Visual Display of Sequence Conservation as an Aid to Taxonomic Classification Using PCR Amplification

Peter K. Rogan, Joseph J. Salvo, R. Michael Stephens
and
Thomas D. Schneider

By comparing corresponding gene sequences from three or more organisms, we can deduce relationships which can be used for biological classification. Portions of a gene with the greatest number of differences (polymorphisms) generally produce the largest amount of useful data for classification of new specimens. In order to identify highly polymorphic regions, we displayed the 28S ribosomal RNA gene as a *sequence logo.*[15] In this representation, the height of letters measures the degree of sequence conservation among several species. We then picked a region which has two conserved motifs surrounding a highly variable domain. The conserved portions are the same in many organisms, so we were able to use the polymerase chain reaction (PCR) technique to amplify the variable portion in the middle. When sequenced and compared, these amplified pieces of DNA will allow taxonomic classifications to be made.

Introduction

Taxonomic classification of organisms requires the study of both the similarities and differences between the organisms. Genetic homology and mutations are analogous to the morphological criteria that have classically been used to deduce evolutionary relationships. Because they are so easy to obtain, nucleic-acid sequences of equivalent genes in different species are now used to determine taxonomic relationships between the species.[20]

If two different species have identical genes, then we have little to say about how they evolved from a common ancestor. When two genes differ at one or more sites, the number of nucleotide changes indicates how much the two organisms have diverged from a common ancestor. This paper describes a method for identifying and then amplifying regions of sequence divergence.

First, regions of DNA that have similar features are located and used to align the sequences. DNA sequence regions that are similar in more than one species are said to be "conserved". This conservation can be visualized with a sequence logo,[15] which displays the nucleotides that are present in multiple aligned sequences. These logos can be used to locate a region of sequence divergence surrounded by two regions of conservation. DNA primers are designed to anneal to the conserved regions for polymerase chain reaction (PCR) amplification.[12] These primers amplify both conserved regions and the divergent DNA sequences between them. The variable region between the two primer sequences can be used to construct dendrograms which depict the taxonomic relationships between different organisms. This method simultaneously satisfies the requirement of PCR for two conserved DNA sequences from which to perform DNA amplification, and the requirement of taxonomy for a highly variable DNA region from which to construct evolutionary trees.

The sequences of those RNA molecules, which are integral components of the ribosome, can be used to identify functional genetic differences between species. All organisms employ ribosomes to carry out the important biological task of protein synthesis, so ribosomal subunits have been structurally and functionally conserved throughout the eons. The sequences of ribosomal RNAs from widely differing species can be aligned and then the differences between these sequences specify the evolutionary or phylogenetic relationships between the organisms.[20]

Methods and Results

Eukaryotic ribosomes contain several conserved RNAs, the largest of which is called the 28S ribosomal RNA (rRNA). In order not to bias the analysis to regions studied previously,[7,20] we chose to use full length 28S sequences in the sequence alignment. Sequences having the broadest possible taxonomic distribution were selected in order to maximize species diversity (sequences of plant, animal, fungal, and protistan origin were used). The corresponding 28S ribosomal DNA (rDNA) sequences were obtained from the GenBank genetic sequence repository.[1] Sequences were aligned by a "rectification algorithm"[3,4] in which the human sequence was chosen to be a reference sequence. The human sequence was aligned in a pairwise fashion with each of the other sequences, and then the modified reference sequence (containing gaps) was realigned with the other sequences again to produce the final alignment. Although other alignment strategies might generate other solutions, they were not explored because our studies produced satisfactory results (see below).

A sequence logo was created from the aligned 28S rDNA sequences, and the region shown in Figs. 1 and 2 was one identified as having two conserved regions surrounding a divergent region. The horizontal axis represents nucleotide positions along the DNA whereas the vertical axis measures the degree of conservation at the same position in the various species. The vertical scale is given in bits of information, which measures the number of choices between two equally likely possibilities. The choice of one base from the four possible bases requires two bits of information.

The two bits correspond to two choices. For example, the first choice could determine whether the base is a purine or a pyrimidine and the second choice would specify which purine or pyrimidine is present. Thus, at position 100 of Fig. 1, all of the 28S sequences have a C so that the position has two bits of conservation. In the logo (Fig. 2), a C appears at position 100 with a height of (almost) 2 bits. (A small sample correction prevents it from being exactly 2 bits high.[16])

For those positions where two equally likely bases occur, there is only one bit of information. This is because a choice of two things from four is equivalent to a choice of one thing from two. Position 86 is an example of this, in which five of the sequences contain A and four have T in Fig. 1. This position is therefore about one bit high in Fig. 2. The relative frequency of the bases determines the relative heights of the letters, and since A is more frequent, it is placed on top. Positions in which all four bases are equally likely not conserved and so have zero heights on the logo. When the frequencies of the bases are other than 0, 50, or 100%, the heights still measure the conservation at each position, but the calculation is a bit more complicated.[9,13,14,16] However, this method permits comparisons of the height of one position with any other.[17,18]

The sequence logo was used to choose the two PCR primers, as shown in Fig. 2, according to the following three criteria:

```
              10         20         30         40
  Hum  -----GTCCG GTGAGCTCTC GCTGGCCCTT GAAAATCCGG GGGAG
  Lem  --GGTGTCCG GTGCGCCCCC GGCGGCCCTT GAAAATCCGG AGGAC
Mrace  ------TCTG GTGCATTCAC AACGATCCTT GAAAATCCAA GGGAA
 Rice  --GGTGTCCG GTGCGCCCCC GGCGGCCCTT GAAAATCCGG AGGA-
Slime  --------GC GGTCTCCTTC CGTTGCCCTA GAAAAGCTGG CAGAT
  Tom  --GGTGTCCG GTGCGCTCCC GGCGGCCCTT GAAAATTCCG GAGGA
 Worm  GTGGTGTCTC GTGCTCTTTG AACGGCCCTT AAAACACCAA GGGAG
 Xlrn  -GGCGTCCGG TGAGCTTCTC GCTGGCCCTT GAAAATCCGG GGGAG
Yeast  --TGGCTCCG GTGCGCTTGT GACGGCCCGT GAAAATCCAC AGGA-

         50         60         70         80         90
  Hum  AGG-- ----GTGTAA ATCTC-GCGC CGGGCCGTAC CCATATCCGC
  Lem  CG--- ----AGTG-- CCGCCCGCGC CCGGTCGTAC TCATAACCGC
Mrace  A---- -----GAATA ATTTTCTCGC CTAGTCGTAC TCATAACCGC
 Rice  ----- ---CCGAGTA CCGTCCACGC CCGGTCGTAC TCATAACCGC
Slime  GGGTG AAACGTGTTG TCCTTCG-GT TGAACCGTAC CTA-ATCCGC
  Tom  CCGAA TGCCGT---- ---TCCACGC CCGGTCGTAC TCATAACCGC
 Worm  GCTAT -------TAA TT---TGCAC TCAATCGTAC CGATATCCGC
 Xlrn  AGG-- ----GTGTAA ATCTCTGCGC CGGGCCGTAC CCATATCCGC
Yeast  ----- ---AGGAATA GTTTTCATGC TAGGTCGTAC TGATAACCGC

             100        110        120        130
  Hum  AGCAGGTCTC CAAGGTGAAC AGCCTCTGGC ATGTTGGAAC AATGT
  Lem  ATCAGGTCTC CAAGGTGAAC AGCCTCTGG- TCGATGGAAC AATGT
Mrace  AGCAGGTCTC CAAGGTGAAA AGCCTCTAG- TTGATAGAAC AATGT
 Rice  ATCAGGTCTC CAAGGTGAAC GACCTCTGGC -CAATGGAAG AATGT
Slime  AGCAGGTCTC CAAGATGAGC AGTCTCTGGC GCATAGAACA AAGTA
  Tom  ATCAGGTCTC CAAGGTGAAC AGCCTCTGG- TCGATGGAAC AATGT
 Worm  ATTAGGTCTC CAAGGTGAAC AGCCTCTAG- TCGATAGAAT AATGT
 Xlrn  AGCAGGTCTC CAAGGTGAAC AGCCTCTGGC ATGTTAGAAC AATGT
Yeast  AGCAGGTCTC CAAGGTGAAC AGCCTCTAG- TTGATAGAAT AATGT

        140        150        160        170
  Hum  AGGTA AGGGAAGTCG GCAAGCCGGA TCCGTAACTT CGG
  Lem  AGGCA AGGGAAGTCG GCAAAATGGA TCCGTAACTT CGG
Mrace  AGATA AGGGAAGTCG GCAAAATAGA TCCGTAACTT CGG
 Rice  AGGCA AGGGAAGTCG GCAAAACGGA TCCGTAACTT CGG
Slime  GCGTA AGGGAATTCG GCAAGCCGGA TTCGTAACTT CGG
  Tom  AGGCA AGGGAAGTCG GCAAAATGGA TCCGTAACTT CGG
 Worm  AGGTA AGGGAAGTCG GCAAACTAGA TCCGTAACTT CGG
 Xlrn  AGGTA AGGGAAGTCG GCAAGTCAGA TCCGTAACTT CGG
Yeast  AGATA AGGGAAGTCG GCAAAATAGA TCCGTAACTT CGG
```

Fig. 1. Aligned sequences for part of the 28S rDNA from nine species. Hum: *Homo sapiens* (humans); Lem: *citrus limon* (lemon); Mrace: *Mucor recemosis* (zygomycete, a fungus); Rice: *Oryza sativa* (rice); Slime: *Physarum polycephalum* (slime mold); Tom: *Lycopersican esculentum* (tomato); Worm: *Caenorhabditis elegans* (nematode worm); Xlrn: *Xenopus laevis* (South American toad), and Yeast: *Saccharomyces cerevisiae* (baker's yeast).

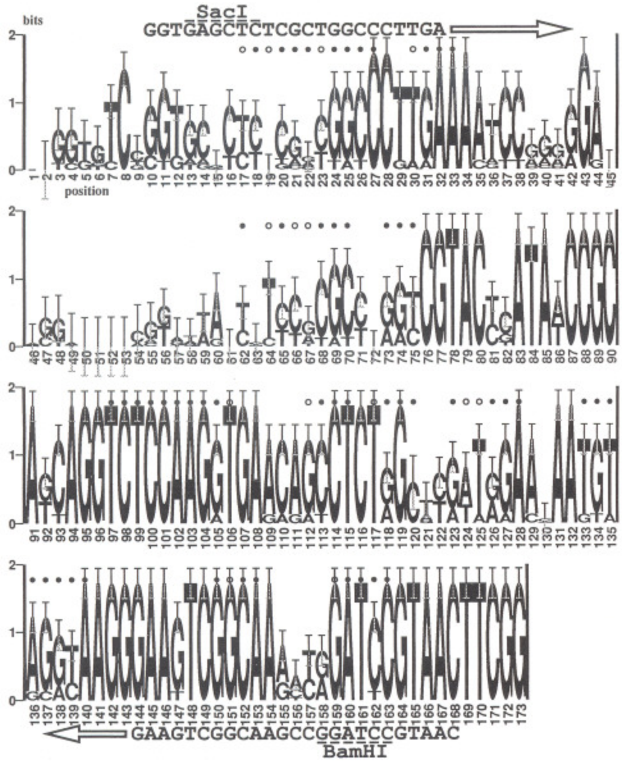

Fig. 2. Sequence logo created from the sequences shown in Fig. 1. The horizontal axis represents the position of nucleotides in the alignment; these correspond to Fig. 1. The vertical axis represents information (i.e., sequence conservation) in bits. Error bars indicate the standard deviation of the height of each stack. (This represents the expected variation of the conservation due to the small (finite) sample of aligned sequences.) See text for further description. The sequence of the *Sac*I primer is as shown. The complement of the *Bam*HI primer is shown for clarity. (The actual sequence of the *Bam*HI primer on the other strand of DNA is found by switching A's with T's and C's with G's and writing the sequence backwards: 5′ GTTACGGATCCGGCTTGCCGACTTC 3′.) The primers are identical to the human 28S rDNA sequences. The 5′ and 3′ terminal coordinates of the PCR product correspond to positions 2698 and 2849 of the human sequence (GenBank entry HUMRGM, accession number M11167). Proposed RNA structure in *X. laevis* 28S RNA[2] is shown by paired bases (•) and G-U base pairs (○). Unmarked bases are single stranded. The sequence logo programs are written in standard (i.e., portable) Pascal,[6] and most have been translated into C. They are available by anonymous ftp (file transfer protocol) from ncifcrf.gov in directory pub/delila. See the README file there for further information. Electronic mail contact: toms@ncifcrf.gov.

- They are in regions of high conservation, and surround regions of low conservation.
- The 3′ termini cover regions that are not variable, so that the primer end which is extended by the DNA polymerase is always properly annealed to the DNA.
- The oligonucleotide primers are not self-complementary and do not base pair to each other.

Because these primers were also designed to guarantee amplification of the human sequence, the 5′ terminus of the left primer was not highly conserved. This had no effect on cross species amplification. The primers contain restriction sites useful for subsequent cloning of the amplification products (results not shown). The primers cover DNA from base 10 through 168 (Fig. 1), so the amplified human product was predicted to be 159 base pairs long. When PCR reactions were carried out on genomic DNA purified from 12 different organisms a major product was observed (Fig. 3). Except for *H. sapiens*, none of the species tested were included in the original sequence alignment. This result demonstrates that the method can be applied to new species. A number of mammalian species were amplified as well as several different classes of fungi. The set of fungal species that were successfully tested is particularly significant, as ribosomal sequences within this subkingdom exhibit a very high degree of diversity.[19] The DNA sequences of the cloned *H. sapiens*, *M. musculus*, and *S. cerevisiae* amplification products corresponded precisely with published reports (data not shown), suggesting that amplification was of high fidelity and that contaminating genomic templates were not present.[10] The phylogenetic relationships inferred from the 28S sequences of the other species were generally compatible with the known taxonomic relationships between these organisms.[10]

Minor amplification products are observed in several of these reactions (*L. catta*, *P. infestans*, *M. musculus*, *P. pinus*, and *S. cerevisiae* lanes). The fragments may have arisen from amplification at sites within these genomes which happen to be complementary to our primers. This possibility is, in part, a function of the annealing temperature of the PCR reaction, which was deliberately chosen to be permissive for amplification of rDNA from a wide variety of extant species. Alternatively, since organisms have many copies of the 28S gene, some of these may have large variations in the amplified regions. In particular, the region selected for amplification is adjacent to a target sequence which is commonly interrupted by a site-specific transposable element

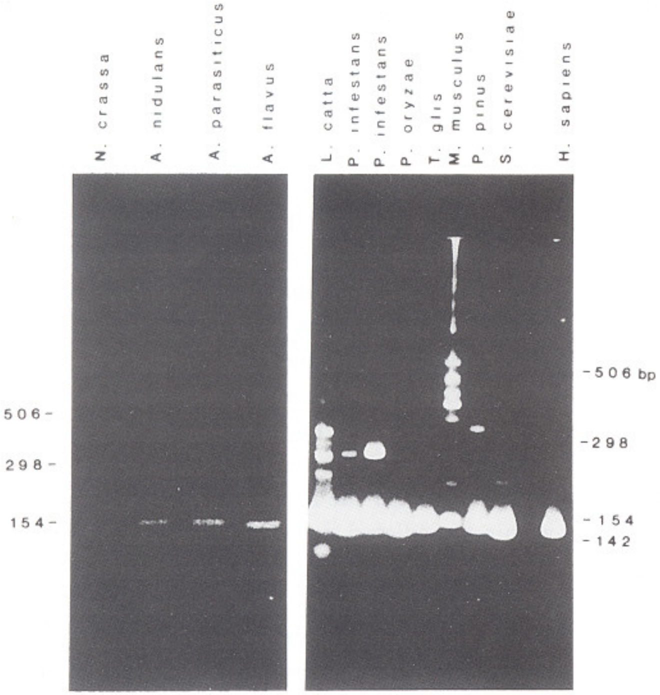

Fig. 3. PCR amplification reaction products. The following conditions were used for PCR amplification: 1 unit of Taq DNA polymerase (Perkin Elmer Cetus) was used. After incubation of oligonucleotide primers and genomic DNA for 4′ at 94°C, 40 cycles were performed at 94°C, 1′; 55°C, 1′; 72°C, 2′ followed by a synthesis step (72°C,7′). The products were separated on a 4% Nusieve agarose gel (Seakem). The gel was stained with ethidium bromide and photographed under ultraviolet light. Reactions (from left to right) are: *Neurospora crassa*, *Aspergillus nidulans*, *Aspergillus parasiticus*, *Aspergillus flavus* (fungal species), *Lemur catta* (brown lemur), *Phytophthora infestans* (Irish potato blight of 1840, strain 506), *Phytophthora infestans* (strain 10126), *Pyricularia oryzae* (rice blast fungus), *Tupaia glis* (tree shrew), *Mus musculus* (domestic mouse), *Pichia pinus* (a soil fungus), *Saccharomyces cerevisiae* (baker's yeast), and *Homo sapiens*.

in some organisms (i.e., insects.[5]) Thus, the evolutionary instability of this region of the 28S gene may give rise to variants which, upon amplification, could produce these minor species-specific PCR products. These artifacts do not interfere with subsequent isolation, cloning, and sequencing of the desired products.[11]

Some nucleotide sequences within the RNA molecule are complementary to one another because the formation of these duplexes is required for normal ribosome function. The proposed secondary structure of base-paired nucleotides in the amplified region of the 28S RNA molecule[2] does not correlate with the information content at each position (Fig. 2). In part, this is because the information measure we used shows conservation at each position and ignores possible cross correlations. For example, if an A is paired to U somewhere in the structure, and then the A is mutated to C during evolution, the other base must become G for the structure to be maintained. However, compensatory mutations are not always found in duplex structures.[8] The fixed bases would display high information content in a sequence logo.

In the 28S RNA sequence that we analyzed, the predicted hairpin structures do not always exhibit a high information content (e.g., positions 62-74), although at least one of the potential duplexes is highly conserved (e.g., positions 97-106). Conversely, although the most variable stretch of nucleotides in this region is found within a predicted single stranded region (positions 33-61), this is not always the case, as some single stranded domains are also highly conserved (e.g., positions 76-96, 164-173). This is consistent with the possibility that these sequences make specific interactions with ribosomal proteins. It is not surprising then, that the 28S sequence logo derived here does not uniformly correlate with the postulated secondary structure.

Discussion

The consensus sequence of an aligned set of sequences is constructed by choosing the most frequent base at each position. Consensus sequences were not used to locate conserved and non-conserved regions because they are not quantitative and they destroy data about the frequency of base occurrences. Because the sequence logo measures sequence conservation in a quantitative way, variable regions that may be useful in studying phylogeny can be rationally selected. Although we did not do so, it may be possible to design degenerate PCR primers where no single nucleotide was present in all of the aligned sequences. Such an oligonucleotide should be complementary to most of the targets, but unfortunately it would also be complementary to many other sequences.

In summary, the sequence logo reveals evolutionary and, by inference, functional conservation of specific nucleotide sites in the ribosomal sequence. However, when applied in molecular phylogenetic analysis, it is simply a tool for assessing the most variable or conserved segments in the rDNA sequence. A highly variable domain flanked by highly conserved segments was identified in the sequence logo and the two conserved elements were used to design oligonucleotide primers for PCR amplification. These primers were shown to be suitable for amplification of the variable region in a wide variety of eukaryotes. The DNA sequence of these PCR amplification products can be used to determine taxonomic relationships or to quickly place a previously unidentified species on the evolutionary tree.

References

1. Burks, C., Cassidy, M., Cinkosky, M.J., Cumella, K.E., Gilna, P., Hayden, J.E.-H., Keen, G.M., Kelley, T.A., Kelly, M., Kristofferson, D. and Ryals, J., "GenBank", *Nucl. Acids Res.* **19**, 2221–2225 (1991).
2. Clark, C.G., Tague, T.W., Ware, V.C. and Gerbi, S.A., "*Xenopus laevis* 28S ribosomal RNA: Secondary structure model and its evolutionary and functional implications", *Nucl. Acids Res.* **12**, 6197–6220 (1984).
3. Feng, D.F. and Doolittle, R.F., "Progressive sequence alignment as a prerequisite to correct phylogenetic trees", *J. Mol. Evol.* **25**, 351–360 (1987).
4. Higgins, D.G. and Sharp, P.M., "Fast and sensitive multiple sequence alignments on a microcomputer", *CABIOS* **5**, 151–153 (1989).
5. Jakubczak, J.L., Burke, W.D. and Eickbush, T.H., "Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects", *Proc. Natl. Acad. Sci. USA* **88**, 3295–3299 (1991).
6. Jensen, K. and Wirth, N., *Pascal User Manual and Report* (Springer-Verlag, 1975).
7. Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L. and Pace, N.R., "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses", *Proc. Natl. Acad. Sci. USA* **82**, 6955–6959 (1985).
8. Michel, F. and Westhof, E., "Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis", *J. Mol. Biol.* **216**, 585–610 (1990).
9. Pierce, J.R., *An Introduction to Information Theory: Symbols, Signals and Noise*, 2nd ed. (Dover Publications, 1980).
10. Rogan, P.K., Salvo, J.J. and Tooley, P.W., "Amplification of 28S ribosomal DNA with universal PCR primers", *Proceedings of the IV International Congress of Systematic and Evolutionary Biology*, University of Maryland, College Park, MD, Vol. 2 (1990) pp. 393.
11. Rogan, P.K. and Salvo, J.J., "High-fidelity amplification of ribosomal gene sequences from South American mummies", *Ancient DNA*, eds. B. Herrmann and S. Hummel (Springer-Verlag, 1994).

12. Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N., "Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia", *Science* **230**, 1350–1354 (1985).
13. Schneider, T.D., "Theory of molecular machines. I. Channel capacity of molecular machines", *J. Theor. Biol.* **148**(1), 83–123 (1991a). (Note: The figures were printed out of order. Figure 1 is on pp. 97).
14. Schneider, T.D., "Theory of molecular machines. II. Energy dissipation from molecular machines", *J. Theor. Biol.* **148**(1), 125–137 (1991b).
15. Schneider, T.D. and Stephens, R.M., "Sequence logos: A new way to display consensus sequences", *Nucl. Acids Res.* **18**, 6097–6100 (1990).
16. Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A., "Information content of binding sites on nucleotide sequences", *J. Mol. Biol.* **188**, 415–431 (1986).
17. Shannon, C.E., "A mathematical theory of communication", *Bell System Tech. J.* **27**, 379–423, 623–656 (1948).
18. Shannon, C.E. and Weaver, W., *The Mathematical Theory of Communication* (University of Illinois Press, 1949).
19. Walker, W.F., "5S and 5.8S ribosomal RNA sequences and protist phylogenetics", *Biosystems* **18**, 269–278 (1985).
20. Woese, C.R., "Bacterial evolution", *Microbiol. Rev.* **51**, 221–271 (1987).

Glossary

***5′ and 3′*:** The ends of a single strand of DNA are named 5′ and 3′ after the names of carbon atoms in the deoxyribose sugar. These define direction along the DNA. In living organisms, the strands of DNA are always synthesized from the 5′ end toward the 3′ end by DNA polymerases. Ends are also called "termini".

***complement*:** In double stranded DNA or RNA, the bases on opposite strands are said to be "complementary" because their surfaces fit together nicely. A is complementary to T (or U) and G is complementary to C.

***DNA*:** Deoxyribonucleic acid, the genetic material. There are four chemical letters (called bases) in DNA: A, C, G, and T. These are connected together by alternating deoxyribose sugar and phosphate groups to form long molecules. A gene is composed of a series of these bases. The order of bases in a DNA is called its "sequence", and "sequencing DNA" means to read the DNA letters. Usually, two strands of DNA are wound around each other to form the famous "double helix". It looks like a twisted ladder, with the rungs made of two bases called "base pairs". If there is an A on one side of a rung, there will be a T on the other (and *vice versa*). Also, if there is a C on one side of a rung, there will be a G on the other (and *vice versa*). During DNA replication, each strand is copied by an enzyme called DNA polymerase. This produces two "daughter" DNA molecules identical to the original parent molecule.

***enzyme*:** A protein which catalyzes (speeds up) a chemical reaction. An example is DNA polymerase, which creates DNA strands by polymerizing the four bases together.

***eukaryote*:** (literally, true nucleus) A living organism whose cells contain membrane bounded nuclei in which their DNA is held. Examples are humans and plants. Procaryotes (literally, before nuclei) do not have nuclei. Examples are bacteria and their viruses.

***gene*:** A segment of DNA which codes for an RNA or protein. Genes perform distinct functions in the cell.

***GenBank*:** One of the three world repositories for genetic sequence information. (See Burks, C. *et al.*, *GenBank. Nucl. Acids Res.* **19**, 2221–2225 (1991).)

***interstitial region*:** In the context of amplifying a piece of DNA by PCR, the region between two PCR primers.

***PCR*:** See polymerase chain reaction.

***polymerase chain reaction*:** (PCR) A method of making many copies of ("amplifying") a DNA. PCR can be started even from a single molecule. The method relies on the ability of the enzyme DNA polymerase to make copies of DNA in the test tube. To start a PCR reaction, one puts some DNA polymerase into a tube, along with the four bases, some DNA and two DNA primers. Primers are pieces of synthetic DNA 15 to 20 bases long. The solution is heated to make the DNA open up. This allows the primers to stick to the DNA when the solution is cooled. They stick in places where their bases match (complement) the bases of the DNA (A to T and C to G). The DNA polymerase binds to the primer/DNA complex and starts replicating the DNA. The two primers are chosen so that they are not too far apart, and so that the DNA polymerases which start the replications will be going toward each other. Because they are on different strands, they will not collide, but will pass each other by. This produces two molecules identical to the parent molecule, but starting at the primers. The solution is heated up again to separate all the strands, and when it is cooled, the primers stick to the DNA again. After a series of heating and cooling cycles, the region of DNA between (and including) the primers is replicated far faster than any other piece of DNA. It is therefore "amplified". The cyclic heating and cooling is performed by a machine, and special heat insensitive DNA polymerases (from hot springs bacteria) are often used.

***protein*:** A long string of amino acids, which folds up into a particular shape. Proteins can be used for structural purposes (e.g., keratin in hair, nails and horns) or they can do things (e.g., cenzymes like DNA polymerase).

***ribosome*:** The large enzyme in cells which translates RNA into protein. The ribosome is made up of many pieces of RNA and protein.

***RNA*:** Ribonucleic acid, a copy of the genetic material (see DNA). In RNA there are four chemical letters (called bases): A, C, G and U (instead of T). Also, the connecting backbone has ribose instead of deoxyribose. RNA is a copy of the DNA which usually survives for a only few minutes in the cell before being broken down. Ribosomal RNAs, however, last indefinitely. The DNA copy of the ribosomal RNA is called an rDNA. RNA usually has only one single strand, but small portions of it can loop back on themselves to make a so-called "hairpin". These twist together like the structure in DNA. Hairpin loops only form if sequences of at least four base pairs in length are complementary. RNA molecules fold up into complex three-dimensional structures. We usually do not know what these "tertiary" structures look like in detail, but maps showing which parts contact which other parts by complementary base pairing — called the "secondary structure" — can often be made. The "primary structure" is the sequence itself. Some rather complex secondary structures have been discovered in RNA, especially in the ribosomal RNAs.

***taxonomy*:** The classification of organisms into related groups. The final product is an evolutionary tree. By counting the number of changes between sequences from two species, we can estimate how long it has been since they had a common ancestor.