Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 346 (2005) 215–219

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Eighty percent of proteins are different between humans and chimpanzees

Galina Glazko, Vamsi Veeramachaneni, Masatoshi Nei, Wojciech Makałowski*

Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802, USA
Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

Received 12 August 2004; received in revised form 1 October 2004; accepted 5 November 2004
Available online 29 January 2005
Received by T. Gojobori

Abstract

The chimpanzee is our closest living relative. The morphological differences between the two species are so large that there is no problem in distinguishing between them. However, the nucleotide difference between the two species is surprisingly small. The early genome comparison by DNA hybridization techniques suggested a nucleotide difference of 1–2%. Recently, direct nucleotide sequencing confirmed this estimate. These findings generated the common belief that the human is extremely close to the chimpanzee at the genetic level. However, if one looks at proteins, which are mainly responsible for phenotypic differences, the picture is quite different, and about 80% of proteins are different between the two species. Still, the number of proteins responsible for the phenotypic differences may be smaller since not all genes are directly responsible for phenotypic characters.

Keywords: Human; Chimpanzee; Genetic distance; Protein identity; Nucleotide identity

1. Introduction

In terms of nucleotide differences, the human is closer to the chimpanzee than to any other hominoid species. The early genome comparison by DNA hybridization suggested a nucleotide difference of 1–2% (Kohne, 1970; Sibley and Ahlquist, 1984). Recently, direct nucleotide sequencing confirmed this estimate (Goodman, 1995; Chen and Li, 2001; Ebersberger et al., 2002; Watanabe et al., 2004).

However, a large portion (about 98%) of the human genome is known to be non-protein-coding DNA, and the estimate of 1–2% nucleotide difference is largely based on the comparison of non-protein-coding DNA, which has little effect on phenotypic characters. Therefore, for the general public who are interested in phenotypic differences, this is clearly misleading. A better way of measuring the genetic difference is to consider functional genes or proteins as the units of comparison, because these are the genetic units that control phenotypic characters. To do this, we compiled 127 human and chimp orthologous proteins (44,000 amino acid residues) from GenBank. Only 25 (20%) of these proteins showed the identical amino acid sequence between humans and chimpanzees. In other words, the proportion of different proteins was 80%, in contrast to the 1–2% difference at the nucleotide level. How these differences are related to the morphological differences is unclear at present, but it is quite possible that a large proportion of phenotypic differences are caused by a relatively small number of

Abbreviations: PAM, percent of accepted mutations; MHC, major histocompability complex; d_s, synonymous substitution distance; d_n, nonsynonymous substitution distance; PBL, Pamilo–Bianki–Li method; mNG, modified Nei-Gojobori method; NCBI, National Center for Biotechnology Information; GO, Gene Ontology.

* Corresponding author. Department of Biology, Pennsylvania State University, 514 Muelle Lab, University Park, PA 16802, USA.
E-mail address: wojtek@psu.edu (W. Makałowski).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.11.003

regulatory mutations (King and Wilson, 1975) or major effect genes (Nei, 1987).

2. Materials and methods

2.1. Apes protein sequences

All human protein sequences known as of March 2003 were compared with all ape proteins available at that time. The human and ape protein data set was prepared as follows. First, all the ape proteins were downloaded from GenBank, and then all the identical proteins and those that are substrings of other proteins were merged. Finally, since we were interested only in full-length proteins, we checked all the sequences for the following criteria: length of at least 10 amino acids with the initiation codon of methionine. It resulted in the following data set—human: 71,334; chimpanzee: 384; gorilla: 157; orangutan: 152; and gibbon: 82 proteins.

2.2. Assignment of orthology relationships

To find ape orthologs of human proteins, the BLAST search of each human protein sequences against all ape sequences was performed with the following parameters: E-value cut-off (e), 0.001; matrix (M), PAM30; gap extension penalty (E), 1; gap opening penalty (G), 9; number of database sequences to show one-line description for (v), 10; number of database sequences to show alignment for (b), 10; and low complexity filtering for lookup table only (F), “m L”. The resulting alignments were manually analyzed. Multigene families such as major histocompability complex (MHC), immunoglobulin, olfactory receptor, and KIR receptor gene families were excluded from the analysis because of difficulties in detecting orthologous relationships. Mitochondrial proteins were also excluded from the analysis. The final ortholog data set consisted of 137 human, 127 chimpanzee, 60 gorilla, 56 orangutan, and 31 gibbon sequences. Here, some human genes were orthologous only to some ape genes.

2.3. Statistical analysis

Orthologous sequence pairs were transformed into 137 multiple species orthology groups using the single linkage approach, and multiple sequence alignments were obtained by using ClustalW with default parameters. Using these alignments, we computed the number of amino acid and nucleotide differences per site. We also computed the synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site using the modified Nei–Gojobori (mNG) and the Pamilo–Bianki–Li (PBL) methods (see Nei and Kumar, 2000). The statistical test (Z test) of the difference $d_N - d_S$ was conducted by computing the standard error of $d_N - d_S$ analytically or with the boostrap test. All these computations were done by using the computer program MEGA2 (Kumar et al., 2001). We did not use the Goldman–Yang method, because the statistical test of $d_N - d_S$ was not available in the program PAML (Yang, 2003).

2.4. Database of orthologous genes

All the results are stored in a *mysql* database. With a simple web interface, readers can have access to the following information: lists of identical protein sequences between hominoid species pairs, identical coding DNA sequences between them, and the proteins that do not appear in ortholog sets. We also provide a link to a table with orthologous gene groups from which multiple amino acid sequence alignment is available, along with links to original sequence information via Entrez system at the National Center for Biotechnology Information (NCBI). Finally, there is a simple search tool that enables a keyword search based on information included in definition line of the original records. The database is accessible at http://warta.bio.psu.edu/ApesDB/.

3. Results and discussion

3.1. Statistical properties of the data

Using the orthologs specified in Section 2.2, we selected 411 hominoid orthologous genes clustered in 137 orthology groups. The number of orthologous genes for each species pair is summarized in Table 1. Five-species orthology groups were established only for 18 proteins. The data set used here appears to be a random set of proteins, because the size distribution of the proteins was similar to that of the entire set of human proteins (see Fig. 1). The list of all proteins used in this study is provided on the project web site http://warta.bio.psu.edu/ApesDB/. Only 25 out of the 127 chimpanzee proteins (20%) were identical to their human orthologs (see Table 2 for the list of these proteins). This is approximately in accordance with a random distribution of the 0.6% nonsynonymous substitutions across proteins of average length of 330 amino acids (V.V. and W.M., unpublished simulation results). Interestingly, there are several genes that showed the identical nucleotide

Table 1
Number of orthologous gene pairs used in this study

	Human	Chimpanzee	Gorilla	Orangutan
Chimpanzee	127			
Gorilla	60	56		
Orangutan	56	52	44	
Gibbon	31	26	22	24

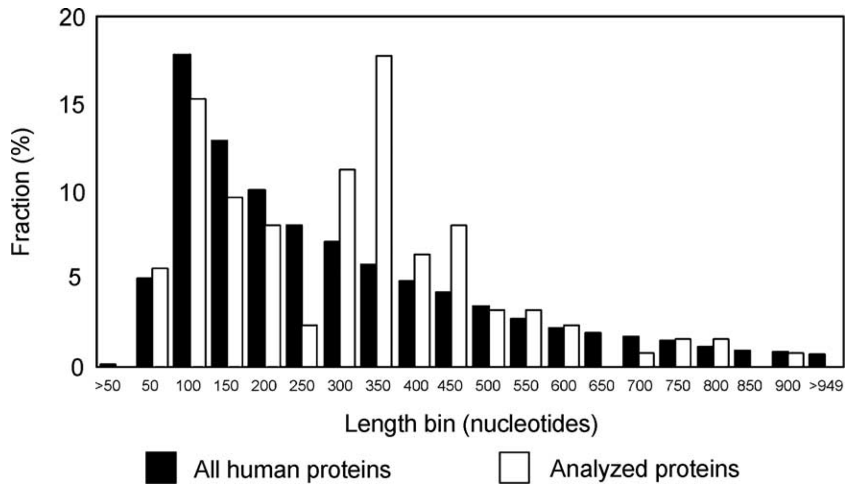

Fig. 1. Length distribution of all human proteins (black histograms) against proteins in the analyzed data set (white histograms).

sequence between different species (see Table 3). For example, both amino acid and nucleotide (coding part) sequences of beta-2 microglobulin were identical among human, chimp, and gorilla genomes, while the interleukin-2 precursor gene sequences from human and gibbon were identical. (Data for other three species were not available.) The overall nucleotide sequence identity between human and other hominoid species was in accordance with earlier studies (e.g., Chen and Li, 2001; Ebersberger et al., 2002; Watanabe et al., 2004, Kitano et al., 2004).

Table 2
List of identical proteins in humans and chimpanzees

Gene name	Protein length (amino acids)
Beta-2-microglobulin	119
Opsin 1	348
Poly(A) binding protein	382
5-Hydroxytryptamine (serotonin) receptor 1B	390
5-Hydroxytryptamine (serotonin) receptor 1E	365
A-gamma globin	147
Alpha 2 globin	142
Beta 1,3-galactosyltransferase polypeptide 1	326
Beta defensin 1	68
Epsilon globin	147
G-gamma globin	147
Histamine receptor H2	359
Lysozyme precursor	148
Superoxide dismutase 1	154
Hemochromatosis protein isoform 1 precursor	348
Dopamine receptor D2	443
G protein-coupled receptor 15	360
Renin precursor	406
Mitogen-activated protein kinase 14,	360
Chemokine (C-X-C motif) receptor 4 (fusin)	352
Epididymal secretory protein	151
CD81 antigen	236
Triosephosphate isomerase 1	249
DEAD-box protein	428
Ubiquitin B	229

Beta-2-microglobulin has the identical nucleotide sequence as well.

3.2. *Evolutionary analysis of human and chimpanzee orthologs*

Our comparison of human and chimp proteins showed that the 80% of proteins are different between humans and chimps. This comparison is biologically meaningful because even one amino acid substitution may have significant effects on phenotypic differences. We were then interested in what kind of proteins maintain the interspecific identity. In general, one would expect that the same functional category of proteins shows a similar level of protein identity. We therefore classified the human/chimpanzee orthologous proteins using Gene Ontology (GO) classification (Harris, 2004). Our data set could be roughly divided into four functional GO categories: enzymes, signal transduction proteins, transporters, and others. The percentages of proteins showing 100%, 99%, 98%, and <98% sequence identity are given in Table 4 for different categories. It is clear that signal transduction is highly conserved, whereas transporter proteins are least conserved. However, these results could be biased because the number of proteins used is small.

Table 3
Number of genes with identical amino acid (above diagonal) and nucleotide (below diagonal) sequences

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	x	25	11	7	2
Chimpanzee	1	x	8	6	2
Gorilla	2	3	x	5	2
Orangutan	1	1	1	x	1
Gibbon	1	1	0	0	x

Table 4
Percentage of proteins showing 100%, 99%, and 98% sequence identity between humans and chimps for different functional categories

Functional categories	Level of amino acid identity			
	100%	99%	98%	Less than 98%
Enzymes	22.7	30.6	26.1	20.0
Signal transduction	45.5	52.8	56.5	45.0
Transporters	22.7	8.3	4.3	0.0
Others	9.1	8.3	13.0	35.0

We also computed the rates of synonymous (d_N) and nonsynonymous (d_S) nucleotide substitutions between humans and chimpanzees. A majority of proteins showed the relationship $d_N<d_S$, but there were 16 proteins for which d_N was higher than d_S (Table 5). However, only three of them showed that d_N is significantly greater than d_S by either the mNG or the PBL method. The results of the Z test using analytical or bootstrap standard errors were virtually identical. The d_N-d_S difference for protamine 1 was significant by both methods, but the function of this protein P1 is to bind and condense sperm DNA during the course of sperm nucleus condensation in spermatogenesis. For this reason, this protein is highly basic and contains a high proportion of argine residue (about 50%). When arginines mutate to nonbasic amino acids at some sites, other nonbasic amino acids tend to mutate to argines to maintain a high level of basicity (Rooney et al., 2000). Furthermore, we found $d_N<d_S$ in most other vertebrates species groups (Rooney et al., 2000). Therefore, the relationship $d_N>d_S$ observed here does not seem to be related to any phenotypic evolution.

The mNG method showed $d_N>d_S$ for glycophorin A. A similar result was obtained by Wang et al. (2003). These authors suggested that the higher d_N/d_S ratio than 1 for this gene is caused by the interaction of this receptor protein with the malaria parasites. A significantly higher d_N/d_S ratio than 1 was also observed for the BRCA1 gene by the PBL method. A similar result was reported by Huttley et al. (2000). However, the analysis of Hurst and Pal (2001) for the entire region of this gene showed the relationship $d_N<d_S$. Some mutations at this locus are known to cause breast cancer, but the normal function does not appear to be well understood (Narod and Foulkes, 2004). Therefore, it is difficult to relate the difference of this gene to the phenotypic differences between humans and chimpanzees.

For the above reason, it is not easy to understand the genetic basis of phenotypic differences between the two species at this stage. It seems that the phenotypic differences are controlled by a small proportion of genes, either by the regulatory genes (King and Wilson, 1975) or by major effect genes (Nei, 1987).

4. Conclusions

(1) Although nucleotide sequence identity between humans and chimpanzees is very high, only 20% of proteins are identical between the two species, and 80% of proteins are different.
(2) Even the 80% protein differences appear to be too small to explain the phenotypic differences. It seems that the phenotypic differences are controlled by a small proportion of genes, either by regulatory genes or by major effect genes.
(3) A larger number of genes than ours need to be studied to understand the genetic basis of phenotypic differences.

Table 5
Genes showing $d_N>d_S$ in the comparison of human and chimpanzee genes

Gene name	Modified Nei–Gojobori				Pamilo-Bianchi-Li			
	d_N	d_S	Z-boot	Z-anal	d_N	d_S	Z-boot	Z-anal
Protamine 1	0.088	0.000	2.642*	2.952*	0.104	0.000	2.345*	2.821*
Glycophorin A	0.047	0.012	2.522*	2.262*	0.045	0.020	1.407	1.203
Protamine 2	0.053	0.023	1.370	1.306	0.055	0.037	0.609	0.506
SRY	0.016	0.006	1.588	1.179	0.016	0.005	1.758	1.468
EP2_variantC	0.009	0.000	1.471	1.412	0.008	0.000	1.365	1.397
Rhesus-like	0.037	0.029	0.791	0.760	0.036	0.025	1.155	1.079
EP2_variantE	0.006	0.000	1.008	0.999	0.007	0.000	1.073	0.997
TGIF-like	0.010	0.005	0.784	0.744	0.010	0.004	1.077	1.036
EP2_variantD	0.004	0.000	0.964	0.999	0.004	0.000	0.988	0.998
Rh50	0.009	0.006	0.687	0.676	0.009	0.005	0.967	0.885
BRCA1	0.008	0.005	1.593	1.588	0.008	0.004	2.427*	2.353*
Apolipoprotein_AII	0.015	0.011	0.239	0.237	0.015	0.009	0.572	0.423
Glycoprotein alpha-2	0.009	0.006	0.605	0.571	0.010	0.005	0.987	0.863
Interleukin-4	0.003	0.000	1.068	0.999	0.004	0.000	0.931	0.998
STRL33	0.003	0.000	1.455	1.414	0.003	0.000	1.484	1.412
Interleukin-8 receptor	0.006	0.003	0.584	0.582	0.006	0.003	0.820	0.798

Z-boot: the standard error for this test was obtained by the bootstrapping. Z-anal: the standard error for this test was obtained by the analytical formula.
* Significant at the 5% level.

Acknowledgement

We thank Jongmin Nam for checking some of our computations. This study was supported by NIH grant GM020293-33 to MN.

References

Chen, F.C., Li, W.-H., 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68, 444–456.

Ebersberger, I., Metzler, D., Schwarz, C., Paabo, S., 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. Am. J. Hum. Genet. 70, 1490–1497.

Goodman, M., 1995. Epilogue: a personal account of the origins of a new paradigm. Mol. Phylogenet. Evol. 5, 269–285.

Harris, M.A., 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258–D261.

Hurst, L.D., Pal, C., 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. 17, 62–65.

Huttley, G.A., et al., 2000. Adaptive evolution of the tumour suppressor *BRCA1* in humans and chimpanzees. Nat. Genet. 25, 410–413.

King, M.C., Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. Science 188, 107–116.

Kitano, T., Liu, Y.H., Ueda, S., Saitou, N., 2004. Human-specific amino acid changes found in 103 protein-coding genes. Mol. Biol. Evol. 21, 936–944.

Kohne, D.E., 1970. Evolution of higher-organism DNA. Q. Rev. Biophys. 3, 327–375.

Kumar, S., Tamura, K., Jakobsen, I.B., Nei, M., 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17, 1244–1245.

Narod, S.A., Foulkes, W.D., 2004. BRCA1 and BRCA2: 1994 and beyond. Nat. Rev., Cancer 4, 665–676.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Oxford.

Rooney, A.P., Zhang, J., Nei, M., 2000. An unusual form of purifying selection in a sperm protein. Mol. Biol. Evol. 17, 278–283.

Sibley, C.G., Ahlquist, J.E., 1984. The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. J. Mol. Evol. 20, 1–15.

Wang, H.-Y, Tang, H., Shen, C.-K., Wu, C.-I., 2003. Rapidly evolving genes in human: I. The glycophorins and their possible role in evading malaria parasites. Mol. Biol. Evol. 20, 1795–1804.

Watanabe, H., et al., 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature 429, 382–388.

Yang, Z., 2003. Phylogenetic Analysis by Maximum Likelihood (PAML) V. 3.14. University College, London.