein A-II	303	APOAII	JBC 260:15222–15231 (1985)	APOAII	JMB 187:325–340 (1986)
Apolipoprotein A-IV	1173	APOA4A	JBC 262:7973–7981 (1987)	APOA03	JBC 261:13268–13277 (1986)
+Apolipoprotein E	927	APOE3	BBRC 130:1261–1266 (1985)	APOEA	JBC 261:13777–13783 (1986)
Liver arginase	969	ARGL	PNAS 84:412–415 (1987)	ARGL	JBC 262:6280–6283 (1987)
Asialoglycoprotein receptor H1 major	852	ASGPR1	JBC 260:1979–1982 (1985)	RHL1	JBC 260:12523–12527 (1985)
Atrial natriuretic factor	450	ANFA	Nature 312:656–658 (1984)	ANF	JBC 260:4568–4571 (1985)
Na^+/K^+ ATPase alpha-subunit	3072	ATPAR	JB 100:389–397 (1986)	ATPA1	BIOCH. 25:8125–8132 (1986)
Na^+/K^+ ATPase beta-subunit	912	ATPBR	NAR 14:2833–2844 (1986)	ATPBSA	MCB 6:3884–3890 (1986)
Beta-2-microglobulin gene	360	B2M1	J.IMM. 139:3132–3138 (1987)	B2MR	NAR 15:7638–7638 (1987)
Beta-adrenergic receptor (beta-2)	1242	BAR	NAR 15:3636–3636 (1987)	ADBC	PNAS 84:8296–8300 (1987)
Bone gla protein (BGP)	297	BGPG	EMBO J. 5:1885–1890 (1986)	BGPR	EMBO J. 5:1885–1890 (1986)
Calcitonin	396	CALCR2	PNAS 82:1994–1998 (1985)	CAL2	MCB 4:2151–2160 (1984)
Calcitonin related peptide	387	CALCR2	PNAS 82:1994–1998 (1985)	CAL2	MCB 4:2151–2160 (1984)
Catalase	1584	CATG01	NAR 14:5321–5335 (1986)	CATL	PNAS 83:313–317 (1986)
Cholecystokinin	348	CCK2	Gene 50:353–360 (1986)	CCK	PNAS 81:726–730 (1984)
C-mos	1002	CMOS	PNAS 79:4078–4082 (1982)	CMOS	NAR 12:2147–2156 (1984)
C-myc	1320	MYCC	EMBO J. 3:383–387 (1984)	MYC	NAR 15:6419–6436 (1987)
Corticotropin-releasing factor	564	CRF	EMBO J. 2:775–779 (1983)	CRF	FEBS Lett 191:63–66 (1985)
Crystallin gamma 2-2 (D)	525	CRYGQ3	MCB 5:1408–1414 (1985)	CRYG	Gene 74:45–50 (1988)
Brain creatine kinase	1146	CKBBA	BBRC 144:1116–1127 (1987)	CKB	Gene 39:263–267 (1985)
Delta-aminolevulinate dehydratase	993	ALAD	PNAS 83:7703–7707 (1986)	ALAD	NAR 14:10115–10115 (1986)
Serum vitamin D-binding protein	1422	DBP	JCI 76:2420–2424 (1985)	DBP	JBC 261:3441–3450 (1986)
Elastase IIA	810	ELAP2A	DNA 6:163–172 (1987)	ELAII1	JBC 259:14271–14278 (1984)
Enkephalin A	792	ENK1	NAR 10:7905–7918 (1982)	ENK2	PNAS 81:7651–7655 (1984)
Enolase alpha	1305	ENOA	PNAS 83:6741–6745 (1986)	NNE	NAR 13:4365–4378 (1985)
Estrogen receptor	1788	ESTR	Nature 320:134–139 (1986)	ERR	NAR 15:2499–2513 (1987)
Liver fatty acid binding protein	384	FABPL	JBC 260:3413–3417 (1985)	FABPL	JBC 258:3356–3363 (1983)
Ferritin light chain	528	FERL	JBC 260:11755–11761 (1985)	FERL	JBC 259:4327–4334 (1984)

Appendix. Continued

Gene[a]	No. of sites[b]	Human		Rat	
		Name[c]	References[d]	Name	References
Fibrinogen alpha A	1623	FBRAA	PNAS 80:3953–3957 (1983)	FBA5E	JMB 185:1–19 (1985)
Fibrinogen gamma	1314	FBRG	BIOCH. 24:2077–2086 (1985)	FIBG1	NAR 15:2774–2776 (1987)
Gap junction protein	852	GAPJR	JBC 103:767–776 (1986)	GFPR	JBC 103:123–134 (1986)
Globin alpha adult	429	HBA4	JMB 184:7–21 (1985)	HBAM	BBRC 146:618–624 (1987)
Glucagon	537	GG	Nature 304:368–371 (1983)	GLU2	JBC 259:14082–14087 (1984)
Glucocorticoid receptor alpha	2325	GCRA	Nature 318:635–641 (1985)	GCR	Cell 46:389–399 (1986)
Glucuronidase beta	1941	GLCB	PNAS 84:685–689 (1987)	GLCB	PNAS 83:7292–7296 (1986)
Glucose transporter protein	1479	GLUTRN	Science 229:941–945 (1985)	GLUTRN	PNAS 83:5784–5788 (1986)
Glutathione S transferase	669	GST2	PNAS 84:2377–2381 (1987)	GSTYC2	PNAS 83:9393–9399 (1986)
Glyceraldehyde-3-phosphate dehydrogenase	1002	G3PD	NAR 12:9179–9189 (1984)	GAPDHB	NAR 13:1431–1442 (1985)
Growth hormone (hGH; somatotropin)	648	GH	PNAS 82:699–702 (1985)	GH1	PNAS 78:4867–4871 (1981)
Growth hormone releasing factor	297	GRFP2	PNAS 82:63–67 (1985)	GHRF2	Nature 314:464–467 (1985)
*Insulin	333	INS01	NAR 10:2225–2240 (1982)	INSI	MCB 5:2090–2103 (1985)
Insulin-like growth factor II (IGF-II)	543	GFI2	Nature 310:775–777 (1984)	INGFII	NAR 13:1119–1134 (1985)
Insulin-like growth factor IA	462	GFIAB1	JBC 261:4828–4832 (1986)	GFIL1	JBC 262:7894–7900 (1987)
Interferon gamma immune	465	IFNG	NAR 11:1819–1867 (1983)	IFNG1	EMBO J. 4:761–767 (1985)
*Interferon alpha	555	IFNAD	Gene 27:87–99 (1984)	IFNA1	NAR 12:1227–1242 (1984)
Beta protein kinase C type II	2022	PKB	Science 233:859–866 (1986)	PKCII	Cell 46:491–502 (1986)
Kininogen high molecular weight	1839	KIN01	JBC 260:8601–8609 (1985)	KINKH	JBC 262:2345–2351 (1987)
Kininogen low molecular weight	1266	KIN01	JBC 260:8601–8609 (1985)	KINKL	JBC 260:12054–12059 (1985)
Lactate dehydrogenase-A	999	LDHA	EJB 147:9–15 (1985)	LDH	NAR 13:711–726 (1985)
Lipocortin I	1041	LIPCR	Nature 320:77–81 (1986)	LCI	NAR 15:7637–7637 (1987)
Luteinizing hormone beta (LH)	426	LHB	Nature 320:77–81 (1986)	LHB	NAR 15:7637–7637 (1987)
Metallothionein-II	186	MET2	Nature 299:797–802 (1982)	MT12C	MCB 6:302–314 (1986)
*Metallothionein-I	186	METIE	JBC 260:7731–7737 (1985)	MT12C	MCB 6:302–314 (1986)
Fast myosin alkali light chain MLC1-f	576	MLC1F	NAR 15:4989–4989 (1987)	MLC131	JBC 259:13595–13604 (1984)
Fast myosin alkali light chain MLC3-f	453	MLC3F	NAR 15:4989–4989 (1987)	MLC132	JBC 259:13595–13604 (1984)
Neuropeptide Y	294	NPY	PNAS 81:4577–4581 (1984)	NPY	PNAS 84:2532–2536 (1987)
Ornithine aminotransferase	1320	OAT	PNAS 83:1203–1207 (1986)	OTA	JBC 260:12993–12997 (1985)
Ornithine transcarbamylase	1065	OTC	Science 224:1068–1074 (1984)	OTC	DNA 4:147–156 (1985)
Oxytocin-neurophysin I	375	OTNPI	JBC 260:10236–10241 (1985)	OXTNP	PNAS 81:2006–2010 (1984)
Pancreatic phospholipase A2	441	PLA2A1	DNA 5:519–527 (1986)	PPLA2	JB 99:733–739 (1986)
Pancreatic polypeptide	216	PPP	PNAS 82:1536–1539 (1985)	PPP	JBC 263:2990–2997 (1988)
Parathyroid hormone	348	PTH2	PNAS 80:2127–2131 (1983)	PTH3	JBC 259:3320–3329 (1984)
Phenylalanine hydroxylase	1356	PHH	Nature 327:333–336 (1987)	PHH	JBC 261:4148–4153 (1986)
Polymerase beta	957	POLB	BIOCH. 27:901–909 (1988)	POLB	PNAS 83:5106–5110 (1986)
Prolactin	678	PRL	JBC 256:4007–4016 (1981)	PRLSDM	JBC 255:6502–6510 (1980)
Proopiomelanocortin	708	POMC	NAR 11:6847–6858 (1983)	POMC2	FEBS Lett 193:54–58 (1985)
*Pulmonary surfactant-associated protein A	747	PSPA	JBC 261:9029–9033 (1986)	PSPA	BBRC 144:367–374 (1987)
C-Raf	1947	RAFR	NAR 14:1009–1015 (1986)	RAFA	MCB 7:1226–1232 (1987)
C-Ha-ras1	570	RASH	PNAS 81:5384–5388 (1984)	RASH1C	MCB 6:1706–1710 (1986)
Retinol binding protein	600	RBP	NAR 11:7769–7776 (1983)	RBP1	JBC 260:11476–11480 (1985)
Cellular retinol binding protein	408	RBPC	BBRC 130:431–439 (1985)	RBPC	PNAS 84:3209–3213 (1987)
Renin	1203	REN01	PNAS 81:5999–6003 (1984)	REN	PNAS 84:5605–5609 (1987)
Relaxin	555	RELAX1	Nature 301:628–631 (1983)	RELAX	Nature 291:127–131 (1981)
Ribosomal protein S17	408	RPS17	PNAS 83:6907–6911 (1986)	RPS17	Gene 35:289–296 (1985)
Somatostatin	351	SOMI	Science 224:168–171 (1984)	SOM141	JBC 260:8145–8156 (1985)
Superoxide dismutase copper–zinc	465	SODG1	EMO J. 4:77–84 (1985)	SODR	NAR 14:6746–6746 (1986)
Synaptophysin (p38)	888	SYNPR	NAR 15:9607–9607 (1987)	SYNPR	NAR 15:9607–9607 (1987)

Appendix. Continued

Gene[a]	No. of sites[b]	Human		Rat	
		Name[c]	References[d]	Name	References
Thy-1 glycoprotein	486	THY1A	PNAS 82:6657–6661 (1985)	THY1G	Nature 313, 485–487 (1985)
Thyrotropin beta	417	TSH1	FEBS 188:394–400 (1985)	TSHBM	BBRC 128:1152–1158 (1985)
Transforming growth factor alpha	480	TGFAM	Cell 38:287–297 (1984)	TGFA	Nature 313:489–491 (1985)
Translational initiation factor 2	948	EIF2A	JBC 262:1206–1212 (1987)	EIF2A	JBC 262:1206–1212 (1987)
Transthyretin	444	PALA	BBCR 125:636–642 (1984)	PALTA	JBC 260:6481–6487 (1985)
Tropomyosin fibroblast muscle type	855	TROPA	PNAS 82:7835–7839 (1985)	TRO01	MCB 6:3582–3595 (1986)
*Tubulin alpha	1356	TUBAG	NAR 13:207–223 (1985)	TUBAL1	Nature 300:330–335 (1982)
*Tubulin beta	1338	TUBBM	JMB 182:11–20 (1985)	TUBB15	EMBO J. 4:3667–3673 (1985)
Tyrosin hydroxylase	1494	HTH1R	Nature 326:707–711 (1987)	TOHA	PNAS 82:617–621 (1985)
Arginine vasopressin neurophysin II	495	VPNP	JBC 260:10236–10241 (1985)	VPNP	JBC 258:14061–14064 (1983)

[a] Homologous genes with an asterisk are present in many copies in the genome (in only one species or in the two species). So we cannot be sure to have compared orthologous genes

[b] Genes were aligned and gaps were eliminated. The sizes used for the study are given in the table

[c] Genes are extracted from Genbank (release 57). Their name in the sequence data bank is mentioned. Names underlined are gene clusters

[d] The references are given by information in Genbank (interrogation by ACNUC). If many references are associated with a gene, the most recent is mentioned in this table. The journal abbreviations are: JBC = Journal of Biological Chemistry, MCB = Molecular and Cellular Biochemistry, BIOCH. = Biochemistry, BBRC = Biochemical and Biophysical Research Communications, JMB = Journal of Molecular Biology, BJ = Biochemical Journal, JB = Journal of Biochemistry, J.IMM = Journal of Immunology, JCI = Journal of Clinical Investigation, NAR = Nucleic Acids Research, PNAS = Proceedings of the National Academy of Science USA, EJB = European Journal of Biochemistry