EVOLUTION OF THE ets GENE FAMILY

Vincent LAUDET*, Christian NIEL, Martine DUTERQUE-COQUILLAUD,
Dominique LEPRINCE and Dominique STEHELIN

CNRS UA 1160, Institut Pasteur, 1 Rue Calmette, 59019 LILLE Cedex, FRANCE

Received November 24, 1992

Over the past few years a variety of genes have been described whose protein products share similarity with that of the c-ets-1 proto-oncogene, the cellular counterpart of the v-ets oncogene of the avian E26 retrovirus. This so-called "ets family" of transcription factors includes at least a dozen members present in several organisms. We have questioned the common evolutionary origin of these various gene products. By constructing phylogenetical trees with different methods, we show that the ets family is very ancient since the duplication of the various groups of ets related proteins occurred before the Arthropods / Vertebrates split (ca. 500 million years). © 1993 Academic Press, Inc.

The v-ets oncogene of the E26 retrovirus as well as its cellular counterpart the chicken c-ets-1 gene is the founder of a still growing family of transcription factors: the ets family (1-3). In addition to c-ets-1, this family comprises its proximal relative the c-ets-2 gene, the highly related erg and fli-1 genes, the mammalian elk-1, sap-1, elf-1, gabpα, spi-1, spi-B and pea3 genes as well as the Drosophila E74, D-elg, D-ets-2, D-ets-3, D-ets-4, D-ets-6 and pok/yan genes. The signature of the ets family is a stretch of 85 amino acids called the ETS domain which is largely conserved between all family members (3). This domain has been shown to be required for the DNA-binding activity of the various ets family members which all bind slightly divergent variations of a core motif C/A GGAA/T G/C.

Structural and functional data have allowed the emergence of a rough classification for the members of the ets family. For example, by comparison of its "ETS" domain, the spi-1/PU-1 gene displays the lowest homology to c-ets-1 and has thus been classified as the most divergent member of the family as also revealed by some evolutionary data (4). This report also scrutinizes the relationships between *ETS*, *ERG* and *ELG* groups of genes. Several other reports have emphasized the possible relationship between c-ets-1 and c-ets-2 genes on one hand and erg and fli-1 genes on the other hand, which exhibit similar chromosomal location, namely human chromosome 11q23-24 for c-ets-1 and fli-1 and 21q22 for c-ets-2 and erg (5,6). It has also been

* To whom correspondence should be addressed.

0006-291X/93 $4.00

shown that, in addition with the ETS domain, other parts of the proteins are conserved. This is the case for the N-terminal portion of c-ets-1 and c-ets-2 genes as well as erg, fli-1, gabpα, elg and pok genes which bear an amino acid region of weak homology (3, 7, 20, 21). Since the first evolutionary studies (4), numerous new genes belonging to the ets family have been described. To better understand the relationship between all these members, we thus decided to study the evolutionary history of the various ets family genes.

MATERIALS AND METHODS

Sequence source and alignment. All available Drosophila sequences were used even those that are closely related to mammalian ets family genes. The references used for the sequences are the following: HETS1(8), CETS1(9), VETS (2, 36), XETS1A, XETS1B (10), HETS2 (11), CETS2 (12), XETS2A, XETS2B (13), SUETS2 (14), DETS2 (15), ERG (16), FLI-1 (5), SUERG (4), DETS3, DETS4, DETS6 (17), GABPα (18), DELG (7), POK (19) or YAN (20), SAP1 (21), ELK (22), PEA3 (23), ELF1 (24), E74 (25), SPI1 (26, 27), SPIB (28). The ETS domain sequences were used to conduct a computer alignment procedure using the CLUSTAL V package available on the CITI-2/Bisance network.

Construction of phylogenetic trees. Our method was mainly identical to that used by Laudet et al. (29). The percent divergence values for all pairwise comparisons of the aligned sequences were calculated by dividing the number of different residues by the total number of residues. Gaps were treated as mismatches. All values were transformed into distance (d) with Poisson correction d = ln (1-S) where S is the proportion of sites that differ. These values were used to construct phylogenetic trees by the Fitch least square method (30). In parallel, we used the Neigbour-Joining (NJ) method (31) as well as the classical UPGMA method. For clarity the names of the groups of genes have been written in italized capital letters (*ETS*) throughout the paper. The ETS domain has been written in capital letters (ETS) and the various genes in small underlined letters.

RESULTS AND DISCUSSION

Alignment of ets family sequences.

For c-ets-1 and c-ets-2 genes whose homologues in various species have been cloned, we have used the human, chicken, Xenopus, sea urchin and Drosophila sequences. This allowed us to estimate grossly the dates of the different duplications which took place during the ETS domain evolution. We, first, have performed an alignment of the ETS domain (Fig. 1) of all the proteins of the family. At the N-terminal end, the ETS domain starts with a glycine (amino-acid 375 in $p68^{c\text{-}ets\text{-}1}$) corresponding to the beginning of an exon in the c-ets-1 gene. Upstream of this residue (which corresponds also to the beginning of the elk-1 and sap-1 proteins) i.e. in the preceding exon, there is no homology between the various ets family genes except inside each group of genes such as c-ets-1 and c-ets-2 or erg and fli-1 (4). In the C-terminal part, we have limited the studied region at the end of the homology between all the ets family genes (glycine residue 467 in $p68^{c\text{-}ets\text{-}1}$). Our C-terminal limit corresponds to a structural point since it is the end of the published reading frame for the sea urchin c-ets-2 gene. The alignment of Figure 1 shows little gaps except at the beginning of the domain where the alignment is very tedious (see alignments of refs 3, 4, 7, 21). The alignment also shows that in the ETS domain, 17 (i.e. 18 %) amino acids are perfectly conserved between all family members and that 14 other residues are conserved in all but one to four ets family gene products (stars in Fig.1). It is noteworthy that the conserved amino acids are equally distributed in the N-terminal and C-terminal parts of the

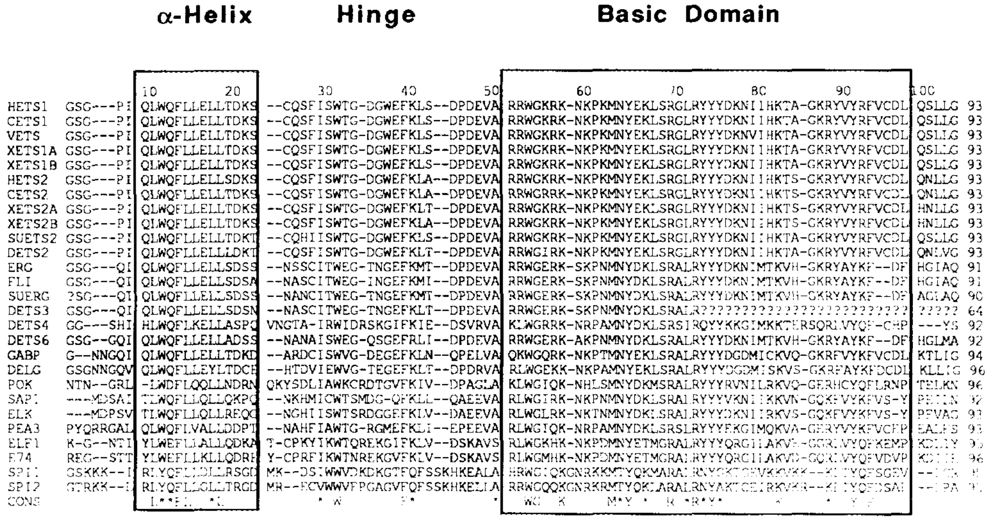

Figure 1. Alignment of the ETS domain of all members of the ets family. The invariant amino acids are noted in the "CONS" line. The residues conserved in all but one to four gene products are marked with a star. The α-helix domain as well as the basic region are boxed. The number of amino acids present in each gene is indicated at the right. ? indicates unknown amino acids.

domain with a central region more divergent. Recently, Wang et al. (32) have suggested that the ETS domain may be divided into two regions: a N-terminal α-helix and a basic C-terminal region. It is interesting to note that the strongly conserved amino acids we point out in Figure 1 cluster in these two regions, the "hinge" connecting them being more divergent. For example there is 57% identity between the most divergent ets family members c-ets-1 and spi-1 in the α-helix region, 35% identity in the basic domain but only 14% in the hinge region. Only two amino acids located in the "hinge" region are common to all ets family members. One of these is a tryptophan residue conserved in ets and myb proteins (3). However, this tryptophan was shown to be dispensable for the DNA binding activity of the c-ets-1 protein (32). These data suggest that the ETS domain is composed of two different conserved regions connected by a more divergent "hinge". Whether these regions correspond to structural sub-domains remains to be addressed by a structural analysis using NMR.

Generation of a phylogenetic tree.

Using this alignment, we calculated distance matrix and Fitch Least Square, Neighbor Joining (NJ) and UPGMA trees as in Laudet et al. (29). We also used the PROTPARS program to calculate the most parsimonious trees relating these sequences. This analysis revealed that the ets family can be divided into 9 groups of genes namely *SPI*, *POK*, *D-ETS-4*, *ELF*, *ELK*, *PEA3*, *ELG*, *ERG* and *ETS* (Fig.2). The various trees constructed with these methods have very similar topologies (data not shown). The major differences between the trees were located at the level of the branching order of the *ERG* and *ELG* groups. It is important to note that for all the trees, the branching order inside the groups was invariable. This strongly argues in favour of the validity of our analysis.

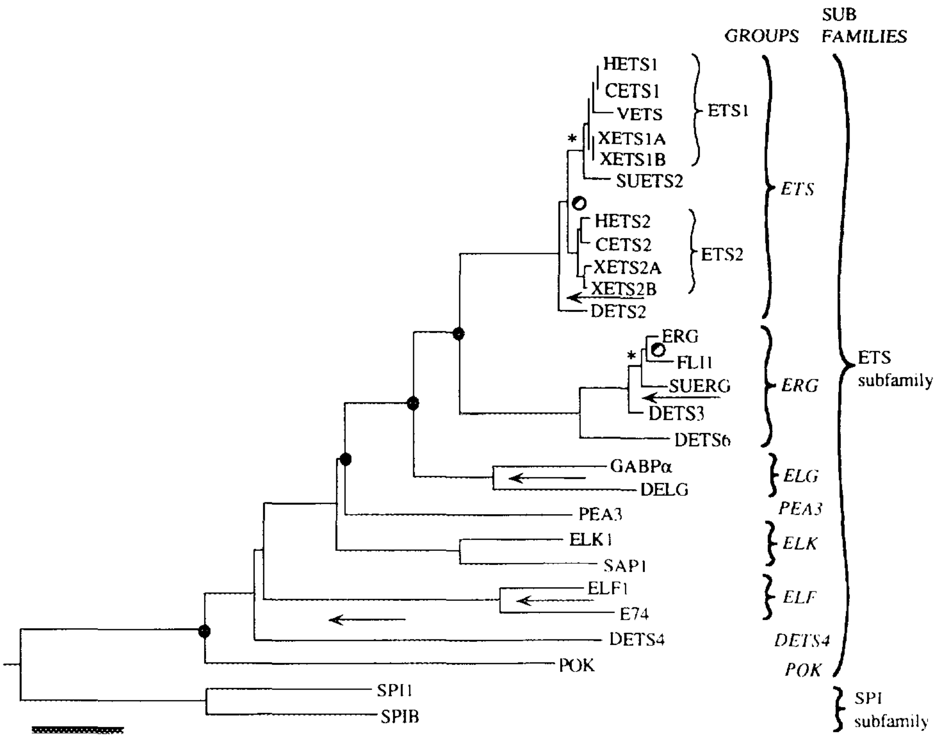

Figure 2. Rooted phylogenetic Fitch tree for all members of the ets family. Drosophila genes are underlined. Arrows point out the mammalian and Drosophila genes which cluster together. The shaded circles in the *ETS* and *ERG* groups show the duplication of the "uncommited" *ETS* and *ERG* ancestors to give c-ets-1 and c-ets-2 and erg and fli-1 genes. The stars indicate the split between sea urchin genes and the other genes. Black points indicate the conflicting branchings between the various programs used.

The trees calculated by the Fitch and N.J. methods are normally unrooted, i.e. the ancestor gene is not determined. We nevertheless observed that when the HMG box motif, which harbours a very low level of homology with the ETS domain (33), was used as an outgroup, the root of the ETS tree was always localized at the level of the *SPI /POK* groups divergence. Thus, the ets family may be divided into two subfamilies: the ets subfamily and the spi subfamily (Fig. 2). This organization is consistent with the relatively weak homology between spi subfamily genes and other ets family genes and with the independent rooting of ets family trees (4). The ets subfamily can be, in turn, divided in eight groups (Fig. 2). This clustering pattern is consistent with other reports (4, 7, 32).

Several Vertebrate ets family genes have a Drosophila homologue: E74 for elf-1, D-elg for gabpα, D-ets-3 for erg and fli-1, D-ets-2 for c-ets-1 and c-ets-2. This suggests that all the divergences giving rise to the various groups of genes took place before the separation of Arthropods and Vertebrates which occurred at least 500 million years ago (34). One of these Drosophila product, E74 , behaves, like the Vertebrate gene products, as a transcriptional regulator. Thus, we can conclude that at least 500 million years ago, the evolution of the ets family was nearly complete with all the groups already defined. The careful examination of the tree at the level of c-ets-1 and c-ets-2 or at the level of the erg and fli-1 genes leads to the conclusion that some duplications took place after the Arthropods / Vertebrates split. Indeed, Figure 2 clearly shows that the Drosophila D-ets-2 gene, primarily described as a Drosophila c-ets-2 homologue is in fact as closely related to c-ets-1 as to c-ets-2. This strongly suggests that the duplication of the ets ancestor to give c-ets-1 and c-ets-2 occurred after the Arthropods/Vertebrates divergence. The same observation could be made for erg, fli-1 and D-ets-3 genes. This is an interesting point since the chromosomal location of *ETS* and *ERG* genes, in addition to the evolutionary tree (5, 6, Fig. 2) allows to suggest that the duplications of *ETS* and *ERG* ancestor genes occured at the same time. It is clear that this divergence had already occurred when the Amphibians lineage diverged from the other Vertebrates since c-ets-1 and c-ets-2 genes have been cloned in *Xenopus laevis*. Thus, during the ca. 100 million years period between the Arthropods / Vertebrates lineages split and the apparition of the amphibians, a duplication of the ancestral c-ets and erg genes occurred to give rise to the four genes we presently know. This period corresponds to the emergence of the Vertebrates lineages and it is tempting to speculate that the ancestral *ETS* and *ERG* genes duplication arose during early Vertebrates evolution(Fig. 3). This conclusion seems to be confirmed by the fact that the sea urchin erg gene appeared to be an "uncommited" erg gene as D-ets-3. Nevertheless, the fact that the sea urchin ets gene primarily described as a c-ets-2 homologue (because it was cloned using a c-ets-2 probe) roots in fact inside the c-ets-1 genes indicates that the situation is more complex. Unfortunately, only the exons coding for the ETS domain are known for the Drosophila and sea urchin *ETS* and *ERG* genes and we cannot compare the other parts of these genes with the other c-ets-1, c-ets-2, erg and fli-1 genes. Since

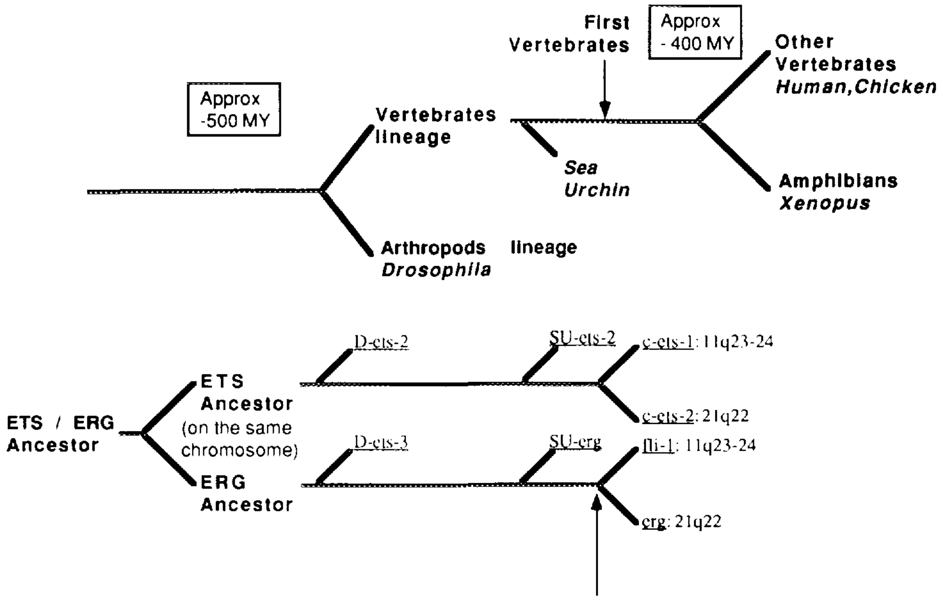

Figure 3. A model of *ETS* and *ERG* gene duplications. From the tree of fig.3 it is obvious that before the Vertebrates / Arthropods split a "uncommited" *ETS/ERG* precursor was duplicated into *ETS* and *ERG* ancestors lying on the same chromosome. Then, Drosophila and sea urchin, respectively,diverged from these ancestors. Finally,at an unknown moment (probably during early Vertebrates evolution),these ancestors duplicate together to give rise to the four genes presently known. Individual genes are written in small and underlined letters and species discussed in this paper are indicated in italic letters.

the homology between c-ets-1 and c-ets-2 or erg and fli-1 is largely significant outside the ETS domain, such a comparison would allow us to strengthen our conclusions. To date and with the molecular data available, it seems logical to conclude that in a first step the *ETS / ERG* ancestor was duplicated to give the *ETS* ancestor and the *ERG* ancestor. The two genes resulting from this duplication occuring more than 500 millions years ago remained closely linked on the same region of the genome. Then all this locus containing the two "uncommitted" *ETS* and *ERG* genes was duplicated after the Arthropods / Vertebrates split, on the lineage which will give rise to the Vertebrates. The result of this last duplication was the two tandem genes we presently know: ets-1 and fli-1 on human chromosome 11 and c-ets-2 and erg on human chromosome 21. Thus, the sea urchin ets and erg genes are "uncommited" precursors *i.e.* represent the "not yet duplicated" ancestor (see Fig. 3). This model should be easy to test since it implies that there is only one *bona fide ETS* and *ERG* genes in Drosophila and sea urchin genomes.

The position of the ETS domain in some genes is interesting to study since in some ets-family members (elk-1, sap-1, elf-1 and pok) this domain is not in the C-terminal part of the protein. Since the *SPI* subfamily genes as well as the majority of the ets sub-family members have a C-terminal ETS domain it is very likely that some independent events arose which have modified the position of the ETS domain in these genes. One possibility is that some exon shuffling occurred to give the presently known members of the *ELK*, *ELF* and *POK* groups.The present situation in the *ELF* group gives a good argument to this model since E74 gene has a C-terminal ETS domain. This suggests that the exon shuffling event in elf-1 occured only in this gene and after the Arthropods / Vertebrates split.

It is interesting to compare the evolutionary history of the ets family to the history of other families of transcription factors. To date two such families have been extensively analyzed on an evolutionary point of view: the Hox gene family (35) and the nuclear receptor superfamily (29). In these two cases, we have found the same theme with some variations: a first "burst" of gene duplications at the origin of the families to give the various classes and a second one later on, which has been positioned at the beginning of Vertebrates evolution for the Hox genes and nuclear receptors. It is interesting to emphasize that our work clearly suggests that the duplication of the *ETS* and *ERG* ancestors took place also during Vertebrates emergence. After these "bursts" all appears as if the function of these genes became "locked" and that the evolution of these families was nearly complete at least on terms of gene number and overall organization. The discovery of ETS genes in early organisms should shed light on the evolutionary history of this family of transcription factors.

ACKNOWLEDGMENTS

We are very grateful to Pascale Crépieux, Jean Coll and Catherine Hänni for constant interest in this project. We also thank Yvan de Launoit, Didier Monté and Jean Marc Vanacker for critical reading of the manuscript. We are grateful to Nicole Devassine and Marîe Christine Bouchez for patient typing.

REFERENCES

1 - Leprince, D., Gegonne, A., Coll, J., de Taisne, C., Schneeberger, A., Lagrou, C. & Stehelin, D. (1983). *Nature (London)*, **306**, 395-397.

2 - Nunn, M.F., Seeburg, P.H., Moscovici, C. & Duesberg, P.H. (1983). *Nature (London)*, **306**, 391-395.
3 - MacLeod, K., Leprince, D. & Stehelin, D. (1992). *Trends Biochem. Sci.*, **17**, 251-256.
4 - Lautenberger, J.A., Burdett, L.A., Gunnell, M.A., Qi, S., Watson, D.K., O'Brien, S.J. & Papas, T.S. (1992). *Oncogene*, **7**, 1713-1719.
5 - Ben-David, Y., Giddens, E.B., Letwin, K. & Bernstein, A. (1991). *Genes Dev.*, **5**, 908-918.
6 - De Taisne, C., Gégonne, A., Stéhelin, D., Bernheim, A. & Berger, R. (1984). *Nature (London)*, **310**, 581-583.
7 - Pribyl, L.J., Watson, D.K., Schulz, R.A. & Papas, T.S. (1991). *Oncogene*, **6**, 1175-1183.
8 - Watson, D.K., McWilliams, M.J., Lapis, P., Lautenberger, J.A., Schweinfest, C.W. & Papas, T.S. (1988). *Proc. Natl. Acad. Sci. USA*, **85**, 7862-7866.
9 - Duterque-Coquillaud, M., Leprince, D, Flourens, A., Henry, C., Ghysdael, J., Debuire, B. & Stéhelin, D. (1988). *Oncogene Res.*, **2**, 335-344.
10 - Stiegler, P., Wolff, C-M., Baltzinger, M., Hirtzlin, J., Senan, F., Meyer, D., Ghysdael, J., Stehelin, D., Befort, N. & Remy, P. (1990). *Nucleic Acids Res.* **18**, 5298.
11 - Watson, D.K., MacWilliams-Smith, M.J., Kozak, C., Reeves, R., Gearhart, J., Nash, W., Modi,W. & Duesberg, P.H. (1986). *Proc. Natl. Acad. Sci. USA*, **83**, 1792-1794.
12 - Boulukos, K.E., Pognonec, P., Begue, A., Galibert, F., Gesquière, J.C., Stehelin, D. & Ghysdael, J. (1988). *EMBO J.*, **7**, 697-705.
13 - Wolff, C-M., Stiegler, P., Baltzinger, M., Meyer, D., Ghysdael, J., Stehelin, D., Befort, N. & Remy, P. (1991). *Cell Growth Diff.*, **2**, 442-456.
14 - Chen, Z-Q, Kan, N.C., Pribyl, L., Lautenberger, J.A., Moudrianakis, E. & Papas, T.S. (1988). *Dev. Biol.*, **125**, 432-440.
15 - Pribyl, L.J., Watson, D.K., McWilliams, M.J., Ascione, R. & Papas, T.S. (1988). *Dev. Biol.*, **127**, 45-53.
16 - Rao, V.N., Papas, T.K. & Reddy, E.S.P. (1987). *Science*, **237**, 635-639.
17 - Chen, T., Bunting, M., Karim, F.D. & Thummel, C.S. (1992). *Dev. Biol.*, **151**, 176-191.
18 - LaMarco, K., Thompson, C.C., Byers, B.P., Walton, E.M. & McKnight, S.L. (1991). *Science*, **253**, 789-792.
19 - Tei, H., Nihonmatsu, I., Yokokura, T., Ueda, R., Sano, Y., Okuda, T., Sato, K., Hirata, K., Fujita, S.C. & Yamamoto, D. (1992). *Proc. Natl. Acad. Sci. USA*, **89**, 6856-6860.
20 - Lai, Z.C. & Rubin, G.M. (1992). *Cell*, **70**, 609-620.
21 - Dalton, S. & Treisman, R. (1992). *Cell*, **68**, 597-612.
22 - Rao, V.N., Huebner, K., Isobe, M., Ar-Rushdi, A., Croce, C.M. & Reddy, E.S. (1989). *Science*, **244**, 66-70.
23 - Xin, J.H., Cowie, A., Lachance, P. & Hassell, A. (1992). *Genes Dev.*, **6**, 481-436.
24 - Thompson, C.B., Wang, C.Y., Ho, I.C., Bohjanen, P.R., Petryniak, B., June, C.H., Miesfeldt, S., Zhang, L., Nabel, G.J., Karpinsky, B. & Leiden, J.M. (1992). *Mol. Cell. Biol.*, **12**, 1043-1053.
25 - Burtis, K.C., Thummel, C.S., Jones, C.W., Karim, F.D. & Hogness, D.S. (1990). *Cell*, **61**, 85-99.
26 - Ray, D., Culine, S., Tavitian, A., C. Moreau-Gachelin, F. (1990). *Oncogene*, **5**, 663-668.
27 - Goebl, M.G., Moreau-Gachelin, F., Ray, D., Tambourin, P. and Tavitian, A. (1990). *Cell*, **61**, 1165-1166.
28 - Ray, D., Bosselut, R., Ghysdael, J., Mattei, M.G., Tavitian, A. & Moreau-Gachelin, F. (1992). *Mol. Cell. Biol.*, **12**, 4297-4304.
29 - Laudet, V., Hänni, C., Coll, J., Catzeflis, F. & Stéhelin, D. (1992). *EMBO J.*, **11**, 1003-1013.
30 - Fitch, W.M. (1981). *J. Mol. Evol.*, **18**, 30-37.
31 - Saitou, N. & Nei, M. (1987). *Mol. Biol. Evol.*, **4**, 406-425.
32 - Wang, C-Y., Petryniak, B., Ho, I-C., Thompson, C.B. & Leiden, J.M. (1992). *J. Exp. Med.*, **175**, 1391-1399.
33 - Waterman, M.L., Fischer, W.H. & Jones, K.A. (1991). *Genes Dev.*, **5**, 656-669.
34 - Hartland, W.B., Cow, A.V., Llewellyn, P.G., Pickton, C.A.G., Smith, A.G. & Walters, R. (1982). *A Geologic Time Scale*. Cambridge University Press, Cambridge.
35 - Murtha, M.T., Leckman, J.F. & Ruddle, F.H. (1991). *Proc. Natl. Acad. Sci. USA*, **88**, 10711-10715.
36 - Golay,J., Introna,M. & Graf,T. (1988). *Cell*, **55**, 1142-1158.