# A heuristic approach of maximum likelihood method for inferring phylogenetic tree and an application to the mammalian *SOX-3* origin of the testis-determining gene *SRY*

Kazutaka Katoh, Takashi Miyata*

*Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan*

Received 31 August 1999; received in revised form 12 November 1999

Edited by Takashi Gojobori

**Abstract Applying the tree bisection and reconnection (TBR) algorithm, we have developed a heuristic method (maximum likelihood (ML)-TBR) for inferring the ML tree based on tree topology search. For initial trees from which iterative processes start in ML-TBR, two cases were considered: one is 100 neighbor-joining (NJ) trees based on the bootstrap resampling and the other is 100 randomly generated trees. The same ML tree was obtained in both cases. All different iterative processes started from 100 independent initial trees ultimately converged on one optimum tree with the largest log-likelihood value, suggesting that a limited number of initial trees will be quite enough in ML-TBR. This also suggests that the optimum tree corresponds to the global optimum in tree topology space and thus probably coincides with the ML tree inferred by intact ML analysis. This method has been applied to the inference of phylogenetic tree of the SOX family members. The mammalian testis-determining gene *SRY* is believed to have evolved from *SOX-3*, a member of the SOX family, based on several lines of evidence, including their sequence similarity, the location of *SOX-3* on the X chromosome and some aspects of their expression. This model should be supported directly from the phylogenetic tree of the SOX family, but no evidence has been provided to date. A recently published NJ tree shows implausibly remote origin of *SRY*, suggesting that a more sophisticated method is required for understanding this problem. The ML tree inferred by the present method showed that the *SRY*s of marsupial and placental mammals form a monophyletic cluster which had diverged from the mammalian *SOX-3* in the early evolution of mammals.**

*Key words:* Maximum likelihood method; Heuristic approach; Molecular phylogeny; *SRY*; SOX family; Evolution

## 1. Introduction

The maximum likelihood (ML) method [1,2] for inferring the molecular phylogenetic tree is thought to be the most accurate method based on a solid statistical basis. The application of the ML method to actual problems, however, is limited to a small number of operational taxonomic units (OTUs). For more than nine OTUs or so, the problem is unable to solve, given a high speed computer. In order to overcome the inherent difficulty in the intact ML method, heuristic searches for finding the best tree topology would be a realistic strategy capable of handling actual problems with a large number of OTUs in practical time scales. Several heuristic approaches have already been proposed [3–8]. Combining two topology search algorithms, the nearest-neighbor interchange (NNI) method [6,9] and the subtree pruning and regrafting method [9], we recently described a heuristic approach of ML method which was applied to the reconstruction of gene family tree [8]. The heuristic approach is further extended in this paper by applying a more sophisticated topology search algorithm, the tree bisection and reconnection (TBR) method [9], and the new method is applied to the inference of the phylogenetic tree of the SOX family as an example, by which information on the evolutionary origin of the testis-determining gene *SRY* is provided.

The *SRY* encodes a transcription factor containing an ∼80 amino acid DNA-binding domain known as the HMG box [10–12] and is a member of the SOX family comprising a large number of members containing the HMG box in common [13–15]. The *SRY* is expressed for a brief period in the indifferent gonad and initiates male development in mammals [16,17]. The *SRY* homologs have been identified on the Y chromosomes of all eutherian and marsupial mammals examined [10,18]. From the observations that *SOX-3*, a member of the SOX family, is most similar in sequence to *SRY* and is located on the X chromosome in both eutherians and marsupials, together with some aspects of their expressions, it has been proposed that *SRY* was derived from *SOX-3* during the divergence of the Y chromosome from the ancestral X chromosome in the early evolution of mammals [18–20]. This is consistent with evidence that no *SRY* homolog has yet been identified in other classes of vertebrates to date (DDBJ release 37 and PIR database release 60). There is, however, still a possibility that *SRY* was originated from another member of the SOX family by gene duplication, followed by translocation or retroposition [20]. Indeed, a recently reported phylogenetic tree of the SOX family shows a remote divergence of *SRY* from the ancestral lineage of SOX-2/-3 subfamilies, the divergence time being very old, going back to date before the separation of vertebrates and arthropods [21]. There is, however, a possibility of an artifact derived from the procedure of tree inference by the neighbor-joining (NJ) method [22], because the evolutionary rate of *SRY* is remarkably high, as compared with other SOX family members [23–25], and the NJ method tends to infer the older divergence of rapidly evolving lineages than they really are. To understand the evolutionary origin of *SRY*, it is therefore important to re-exam-

*Corresponding author. Fax: (81)-75-753 4223.
E-mail: miyata@biophys.kyoto-u.ac.jp

0014-5793/99/$20.00 © 1999 Federation of European Biochemical Societies. All rights reserved.
PII: S0014-5793(99)01621-X

ine the phylogenetic tree of the SOX family by sophisticated methods.

The heuristic approach of ML analysis working on a PC cluster developed here has been applied to the inference of the phylogenetic tree of the SOX family. We report here that the inferred ML tree shows the mammalian *SOX-3* origin of *SRY*.

## 2. Materials and methods

### 2.1. Sequence data source

Sequence data used in the present analysis were taken from DDBJ release 37 and PIR database release 60. Only data containing the complete HMG box sequences were included in the analysis. For *SRY* genes, extremely rapidly evolving sequences were excluded, except for human *SRY*.

### 2.2. Sequence alignment

Multiple alignment of the amino acid sequences of the SOX family members was carried out by a method developed recently by us (Katoh et al., manuscript in preparation). This method is basically an extended version of the progressive approach of Feng and Doolittle [26]. By improving the calculation procedure of dynamic programming [27], the speed of computation has been greatly improved without sacrificing accuracy and efficiency. The computation time required by the new method is only about one-tenth of that by the standard method (for example, CLUSTAL W [28], a widely distributed multiple alignment program).

### 2.3. Phylogenetic tree inference

Using an elaborate tree topology rearrangement algorithm, called the TBR [9], a heuristic approach has been applied for inferring the ML tree of protein phylogeny [2]. This heuristic ML method consists of performing rearrangements of tree topology for a limited number ($N_I$) of initial trees by the TBR algorithm. The calculation procedure is as follows:

1. On the basis of the bootstrap resampling procedure [29], $N_I$ different initial trees (including that inferred from actual alignment) are generated by the NJ method [22] using the distance matrix estimated by the ML method [6] ($N_I = 100$ for the present case). Random tree topologies generated by connecting branches randomly were also examined as initial trees.
2. For a given initial tree, performing all possible topology rearrangements of the initial tree under TBR, a set of rearranged trees is generated.
3. For these rearranged trees generated from an initial tree, their approximate log-likelihoods are calculated by the method of Adachi and Hasegawa [6], using the JTT model (ProtML version 2.3 in Adachi and Hasegawa's program package MOLPHY), and the top 40% trees are selected by log-likelihood criterion.
4. For the set of trees selected in step 3, the intact ML analysis is performed and only one tree with the largest log-likelihood value among the rearranged trees is selected.
5. The steps 2–4 are repeated until no improvement on the log-likelihood is found. By the above iterative procedure, we finally obtain an optimum tree for a given initial tree. There is a possibility that this optimum tree corresponds to a local optimum in tree topology space.
6. Repeating the steps 2–5 for all different initial trees, we finally obtain the set of $N_I$ independent optimum trees.
7. The best optimum tree is selected from the $N_I$ independent optimum trees in log-likelihood criterion. It might be expected that the best optimum tree corresponds to the global optimum in tree topology space. Thus the best optimum tree is likely to be the ML tree.

In the case of the small number of OTUs, as in the present case, step 3 is skipped and all the trees generated in step 2 are subjected to the intact ML analysis. The above computational procedure was performed on a PC cluster composed of 32 Pentium III 500 MHz processors. Hereafter, we will designate the heuristic ML analyses based on TBR as ML-TBR. We will also distinguish the ML-TBR as ML-TBRb or ML-TBRr, depending on the difference (bootstrap trees or random trees) of the initial trees used.

The phylogenetic tree was also inferred by ProtML program developed by Adachi and Hasegawa [6], a heuristic ML method based on a simple tree topology rearrangement strategy, in which only NNI of branches are considered. The ProtML was also applied to the present problem and compared with our ML-TBR.

## 3. Results and discussion

### 3.1. Phylogenetic tree of the SOX family members

A variety of members belonging to the SOX family have already been identified from a diverse group of vertebrates and they are classified into several subfamilies which diverged by gene duplications in the early evolution of animals, possibly before the separation of vertebrates and arthropods [21]. In contrast, no *SRY* gene has been isolated from vertebrates other than mammals to date (DDBJ release 37 and PIR database release 60). The SOX family members share a highly conserved HMG box of ∼80 amino acids in common, but no appreciable sequence similarity is observed in regions outside the HMG box, except for closely related species. To understand the evolutionary origin of the *SRY* gene, it is important to compare sequences that are closely related to mammalian *SRY*s. According to the NJ tree of the SOX family members [21], mammalian *SRY*s are closely related to *SOX-2* and *SOX-3*, and the *SOX-9* and *SOX-11* are distantly related to SRY/SOX-2/-3 subfamilies. Thus, we compared these sequences, together with closely related sequences which are not included in the NJ tree by Soullier et al. [21]. Slowly evolving *SRY* sequences were used for tree inference, except for human sequence.

Alignment of the amino acid sequences was carried out for a region of the HMG box corresponding to amino acid sites 59–134 in human *SRY* sequence (comprising 75 amino acid sites in total, excluding gap positions) by the method described in Section 2. On the basis of the alignment, the phylogenetic tree of the SRY/SOX-2/-3 subfamilies was inferred by a heuristic ML method, ML-TBR, described in Section 2, together with the ProtML [6], using human *SOX-9* and *SOX-11* as an outgroup.

Fig. 1 shows the phylogenetic tree of the SRY/SOX-2/-3 subfamilies inferred from the ML-TBR, using 100 different NJ trees generated by bootstrap resamplings as initial trees ($N_I = 100$). The ML tree shows close association of marsupial and placental *SRY*s to form a monophyletic group. The NJ tree inferred from the same data set also shows monophyly of the mammalian *SRY*s (data not shown), although the bootstrap probability of monophyly is very low (25%). The NJ tree by Soullier et al. [21] shows separate origins of the marsupial and placental *SRY*s. An interesting result on the evolutionary origin of *SRY* was obtained from Fig. 1. The tree demonstrates that the common ancestor of the mammalian *SRY*s originated from the mammalian *SOX-3* lineage. The same tree topology was also obtained when randomly generated trees were used as initial trees. Furthermore, the ProtML also predicted the same tree in both cases of initial trees (bootstrap trees and random trees). This result is consistent with the model that *SRY* was derived from *SOX-3* [18–20]. This is also consistent with evidence that no *SRY* homolog has yet been identified in other classes of vertebrates. In addition, as Fig. 1 shows, the branch lengths of *SRY*s are extremely long, as compared with those of other members, implying the rapid accumulation of amino acid substitutions in *SRY* lineages

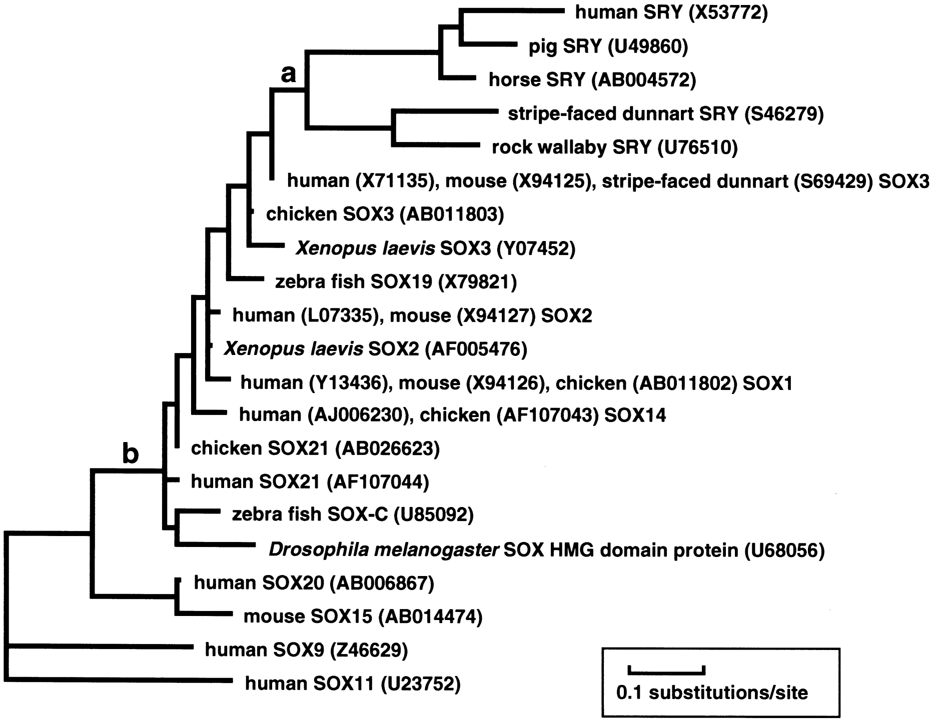
human SRY (X53772)
pig SRY (U49860)
a
horse SRY (AB004572)
stripe-faced dunnart SRY (S46279)
rock wallaby SRY (U76510)
human (X71135), mouse (X94125), stripe-faced dunnart (S69429) SOX3
chicken SOX3 (AB011803)
*Xenopus laevis* SOX3 (Y07452)
zebra fish SOX19 (X79821)
human (L07335), mouse (X94127) SOX2
*Xenopus laevis* SOX2 (AF005476)
human (Y13436), mouse (X94126), chicken (AB011802) SOX1
human (AJ006230), chicken (AF107043) SOX14
b
chicken SOX21 (AB026623)
human SOX21 (AF107044)
zebra fish SOX-C (U85092)
*Drosophila melanogaster* SOX HMG domain protein (U68056)
human SOX20 (AB006867)
mouse SOX15 (AB014474)
human SOX9 (Z46629)
human SOX11 (U23752)
0.1 substitutions/site

Fig. 1. ML tree of SOX family members. From a comparison of the HMG box sequences, the tree was inferred by the heuristic ML method ML-TBR described in Section 2, using human *SOX-9* and *SOX-11* sequences as an outgroup. One hundred initial trees were generated by applying the bootstrap resamplings and the NJ method. Accession numbers of sequences are shown in parentheses.

[23–25]. Recently, O'Neill et al. [30] identified an intron in *SRY* gene from marsupial groups and proposed intron insertion model, assuming that *SRY* was derived from *SOX-3* in the early evolution of mammals before the divergence of marsupial and placental mammals. The present result supports their argument.

The NJ tree inferred from the same data set shows the remote divergence of mammalian *SRY*s. The ancestral branch 'a' of mammalian *SRY*s diverges from the ancestral branch 'b' of *SOX-1, -2, -3, -14, -19, -21* and *Drosophila SOX*. A similar result is also found in the NJ tree by Soullier et al. [21]. The family tree was also examined by the maximum parsimony (MP) method [31], using the same data set. We obtained many MP trees, each of which requires a total of 163 amino acid substitutions. All these MP trees with the smallest number of substitutions show separate origins of marsupial and placental *SRY*s and also the remote divergence of placental *SRY*, which antedates the protostome-deuterostome split. An extremely rapid rate of *SRY* evolution might be responsible for this. Generally, there is a tendency in the NJ tree (and also in the MP tree) to prefer deep branching of rapidly evolving lineages. When the branch 'a' in Fig. 1 is connected with the branch 'b', the log-likelihood value decreases by 4.9 ± 9.5, relative to that of the ML tree of Fig. 1. Because the HMG box is short in amino acid length and is highly conserved, it is not possible to evaluate statistically the reliability of inferred tree at the significance level of 1 S.E.M.; the local bootstrap probability [6] that the mammalian *SOX-3* and *SRY* are clustered is not high enough (70%). Rather, the reliability should be examined by other lines of evidence supporting the tree. Many genes homologous to human Y-linked genes are identified on the X chromosome [32,33], supporting the hypothesis that the mammalian Y chromosome originated from the X chromosome [34–36]. Because *SOX-3* and *SRY* are located on the X and Y chromosomes in marsupial and placental mammals, respectively [18,20,32,33], the phylogenetic position of *SRY* in Fig. 1 is consistent with this hypothesis. In addition, from a comparison of the expression pattern of SOX-related genes, *SOX-1, SOX-2, SOX-3* and *SRY*, Collignon et al. [20] suggested that *SOX-3* is the closest relative of *SRY*. Thus, the present ML tree, together with evidence from chromosomal location and expression pattern, supports the mammalian *SOX-3* origin of *SRY*.

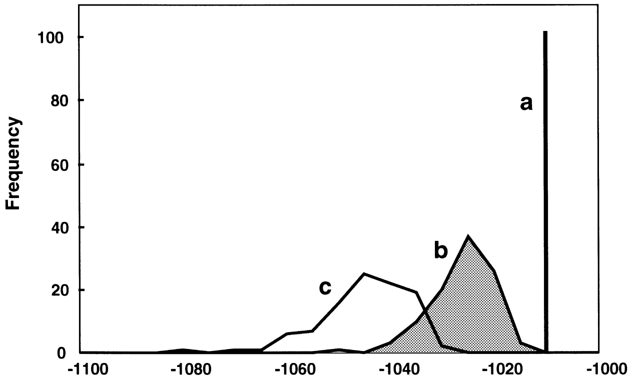

Fig. 2. Distribution of the log-likelihoods of the optimum trees inferred by heuristic ML methods. (a) ML-TBR, (b) ProtML, (c) the log-likelihoods of initial trees. One hundred initial trees were generated by applying the bootstrap resamplings and the NJ method, and for each initial tree, an optimum tree was obtained by the iterative procedures described in Section 2. Note that, in (a), all the 100 optimum trees converged to the best optimum tree with the largest log-likelihood value shown in Fig. 1.

### 3.2. Comparison of heuristic ML methods

As mentioned above, the identical ML tree of SOX family was inferred by ML-TBR and ProtML. Dependence of the initial tree used, however, differs greatly between the two methods. Fig. 2 shows distribution of the log-likelihood value of the optimum trees obtained from the steps 2–5 in Section 2 for each of 100 initial trees. In the ProtML, the log-likelihoods differ greatly for different initial trees used. Only one optimum tree coincided with the best optimum tree shown in Fig. 1, and in the remaining 99 optimum trees, their log-likelihood values were always lower than that of the best optimum tree. This might imply that there are many local optimum points in tree topology space and transition from one local optimum to another is impossible by NNI which is possible to generate only a small variation in tree topology. Note that, in the present case, the optimum tree obtained by using the NJ tree of actual alignment as an initial tree does not correspond to the best optimum tree of Fig. 1. These results suggest that many different initial trees are required to obtain the best tree with the largest log-likelihood value by ProtML.

Interestingly, in ML-TBR, all the optimum trees coincided with the best optimum tree. This may be due to the reason that transition from one local optimum to another with a larger log-likelihood value was facilitated by TBR that could generate a drastically distinct tree topology by a single operation. That is, ML-TBR allows one to search efficiently the ML tree from a limited number of initial trees. Furthermore, the result that all the optimum trees, particularly those obtained by using the random trees as the initial trees, converged on the best optimum tree strongly suggests that the obtained best optimum tree corresponds to the global optimum in tree topology space, and thus, it is highly likely that the best optimum tree coincides with the ML tree obtained by intact ML analysis.

*Acknowledgements:* We thank Drs. K. Kuma, N. Iwabe, H. Suga for discussion and comments. This work was supported by grants from the Ministry of Education, Science, Sports and Culture of Japan.

## References

[1] Felsenstein, J. (1981) J. Mol. Evol. 17, 368–376.
[2] Kishino, H., Miyata, T. and Hasegawa, M. (1990) J. Mol. Evol. 30, 151–160.
[3] Felsenstein, J. (1993) PHYLIP: phylogenetic inference package, Version 3.5c, Department of Genetics, University of Washington, Seattle, WA.
[4] Olsen, G.J., Matsuda, H., Hagstrom, R. and Overbeek, R. (1994) CABIOS 10, 41–48.
[5] Yang, Z. (1995) Phylogenetic analysis by maximum likelihood (PAML), Version 1.1, Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA.
[6] Adachi, J. and Hasegawa, M. (1996) Computer science monographs, No. 28, MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood, The Institute of Statistical Mathematics, Tokyo.
[7] Strimmer, K. and von Haeseler, A. (1996) Mol. Biol. Evol. 13, 964–969.
[8] Suga, H., Hoshiyama, D., Kuraku, S., Katoh, K., Kubokawa, K. and Miyata, T. (1999) J. Mol. Evol. (in press).
[9] Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996) Phylogenetic inference, in: Molecular Systematics (Hillis, D.M., Moritz, C. and Mable, B.K., Eds.), 2nd edn., pp. 407–514, Sinauer Associates, Sunderland, MA.
[10] Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischauf, A.M., Lovell-Badge, R. and Goodfellow, P.N. (1990) Nature 346, 240–244.
[11] Berta, P., Hawkins, J.R., Sinclair, A.H., Taylor, A., Griffiths, B.L., Goodfellow, P.N. and Fellous, M. (1990) Nature 348, 448–450.
[12] Jager, R.J., Anvret, M., Hall, K. and Scherer, G. (1990) Nature 348, 452–454.
[13] Harley, V.R., Jackson, D.I., Hextall, P.J., Hawkins, J.R., Berkovitz, G.D., Sockanathan, S., Lovell-Badge, R. and Goodfellow, P.N. (1992) Science 255, 453–456.
[14] Ferrari, S., Harley, V.R., Pontiggia, A., Goodfellow, P.N., Lovell-Badge, R. and Bianchi, M.E. (1992) EMBO J. 11, 4497–4506.
[15] Capel, B. and Lovell-Badge, R. (1993) The Sry gene and sex determination in mammals, in: Advances in Developmental Biology (Wassarmen, P., Ed.), Vol. 2, pp. 1–35, JAI Press.
[16] Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. and Lovell-Badge, R. (1991) Nature 351, 117–121.
[17] Hacker, A., Capel, B., Goodfellow, P. and Lovell-Badge, R. (1995) Development 121, 1603–1614.
[18] Foster, J.W. and Graves, J.A. (1994) Proc. Natl. Acad. Sci. USA 91, 1927–1931.
[19] Stevanovic, M., Lovell-Badge, R., Collignon, J. and Goodfellow, P.N. (1993) Hum. Mol. Genet. 2, 2013–2018.
[20] Collignon, J., Sockanathan, S., Hacker, A., Cohen-Tannoudji, M., Norris, D., Rastan, S., Stevanovic, M., Goodfellow, P.N. and Lovell-Badge, R. (1996) Development 122, 509–520.
[21] Soullier, S., Jay, P., Poulat, F., Vanacker, J.M., Berta, P. and Laudet, V. (1999) J. Mol. Evol. 48, 517–527.
[22] Saitou, N. and Nei, M. (1987) Mol. Biol. Evol. 4, 406–425.
[23] Whitfield, L.S., Lovell-Badge, R. and Goodfellow, P.N. (1993) Nature 364, 713–715.
[24] Tucker, P.K. and Lundrigan, B.L. (1993) Nature 364, 715–717.
[25] Pamilo, P. and O'Neill, R.J. (1997) Mol. Biol. Evol. 14, 49–55.
[26] Feng, D.F. and Doolittle, R.F. (1987) J. Mol. Evol. 25, 351–360.
[27] Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. 48, 443–453.
[28] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[29] Felsenstein, J. (1985) Evolution 39, 783–791.
[30] O'Neill, R.J., Brennan, F.E., Delbridge, M.L., Crozier, R.H. and Graves, J.A. (1998) Proc. Natl. Acad. Sci. USA 95, 1653–1657.
[31] Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package), version 3.5c, Department of Genetics, University of Washington, Seattle, WA.
[32] Graves, J.A. (1995) Bioessays 17, 311–320.
[33] Lahn, B.T. and Page, D.C. (1997) Science 278, 675–680.
[34] Ohno, S. (1967) Sex Chromosomes and Sex-linked Genes, Springer-Verlag, New York.
[35] Charlesworth, B. (1991) Science 251, 1030–1033.
[36] Hodgkin, J. (1992) Bioessays 14, 253–261.