Large-Scale Search for Genes on Which Positive Selection May Operate

Toshinori Endo, Kazuho Ikeo, and Takashi Gojobori
National Institute of Genetics, and Department of Genetics, School of Life Science, Graduate University for Advanced Studies, Mishima, Japan

We conducted a systematic search for the candidate genes on which positive selection may operate, on the premise that for such genes the number of nonsynonymous substitution is expected to be larger than that of synonymous substitutions when the nucleotide sequences of the genes under investigation are compared with each other. By obtaining 3,595 groups of homologous sequences from the DDBJ, EMBL, and GenBank DNA sequence databases, we found that 17 gene groups can be the candidates for the genes on which positive selection may operate. Thus, such genes are found to occupy only about 0.5% of the vast number of gene groups so far available. Interestingly enough, 9 out of the 17 gene groups were the surface antigens of parasites or viruses.

Introduction

The neutral theory of molecular evolution maintains that fixation of neutral mutations by random genetic drift within a population causes evolution of most genes (Kimura 1983). Adaptive mutations can contribute to improvement of genes more directly than the neutral mutations, though they may occur at an extremely low frequency. Most deleterious mutations, on the other hand, do not contribute to gene evolution because they have a high chance of being lost in a small number of generations even though they may occur at a high frequency.

The genes on which positive selection operates are considered to have an evolutionary character that the number \( d_{NS} \) of nonsynonymous substitutions is larger than that \( d_{S} \) of synonymous substitutions since positive selection prefers nonsynonymous substitutions that cause adaptive amino acid changes rather than synonymous substitutions that do not cause any amino acid changes. By this criterion, it was shown that the highly polymorphic genes for major histocompatibility complex (MHC) classes I and II may be subjected to positive selection by the fact that \( d_{NS} \) is significantly larger than \( d_{S} \) at the antigen recognition sites of the genes (Hughes and Nei 1988, 1989). Similarly, several other genes can be considered to be candidate genes on which positive selection may operate, such as stomach lysozyme genes of ruminants (Irwin and Wilson 1990), visual pigment genes (Yokoyama and Yokoyama 1989), homeobox genes (Kappen, Schughart, and Ruddle 1989), \( \gamma \)-hemoglobin (Fitch et al. 1991), alcohol dehydrogenase genes (Long and Langley 1993), ion channel genes (Strong, Chandy, and Gutman 1993), growth hormone genes (Ohta 1993), \( \alpha \)-antitrypsin genes (Ohta 1994), and abalone sperm lysozin genes (Lee, Ota, and Vacquier 1995). Nevertheless, these studies were conducted for the individual cases, and therefore the proportion of such genes on which positive selection operates was unknown. Thus, it is very important to understand what proportion of the genes on which positive selection operates is among all the sequence data available and what they are, in order to know not only the evolutionary mechanisms of the genes but also a relationship between the evolutionary process and functions of the genes.

With the aim of answering these interesting questions, we conducted systematic search for the candidate genes on which positive selection may operate by applying our newly developed method to be mentioned later to the DNA sequence database and by creating the multiple alignments of 3,595 homologous cDNA sequences.

Materials and Methods

Construction of the cDNA Alignment Database for Protein-Coding Sequences

A total of 24,832 sequences of the protein-coding sequences were extracted from the DNA sequence database, DDBJ release 11, which includes three DNA database of DDBJ, EMBL, and GenBank. All the coding sequences were translated into amino acid sequences. The amino acid sequences were then categorized into appropriate gene groups of homologous sequences by using the computer program, blast (Altschul et al. 1990). In this case, we set 0.5 as the maximum threshold of the expected probability score. The homologies of the sequence groups were then examined by the algorithm of global homology search that was originally developed by Needleman and Wunsch (1970). We used only the sequences that show the similarity score larger than a quarter of the score of identical sequences. These homologous sequences were then multiple-aligned to create the database called “SODHO” (Gojobori et al. 1993). Then, the multiple alignments of homologous amino acid sequences were converted into the corresponding nucleotide sequences. We now call them the cDNA multiple alignments.

Search for Genes on Which Positive Selection Operates

The candidate genes on which positive selection may operate were searched by comparing the estimated number \( d_{NS} \) of nonsynonymous substitutions with that \( d_{S} \) of synonymous substitutions. \( d_{NS} \) and \( d_{S} \) were estimated by Nei and Gojobori’s method (Nei and Gojobori 1986) for all the sequence pairs in the cDNA alignments for each gene group. If \( d_{NS} \)'s were larger than the cor-
FIG. 1.—Strategies for searching for the candidate genes on which positive selection operates.
responding $d_s$'s for more than half of the sequence pairs compared in a particular gene group, the gene group was regarded as the candidate gene on which the positive selection may operate. We excluded from the analysis the gene groups that either $d_s$ or $d_N$ was larger than 1 to avoid the saturation effect of nucleotide substitutions on the estimation process. We also inspected the resulting alignments manually to prevent incorporation of possible inappropriate alignments such as frame shifts occurred in only a tiny part of the DNA sequences. The entire procedures are summarized in figure 1.

Window Analysis of $d_s$ and $d_N$

In order to identify particular regions where positive selection may be operating within the candidate genes, we conducted the window analysis.

A window is defined as a sequence region of 20-codon length on an alignment of homologous genes. In the window analysis, $d_s$ and $d_N$ were estimated for each window along the nucleotide sites codon by codon to test whether $d_N$ was larger than $d_s$ for the window region within the gene. If $d_N$ was larger than $d_s$ in the within-gene region, it will be considered to be the within-gene region on which positive selection may operate.

Results and Discussion

Ratio of Candidate Genes on Which Positive Selection May Operate

A total of 24,832 complete coding sequences were extracted from the DNA sequence database. All of those sequences were classified into 3,595 homologous gene groups except 4,845 unclassifiable sequences. The unclassifiable sequences are defined as the sequences that were not included in any of the gene groups of the cDNA multiple alignments. Surveying the 3,595 homologous gene groups, we searched for gene groups that meet the criterion that $d_s$ is larger than $d_N$ for more than half the pairs in each gene group. Finally, we obtained 17 gene groups as candidate genes on which positive selection may operate (table 1). The candidate genes constituted 0.45% (=17/3,595 × 100%) of all the gene groups examined so far.

Members of the Candidate Genes

Interestingly enough, 9 of the 17 candidate genes were the antigenic surface proteins of parasites and viruses (table 1). They were an outer membrane protein gene of Chlamydia, an envelope protein gene of equine infections anemia virus, a merozoite surface antigen (MSA2) gene of malaria Plasmodium falciparum, an E gene of phages G4, FX174, and S13, a glycopolypeptide gene $gH$ of Pseudorabies virus, a sigma-I protein gene of reovirus, a major surface protein ($msp1\alpha$) of rickettsia Anaplasma marginale, and an invasion plasmid antigen of Shigella. This result might indicate that favoring amino acid changes are taking place on these genes in order to escape from the defensive system of the hosts, such as the immune system of vertebrates. For the remaining gene groups, however, we could not find any common characteristic features.

Identification of the Within-Gene Region Where Positive Selection Operates

With the aim of locating the regions where positive selection operates within the candidate genes, we conducted the window analysis for all 17 candidate genes. Let us describe the results of only two candidate genes as examples.

Case of Merozoite Surface Antigen 2 Genes of Malaria

The first example is the merozoite surface antigen 2 (MSA2) genes of malaria Plasmodium falciparum (fig. 2). Malaria is a well-known parasite that causes malarial fever. MSA2 is known to be one of the genes expressed during asexual life cycle (Smythe et al. 1988). Merozoite is the only stage of a malarial life cycle in which malaria is exposed to the human bloodstream except the time of infection. Thus, it is quite possible that the genes expressed in the merozoite stage are a target of the im-

<table>
<thead>
<tr>
<th>Gene Group</th>
<th>Representative Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merozoite surface antigen (MSA2)</td>
<td>Malaria Plasmodium falciparum</td>
</tr>
<tr>
<td>Major surface protein (msp/α)</td>
<td>Rickettsia Anaplasma marginale</td>
</tr>
<tr>
<td>Outer membrane protein (omp)</td>
<td>Chlamydia</td>
</tr>
<tr>
<td>env</td>
<td>Equine infectious anemia virus</td>
</tr>
<tr>
<td>Citycprotein gH gene</td>
<td>Pseudorabies virus</td>
</tr>
<tr>
<td>E gene</td>
<td>Phages G4, dX174 and S13</td>
</tr>
<tr>
<td>Sigma-I protein gene</td>
<td>Reovirus</td>
</tr>
<tr>
<td>Invasion plasmid antigen gene (ipuC)</td>
<td>Shigella</td>
</tr>
<tr>
<td>Invasion plasmid antigen gene (ipuD)</td>
<td>Shigella</td>
</tr>
<tr>
<td>Egg-laying hormone</td>
<td>Aplysia californica</td>
</tr>
<tr>
<td>Egg-laying hormone A peptide</td>
<td>Aplysia californica</td>
</tr>
<tr>
<td>ATP synthase F, subunit (atp-2)</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>Neomycin resistance protein gene</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>Virulence determinant gene (yadA)</td>
<td>Yersinia</td>
</tr>
<tr>
<td>Prostatic steroid binding protein</td>
<td>Rat</td>
</tr>
<tr>
<td>Neurotoxin</td>
<td>Snake</td>
</tr>
<tr>
<td>CDC6</td>
<td>Saccharomyces cerevisiae</td>
</tr>
</tbody>
</table>

Table 1
The Gene Groups on Which Positive Selection May Operate
immune attack from a human body. In fact, a high degree of polymorphism of MSA2, just like that of the MHC genes, has been reported in various studies (Smythe et al. 1988, 1990, 1991; Elliott et al. 1990; Fenton et al. 1991; Marshall et al. 1991).

Figure 3 shows the result of the window analysis for two alleles of MSA2. It is clear from the figure that there are two regions where the number of nonsynonymous substitutions are high whereas no synonymous substitution is observed. Although the function of the second region is not well understood, the first region corresponds exactly to known antigenic epitopes (fig. 2; Fenton et al. 1991). Thus, it suggests that positive selection is operating on the antigenic epitopes of MSA2 to escape from the host immune system. It also implies that the second region found in this analysis can be another possible antigenic epitope.

It is very interesting that no synonymous substitutions were found over the entire region of MSA2. In the case of MSA1, which are another malaria surface protein genes, a small number of synonymous substitutions were observed (Hughes 1991, 1992). We have to note that the genome of P. falciparum is extremely A+T rich, the G+C content being only 18% at the third positions of the codons (Musto, Rodriguez-Maseda, and Bernardi 1995). It implies that nucleotide substitutions occur mostly between A and T by some kinds of mutation pressure. It is clear from the genetic codes that the nucleotide substitutions between A and T causes mostly amino acid changes even at the third codon positions. Even if nucleotide substitutions should occur from A or T to G or C, it is possible that the mutated nucleotides can return easily to A or T by back mutations. For example, suppose that a synonymous substitution from A to G took place at codon AAA for lysine, changing this codon to AAG for the same lysine. However, it is possible that a back mutation from AAG to AAA easily takes place by strong mutation pressure toward A+T. Thus, we may not be able to observe any synonymous substitutions at all, even though they are really occurring (Musto, Rodriguez-Maseda, and Bernardi 1995).

Case of Major Surface Protein of Rickettsia

Another example is the major surface protein 1α (msp1α) genes of rickettsia Anaplasma marginale. Anaplasma marginale can cause anaplasmosis, a hemoparasitic disease of cattle. The major surface protein 1 of A. marginale contains the neutralization-sensitive epitope (Oberle et al. 1988). The neutralization-sensitive epitope was mapped to the repeated region where variation in length between the clones was observed. In fact, the msp1α gene codes for 78–105-kDa proteins that depend on the number of repeats (Allred et al. 1990). There were two major regions where $d_a$ is larger than $d_s$ (fig. 4). The first region is from site number 150–900 and the second region is from 1060–1600.

The first region corresponds exactly to the known repeat region where the predicted antigenic epitope is located (Allred et al. 1990). Thus, it again suggests that positive selection is operating on the first region within the msp1α gene by favoring amino acid changes to escape from the host immune system. Although the biological significance of the second region is unknown, it is possible that the second region is also another antigenic epitope that has not been discovered yet.

Only a Small Proportion of the Genes on Which Positive Selection Might Operate

In this study, 17 out of 3,595 gene groups were found to be the candidate genes on which positive selection may operate. It means that for only 0.47% of the gene groups examined, positive selection may be operating. This proportion is very small as expected by the neutral theory of molecular evolution (Kimura 1983). However, we have to be very careful about the fact that the candidate genes were identified by examining the
FIG. 5.—The window analysis of the A locus of human class I MHC genes. The numbers of synonymous and nonsynonymous substitutions are plotted by dashed and solid lines, respectively, along the nucleotide sites. The antigen recognition sites are marked by shadowing.

For this reason, all of these genes were not always identified as the candidate genes in the present study.

Common Features Among the Candidate Genes

Nine of 17 candidate genes found by our systematic search were the surface proteins of parasites and viruses. These results suggest a possibility that positive selection may be operating on the genes for parasite and virus surface proteins, in order to escape from the immune system of their hosts by favoring frequent nonsynonymous substitutions. To examine this possibility, we conducted a window analysis in the two candidate genes, MSA2 and msp1α. For these two gene groups, the window analysis successfully identified the two regions where \( d_N \) is larger than \( d_S \) within each of the entire genes. These regions corresponded very well to the antigenic epitopes of these surface proteins. Thus, it is possible that the surface proteins of parasites and viruses can be identified as the candidate genes by showing that \( d_N \) is larger than \( d_S \) in their epitope regions.

As for the candidate genes other than the surface proteins, we have no idea about the common features among these genes. More detailed analysis such as the window analysis for these genes will be required for understanding the biological significance of positive selection on those genes.

Acknowledgments

We thank Drs. Yoshio Tateno, Naruya Saitou, and Hector Musto for their helpful discussion. We also thank anonymous reviewers for their helpful comments.

LITERATURE CITED


Mitiko Go, reviewing editor

Accepted February 7, 1996