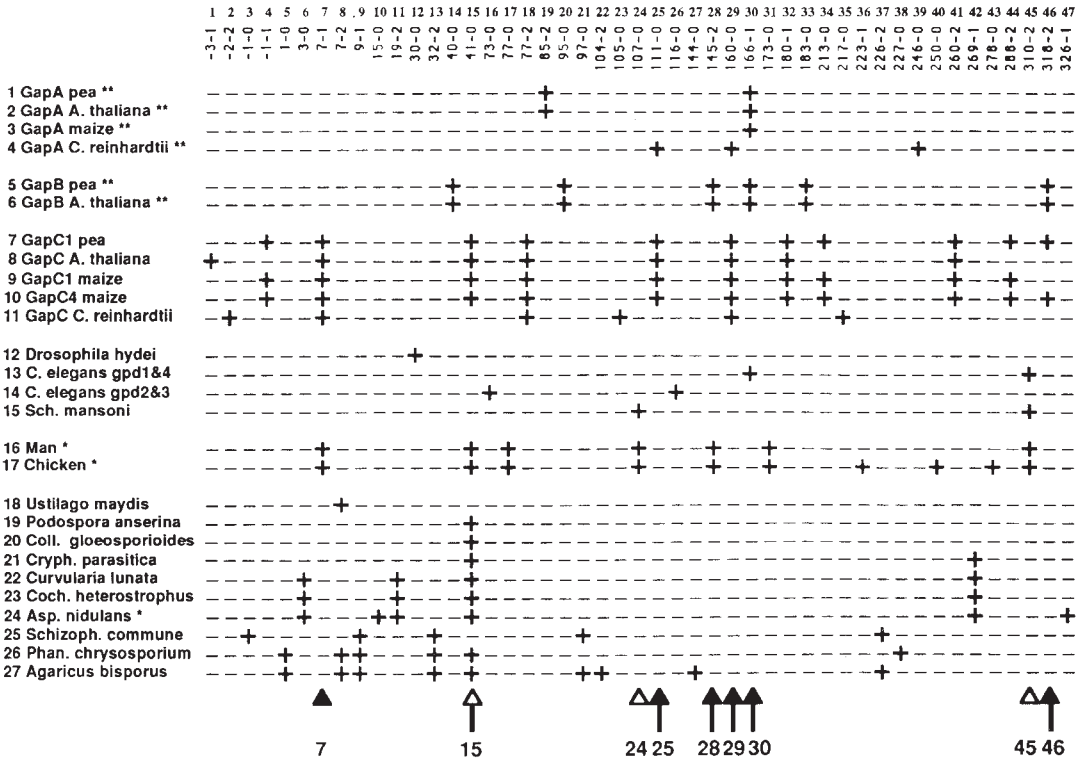
Five identical intron positions in ancient duplicated genes of eubacterial origin

Ralf Kersanach, Henner Brinkmann, Marie-Françoise Liaud, De-Xing Zhang*, William Martin & Rüdiger Cerff†

Institut für Genetik, Universität Braunschweig, Postfach 3329, D-38023 Braunschweig, Germany
* Present address: School of Biological Sciences, University of East Anglia, Norwich NR4 71J, UK
† To whom correspondence should be addressed

In 1985 Cornish-Bowden wrote "although there is now much to suggest that introns are an ancient relic of primordial genes, convincing proof must await the discovery of clearly corresponding intron arrangements in genes that arose by duplication before the separation of prokaryotes and eukaryotes"[1]. Genes for chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases of eukaryotes are descendants of an ancient gene family that existed in the common ancestor of extant eubacteria. During eukaryotic evolution, both genes were transferred to the nucleus from the antecedents of present-day chloroplasts and mitochondria, respectively[2–5]. Here we report the discovery of five spliceosomal introns at positions that are precisely conserved between nuclear genes for this chloroplast/cytosol enzyme pair. These data provide strong evidence in favour of the 'introns early' hypothesis, which proposes that introns were present in the earliest cells, consistent with the idea that introns facilitated the assembly of primordial genes by accelerating the rate of exon shuffling[6–13].

Two introns strictly conserved in nuclear genes encoding chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases (GAPDH; *GapA/GapB* and *GapC*, respectively) have been generally regarded as strong evidence in favour of the 'introns early' hypothesis[10], although the identity of intron positions has also been dismissed as 'parallel insertion of different introns'[14]. To obtain more information on intron conservation in GAPDH genes, we have determined gene structures for *GapA* and *GapC* from the unicellular green alga *Chlamydomonas reinhardtii* and for two new *GapC* genes from higher plants, *GapC1* of pea and *GapC4* of maize. We found three new intron positions that are precisely conserved between chloroplast and cytosolic GAPDH genes. Figure 1 shows that there are 47 known intron positions in 27 different GAPDH genes. Of the ten intron positions found in chloroplast GAPDH genes (*GapA* and *GapB*;



* One intron in leader, ** two introns in N-terminal transit peptide

Introns conserved between GapA(B)/GapC, plants/animals/fungi, plants/animals and vertebrates/non-vertebrates, respectively.

FIG. 1 The distribution of introns in present-day GAPDH genes. The presence of an intron is indicated by a plus sign, its absence by a minus sign. There are 47 known intron positions, defined by codon number and phase, in 27 different genes. For introns separating triplets after the first or second base codon numbers are followed by −1 and −2, respectively. For introns falling between triplets the number of the codon located 3′ of the intron is given followed by −0. Codon positions are numbered with reference to the standard sequence alignment of GAPDH genes[2,3,12,13]. The 27 genes are shown from top to bottom as six separate groups comprising genes encoding: lines 1–4, chloroplast *GapA*; lines 5 and 6, chloroplast *GapB*; lines 7–11, glycolytic *GapC* genes from plants; lines 12–15, non-vertebrates; lines 16 and 17, vertebrates; and lines 18–27, fungi. Arrows and arrowheads at the bottom of the figure indicate intron positions precisely conserved across *GapA(B)/GapC* genes (introns 25, 28, 29, 30, 46) and across *GapC* genes of more or less distant eukaryotic organisms (introns 7, 15, 24, 45), as indicated at the bottom of the figure.

METHODS. Cloning and sequencing of cDNAs and genomic clones encoding *GapA* and *GapC* of *Chlamydomonas reinhardtii*, *GapC4* of maize and *GapC1* of pea has been described[3,13,15,22,23]. cDNA clones of *Chlamydomonas* encoding *GapA* and *GapC* were identified using heterologous probes of maize[3]. The genomic clone encoding *GapC4* of maize was identified with a cDNA probe encoding maize *GapC3* (ref. 24). The genomic clone encoding pea *GapC1* was isolated using the authentic cDNA probe[22]. Intron/exon arrangements were established by comparing the genomic sequences with their corresponding cDNA sequences, except for *Chlamydomonas GapC*, for which intron positions were deduced by homology comparisons with genes from angiosperms. Sources of sequence information: *Chlamydomonas GapA* and *GapC*, pea *GapC1* and maize *GapC4* (lines 4, 7, 10, 11), this work; maize *GapA* (line 3)[12]; pea *GapA* and *GapB* (lines 1, 5)[13]; maize *GapC1* (line 9)[23]; *GapC* genes from *Schizophyllum commune*, *Phanerochaete chrysosporium* and *Agaricus bisporus* (lines 25 to 27)[25]; all other sequences can be found in release 35 of GenBank. *A. thaliana*, *Arabidopsis thaliana*; *C. reinhardtii*, *Chlamydomonas reinhardtii*; *C. elegans*, *Caenorhabditis elegans*; *Sch. mansoni*, *Schistosoma mansoni*; *Coll. gloeosporioides*, *Colletotrichum gloeosporioides*; *Cryph. parasitica*, *Cryphonectria parasitica*; *Coch. heterostrophus*, *Cochliobolus heterostrophus*; *A. nidulans*, *Aspergillus nidulans*; *Schizoph. commune*, *Schizophyllum commune*; *Phan. chrysosporium*, *Phanerochaete chrysosporium*.

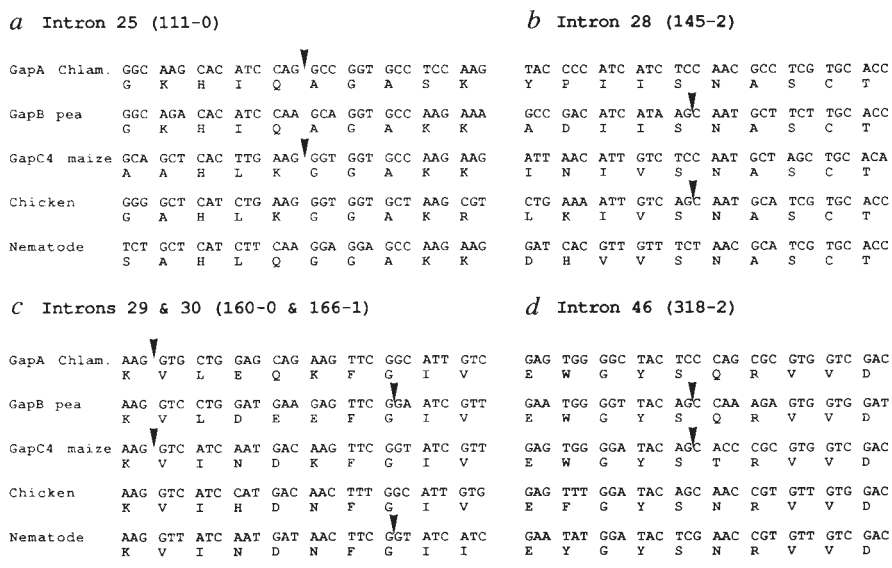


FIG. 2 Conservation of intron-exon junctions between chloroplast and glycolytic GAPDH genes, *GapA(B)* and *GapC*, respectively. The alignments show that the five introns interrupt homologous coding sequences at identical positions. Intron designations and sources of sequence information are as in Fig. 1. For the estimation of the probability that the five intron positions coincide by chance, we considered *GapA/GapB* (999 bp, 6 genes, 10 different introns; see Fig. 1) versus *GapC* (999 bp, 21 genes, 42 different introns; see Fig. 1) as a pairwise comparison and determined a probability of 2.2×10^{-5} according to ref. 26. If fungal genes with their highly non-random intron distributions (Fig. 1, genes 18 to 27) are excluded from this calculation, the probability becomes 2.2×10^{-6}.

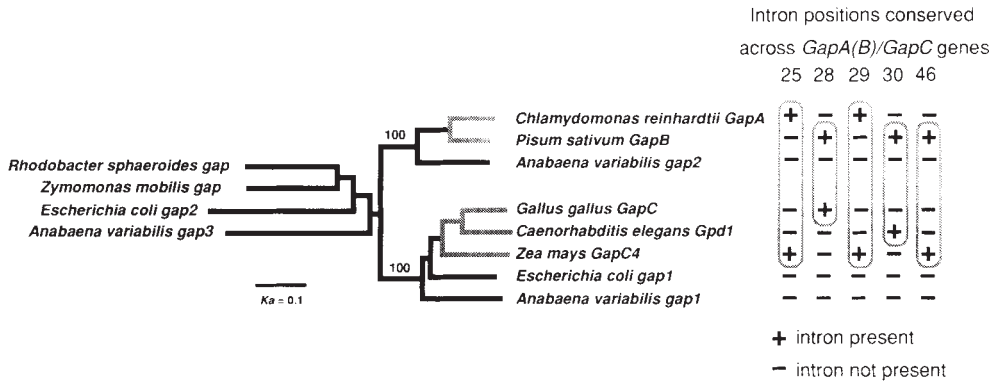


FIG. 3 Phylogenetic tree for GAPDH sequences showing intron conservation patterns. The tree was inferred by the neighbour-joining method[27] from a divergence matrix of non-synonymous substitutions per non-synonymous site, Ka[28]. Branches bearing genes found in eubacterial genomes are shown as solid lines, those bearing genes found in eukaryotic genomes are shown as grey lines. The scale bar indicates 0.1 substitutions per site. The numbers above the *GapA* and *GapC* branches indicate that these were found in 100/100 bootstrap parsimony replicates (PAUP, version 3.0) for the amino-acid alignment; these are the only branches in the figure that are essential to arguments concerning the age of introns conserved across *GapA(B)/GapC* genes (see text). Designations of intron positions correspond to those indicated in Figs 1 and 2. The *E. coli* GAPDH genes were designated in the figure as *gap1* and *gap2* instead of *gapA* and *gapB*, respectively, as in the original literature[29], in order to avoid confusion with the plant *GapA* and *GapB* genes[4]. Sources of bacterial sequences are given in[4], with the exception of Rhodobacter sphaeroides[30]. For all other sources, see Fig. 1 legend.

lines 1 to 6 in Fig. 1), five are precisely conserved in glycolytic GAPDH (*GapC*) genes of plants and animals: introns 28 and 30 (corresponding to positions 145-2 and 166-1) in vertebrates (lines 16 and 17 in Fig. 1) and nematodes (line 13), respectively[12,13], and the three new introns 25, 29 and 46 (positions 111-0, 160-0 and 318-2) in *GapC* genes of higher plants (Fig. 1, lines 7 to 10). In addition to these precisely conserved introns across the *GapA(B)/GapC* boundary, there are four cases of conserved introns between *GapC* genes of different major eukaryotic groups (introns 7, 15, 24, 45) and numerous cases of quasi-conservation or 'slippage' of introns[15,23], in which positions differ by only one to eight bases both across gene classes and major taxonomic groups, such as introns B40-0/C41-0, B95-0/C97-0, C144-0/B145-2, C180-1/B183-0, C-3-1/C-2-2/C-1-0/C-1-1/C1-0 (plants/fungi), C7-1/C7-2 (plants/fungi), C30-0/C32-2 (*Drosophila*/fungi), C77-0/C77-2 (vertebrates/plants), C104-2/C105-0/C107-0 (fungi/*Chlamydomonas*/animals) C226-2/C227-0 (fungi/fungi), where B and C denote *GapB* and *GapC*, respectively.

Figure 2 shows a detailed alignment of the exon sequences flanking the five introns precisely conserved between chloroplast and cytosolic GAPDH genes. A representative set of five genes has been selected from Fig. 1, two genes encoding chloroplast GAPDH (*GapA* of *Chlamydomonas* and *GapB* of pea) and three encoding the cytosolic enzyme (chicken *GapC*, nematode *Gpd1* and maize *GapC4*). Clearly, in all five pairwise comparisons, introns interrupt the *GapA(B)* and *GapC* coding sequences in identical positions. This is also true for intron 25 (position 111-0), where the two flanking codons 110 and 111 have changed from Gln|Ala in *Chlamydomonas GapA* to Lys|Gly in maize *GapC4*. The probability of finding these five identical intron positions across the *GapA(B)/GapC* boundary as a result of independent insertion is extremely low and is estimated to be roughly 2×10^{-5} (Fig. 2 legend).

In Fig. 3, the same five sequences have been incorporated into a phylogenetic tree together with the GAPDH genes of the cyanobacterium *Anabaena variabilis* (genes *gap1, gap2* and *gap3*), the γ-purple bacterium *Escherichia coli* (genes *gap1* and *gap2*) and the α-purple bacteria *Rhodobacter sphaeroides* and *Zymomonas mobilis*. The tree topology and the corresponding bootstrap values clearly show that the duplication event that gave rise to chloroplast and cytosolic GAPDH genes of eukaryotes occurred long before the separation of distinct organismal lineages leading to present day eukaryotes and eubacteria. The subtrees for *GapA(B)* and *GapC* both bear genes found in eukaryotic and eubacterial chromosomes. Thus, either the duplication

event that gave rise to *GapA(B)* and *GapC* genes occurred in ancient eubacteria and the eukaryotic genes are of endosymbiotic origin, as we believe[2–5,12,13], or the gene duplication took place in the common ancestor of eubacteria and eukaryotes. In the former case, the five indisputably identical spliceosome intron positions in *GapA(B)/GapC* (Fig. 2) were occupied in eubacterial GAPDH genes long before the divergence of purple and cyanobacteria (*E. coli* and *A. variabilis*). In the latter case, they were occupied in the ancestral GAPDH gene of the progenote.

Although the five introns conserved across present-day *GapA(B)* and *GapC* genes are removed with the aid of spliceosomes, the data do not indicate how their predecessors in the ancestral GAPDH gene were spliced. The discovery of group II introns in eubacteria[16] has given support to the notion that these elements may have been the precursors from which both spliceosomal introns and small nuclear RNAs in eukaryotic nuclei arose[17,18–21]. Introns in *GapA(B)*- and *GapC*-ancestors within eubacterial chromosomes may have been mechanistically group II, and if so, may have evolved into contemporary spliceosomal introns *in situ*. The conservation of five introns at identical positions in GAPDH genes which were duplicated in ancient eubacteria and perhaps in progenotic DNA, lends strong support to the exon theory of genes[9]. □

Received 11 June; accepted 16 November 1993.

1. Cornish-Bowden, A. *Nature* **313,** 434–435 (1985).
2. Martin, W. & Cerff, R. *Eur. J. Biochem.* **159,** 323–331 (1986).
3. Brinkmann, H., Martinez, P., Martin, W., Quigley, F. & Cerff, R. *J. molec. Evol.* **26,** 320–328 (1987).
4. Martin, W., Brinkmann, H., Savona, C. & Cerff, R. *Proc. natn. Acad. Sci. U.S.A.* **90,** 8692–8696 (1993).
5. Liaud, M.-F. *et al. J. molec. Evol.* (in the press).
6. Dorit, R. L., Schoenbach, L. & Gilbert, W. *Science* **250,** 1377–1382 (1990).
7. Doolittle, F. & Stoltzfus, A. *Nature* **361,** 403 (1993).
8. Tittiger, C., Whyard, S. & Walker, V. K. *Nature* **361,** 470–472 (1993).
9. Gilbert, W. *Cold Spring Harbor Symp. quant. Biol.* **52,** 901–905 (1987).
10. Doolittle, W. F. in *The Hierarchy of Life* (eds Fernholm, B., Bremer, K. & Jörnvall, H.) 65–72 (Elsevier Science, Amsterdam, 1989).
11. Stone, E. M. & Schwartz, R. J. (eds). *Intervening Sequences in Evolution and Development* (Oxford University Press, Oxford, 1990).
12. Quigley, F., Martin, W. F. & Cerff, R. *Proc. natn. Acad. Sci. U.S.A.* **85,** 2672–2676 (1988).
13. Liaud, M.-F., Zhang, D. X. & Cerff, R. *Proc. natn. Acad. Sci. U.S.A.* **87,** 8918–8922 (1990).
14. Palmer, J. D. & Logsdon, J. M. *Curr. Opin. Gen. Dev.* **1,** 470–477 (1991).
15. Liaud, M.-F., Brinkmann, H. & Cerff, R. *Plant molec. Biol.* **18,** 639–651 (1992).
16. Ferat, J.-L. & Michel, F. *Nature* **364,** 358–361 (1993).
17. Cavalier-Smith, T. *Trends biochem. Sci.* **7,** 145–148 (1991).
18. Roger, A. J. & Doolittle, F. *Nature* **364,** 289–290 (1993).
19. Jarrell, K. A., Dietrich, R. C. & Perlman, P. S. *Molec. cell. Biol.* **8,** 2361–2366 (1988).
20. Suchy, M. & Schmelzer, C. J. *Molec. Biol.* **222,** 179–187 (1991).
21. Koch, J. L., Boulanger, S. C., Dib-Hajj, S. D., Hebbar, S. K. & Perlman, P. S. *Molec. cell. Biol.* **12,** 1950–1958 (1992).
22. Brinkmann, H., Cerff, R., Salomon, M. & Soll, J. *Plant. molec. Biol.* **13,** 81–94 (1989).
23. Martinez, P., Martin, W. F. & Cerff, R. *J. molec. Biol.* **208,** 551–565 (1989).
24. Russell, D. A. & Sachs, M. M. *Plant Cell* **1,** 793–803 (1989).
25. Harmsen, M. C. *et al. Curr. Genet.* **22,** 447–454 (1992).
26. Iwabe, N., Kuma, K., Kishino, H., Hasegawa, M. & Miata, T. *J. molec. Evol.* **31,** 205–210 (1990).
27. Saitou, N. & Nei, M. *Molec. Biol. Evol.* **4,** 406–425 (1987).
28. Li, W. H., Wu, C. I. & Luo, C. C. *Molec. Biol. Evol.* **2,** 150–174 (1985).
29. Alefounder, P. R. & Perham, R. N. *Molec. Microbiol.* **3,** 723–732 (1989).
30. Gibson, J. L. & Tabita, F. R. *J. Bact.* **170,** 2153–2158 (1988).

ACKNOWLEDGEMENTS. We thank M. Wettern for the preparation of *Chlamydomonas* mRNA, G. Last for his help with probability calculations, the Deutsche Forschungsgemeinschaft for financial support and the Gesellschaft für Biotechnologische Forschung Braunschweig for use of computer facilities. R. Kersanach is recipient of a doctoral grant from the Deutscher Akademischer Austauschdienst.

GUIDE TO AUTHORS

PLEASE follow these guidelines so that your manuscript may be handled expeditiously.

Nature is an international journal covering all the sciences. Contributors should therefore bear in mind those readers who work in other fields and those for whom English is a second language, and write clearly and simply, avoiding unnecessary technical terminology. Space in the journal is limited, making competition for publication severe. Brevity is highly valued. One printed page of *Nature*, without display items, contains about 1,300 words.

Manuscripts are selected for publication according to editorial assessment of their suitability and reports from independent referees. They can be sent to London or Washington and should be addressed to the Editor. Manuscripts may be dealt with in either office, depending on the subject matter, and will where necessary be sent between offices by overnight courier. High priority cannot be given to presubmission enquiries; in urgent cases they can be made in the form of a one-page fax. All manuscripts are acknowledged on receipt but fewer than half are sent for review. Those that are not reviewed are returned as rapidly as possible so that they may be submitted elsewhere without delay. Contributors may suggest reviewers; limited requests for the exclusion of specific reviewers are usually heeded. Manuscripts are usually sent to two or three reviewers, chosen for their expertise rather than their geographical location. Manuscripts accepted for publication are processed from the London office.

Nature requests authors to deposit sequence and crystallographic data in the databases that exist for this purpose.

Once a manuscript is accepted for publication, contributors will receive proofs in about 4 weeks. *Nature*'s staff will edit manuscripts with a view to brevity and clarity, so contributors should check proofs carefully. Manuscripts are generally published 2–3 weeks after receipt of corrected proofs. *Nature* does not exact page charges. Contributors receive a reprint order form with their proofs; reprint orders are processed after the manuscript is published and payment received.

Categories of paper

Review Articles survey recent developments in a field. Most are commissioned but suggestions are welcome in the form of a one-page synopsis addressed to the Reviews Coordinator. Length is negotiable in advance.

Progress articles review particularly topical developments for a nonspecialist readership. They do not exceed 4 pages in length. Suggestions may be made to the Reviews Coordinator in the form of a brief synopsis.

Articles are research reports whose conclusions are of general interest and which are sufficiently rounded to be a substantial advance in understanding. They should not have more than 3,000 words of text (not including figure legends) or more than six display items and should not occupy more than five pages of *Nature*.

Articles start with a heading of 50–80 words written to advertise their content in general terms, to which editors will pay particular attention. The heading does not usually contain numbers, abbreviations or measurements. The introduction to the study is contained in the first two or three paragraphs of the article, which also briefly summarize its results and implications. Articles have fewer than 50 references and may contain a few short subheadings.

Letters to Nature are short reports of outstanding novel findings whose implications are general and important enough to be of interest to those outside the field. Letters should have 1,000 or fewer words of text and four or fewer display items. The first paragraph describes, in not more than 150 words and without the use of abbreviations, the background, rationale and chief conclusions of the study for the particular benefit of non-specialist readers. Letters do not have subheadings and contain fewer than 30 references.

Commentary articles deal with issues in, or arising from, research that are also of interest to readers outside research.

News and Views articles inform nonspecialist readers about new scientific advances, sometimes in the form of a conference report. Most are commissioned but proposals can be made in advance to the News and Views Editor.

Scientific Correspondence is for discussion of topical scientific matters, including those published in *Nature*, and for miscellaneous contributions. Priority is given to letters of fewer than 500 words.

Preparation of manuscripts

All manuscripts should be typed, double-spaced, on one side of the paper only. An original and four copies are required, each accompanied by artwork. If photographs are included, five sets of originals are required; for line drawings, one set of originals and four good-quality photocopies are acceptable. Reference lists, figure legends and tables should all be on separate sheets, all of which should be double-spaced and numbered. Three copies of relevant manuscripts in press or submitted for publication elsewhere should be included with submitted manuscripts, clearly marked as such. Five copies of revised and resubmitted manuscripts, labelled with their manuscript numbers, are required, together with five copies of a letter detailing the changes made.

Titles are brief and simple. Active verbs, numerical values, abbreviations and punctuation are to be avoided. Titles should contain one or two key words for indexing purposes.

Artwork should be marked individually and clearly with the author's name and, when known, the manuscript number. Ideally, no figure should be larger than 28 by 22 cm. Figures with several parts are to be avoided and are permitted only if the parts are closely related, either experimentally or logically. Unlettered originals of photographs should be provided. Suggestions for cover illustrations, with captions and labelled with the manuscript number, are welcome. Original artwork is returned when a manuscript cannot be published.

Protein/nucleotide sequences should ideally be in the three-letter and not the single-letter code for amino acids. One column width of *Nature* can accommodate 20 amino acids or 60 base pairs.

Colour artwork. A charge of £500 per page is made as a contribution towards the cost of reproducing colour figures. Inability to pay these costs will not prevent publication of essential colour figures if the circumstances are explained. Proofs of colour artwork may be sent to contributors under separate cover from their galley proofs.

Figure legends should not exceed 300 words and ideally should be shorter. The figure is described first, then, briefly, the method. Reference to a method published elsewhere is preferable to a full description. Methods are not described in the text.

References are numbered sequentially as they appear in the text, followed by those in tables and finally by those in figure legends. Only papers published or in the press are numbered and included in the reference list. All other forms of reference should be cited in the text as a personal communication, manuscript submitted or in preparation. Text is not included in reference lists. References are abbreviated according to the *World List of Scientific Periodicals* (Butterworths, London, 1963–65). The first and last page numbers are included; reference to books includes publisher, place and date.

Abbreviations, symbols, units and Greek letters should be identified the first time they are used. Acronyms should be avoided whenever possible and, if used, defined. Footnotes are not used in the text.

Acknowledgements are brief and appear after the reference list; grant and contribution numbers are not allowed.

Supplementary Information is material relevant to Articles or Letters which cannot, for lack of space, be published in full, but which is available from *Nature* on request.

Submission. Manuscripts can be sent to the Editor at 4 Little Essex Street, London WC2R 3LF, UK or at 1234 National Press Building, Washington, DC 20045, USA. A telephone and fax number should be included. Manuscripts or proofs sent by air courier to London should be declared as 'manuscripts' and 'value $5' to prevent the imposition of import duty and value-added tax.