ORIGINAL ARTICLE

Nucleolins from different model organisms have conserved sequences reflecting the conservation of key cellular functions through evolution

Fernando González-Camacho and Francisco Javier Medina

Centro de Investigaciones Biológicas (CSIC), Madrid, Spain

Summary

Sequences available in public protein databases belonging to nucleolin or nucleolin-like proteins have been aligned using public domain software, in order to obtain relevant data regarding the degree of their conservation, which could be a reflection of the degree of conservation of the functions currently attributed to this protein. Nucleolin is known to be a nucleolar multifunctional protein, involved in different steps of pre-rRNA transcription and processing. Three domains are constantly present in all nucleolins, namely a series of acidic/serin (Ac/Ser) sequences, a number of RNA recognition motifs (RRM) and a region rich in glycin and arginin (GAR). The number of motifs present in each one of the three domains is variable. Furthermore, we have characterized in all nucleolins the presence of a bipartite consensus nuclear localization sequence (NLS). The only cases in which this sequence with a definite structure was not totally evident were in the yeast *S. pombe* (a possible monopartite structure) and in the protozoan *T. thermophyla,* in which it appears to be absent. Finally, we have constructed the phylogenetic tree of the 15 species investigated, taking exclusively the data regarding this protein. Interestingly, the tree obtained closely resembles the organization of these taxonomic groups throughout evolution, as it is presently known. We conclude that nucleolin is a highly conserved protein, whose gene was already present in an ancestor eukaryotic species, at an early stage of the evolutionary process, from which it has evolved very slowly. This is a reflection of the fundamental functions carried out by this protein, which were already fixed in the ancestor species.

Keywords: nucleolar protein – nuclear localization sequence – RNA recognition motifs – glycin-arginin-rich – phylogenetic study – evolution

INTRODUCTION

Nucleolin is a nucleolar multifunctional protein which has been described in a large number of eukaryotic organisms. Most of the functions attributed to nucleolin are closely related to ribosome biogenesis, in which it participates in ribosomal gene transcription and pre-rRNA processing. Furthermore, nucleolin is the major nucleolar protein of actively proliferating cells (for reviews, see Olson 1991, Tuteja and Tuteja 1998, Ginisty et al. 1999, González-Camacho and Medina 2004). However, nucleolin has also been involved in the splicing process leading to the formation of

mRNA, as a constituent of the spliceosome particle (Rappsilber et al. 2002), as well as in the shuttling process between nucleolus and cytoplasm (Borer et al. 1989). Furthermore, other than its major nucleolar localization, nucleolin has been found in the nucleoplasm, and also, in a minor proportion, in the cytoplasm (Martín et al. 1992, Medina et al. 2001).

The first studies on nucleolin were made in the seventies, when it was described as the C23 protein in Novikoff hepatoma cells (Orrick et al. 1973, Prestayko et al. 1974). Then, in the late eighties, nucleolin homologues were found and sequenced in various mammalian species, namely hamster (Lapeyre et al. 1987), mouse (Bourbon et al. 1988), human (Srivastava et al. 1990) and rat (Bourbon and Amalric 1990). Studies in other organisms followed these early characterizations (Table 1).

Nucleolin multifunctionality comes from its structural organization. The analysis of the amino acid sequence reveals the presence of three structural domains, whose functional significance has been shown in a number of studies. The N-terminal domain is characterized by the alternation of acidic and basic regions, and it contains the targets of phosphorylation by cdc2 kinase and casein kinase II (CKII) (Caizergues-Ferrer et al. 1987, Peter et al. 1990, Belenguer et al. 1990, De Cárcer et al. 1997). The central domain contains, in mammals, four RNA recognition motifs (RRMs). Between two of these domains is found the bipartite nuclear localization sequence (NLS). Finally, the C-terminal domain consists of a sequence rich in glycins and arginins (GAR). The RRMs specifically interact with the external transcribed spacer (ETS) region in the primary pre-rRNA transcript, whereas the GAR domain is involved in the efficiency of nucleolin binding to RNA, but not in the specificity (Ghisolfi et al. 1992). For detailed and extensive revisions of the structure and function of nucleolin, see Tuteja and Tuteja (1998) and Ginisty et al. (1999).

Several nucleolar proteins, in different eukaryotic organisms, have shown a similar tripartite structural organization and, as far as we know, they are involved in the same functional events in ribosome biogenesis as described for nucleolin. This set of nucleolar multifunctional proteins has received the name: "nucleolin-like proteins".

The sequences of proteins described as nucleolin or nucleolin-like are available in the protein databases Swiss-Prot and TrEMBL (http://www.ebi.ac.uk/swissprot/access.html). Now, the number of complete sequences is 15, coming from a variety of organisms. All of them are eukaryotic, belonging to a wide range of taxonomic groups, namely Chordates *(Homo sapiens, Mesocricetus auratus, Mus musculus, Rattus norvegicus, Gallus gallus, Xenopus laevis, Cyprinus carpio),* Prochordates *(Ciona intestinalis),* Yeast *(Saccharomyces cerevisiae, Schizosaccharomyces pombe),* Plants *(Arabidopsis thaliana, Medicago sativa, Pisum sativum, Nicotiana tabacum),* and Protozoa *(Tetrahymena thermophila).*

The availability of these sequences and the wide variety of organisms represented constitute a good starting point in the search for the identity of this group of proteins and the actual homologies between the different members.

NUCLEOLINS FROM CHORDATES, PROCHORDATES AND PROTOZOA

As indicated above, the earliest studies performed on nucleolin were carried out on mammals (hamster, mouse, rat and human). Taking the amino acid sequences of these nucleolins and aligning them by using the software "Clustal-W" (Thompson et al. 1994) (http://www.ebi.ac.uk/clustalw) results in: 77.1% of total sequence identity, 10.62% of high identity, 4.97% of weak identity and 7% of differences (Fig. 1). The alignment clearly shows that the four proteins present the modular structure in three structural and functional domains, as has been described earlier for CHO nucleolin (Lapeyre et al. 1987). Therefore, the four proteins correspond to a common structural and functional model.

The structural organization of human nucleolin in functional domains is well known, as it is in other organisms. Nevertheless, the nucleolar localization sequence (NLS) is only evident in the sequence of human nucleolin (Srivastava et al. 1990), while it is not so clear in other species. This sequence is formed by three components: first, a motif of four constant amino acids (three lysins and one arginin), second, a spacer tract constituted by 11 amino acids, and third, a second constant motif of four amino acids (three lysins and one glutamin). The multiple alignment that we have obtained allowed us to localize the sequence described by Srivastava and co-workers and to compare it with the rest of the aligned sequences. The recognized sequences were shown to be identical to the human in relation to the bipartite signal, whereas the spacer sequence showed slight variations (Fig. 2).

In general, the four mammalian nucleolins follow exactly the same scheme and share a very high proportion of their amino acid sequences; in fact, the identity rate is 93%. However, when comparing mammalian nucleolins with the nucleolin-like proteins characterized in other vertebrates, the identity is less apparent, although

their organization in three domains appears evident as a common conserved feature.

The analysis based on multiple alignment and the similarities found in known domains showed that the GAR domain, the RRMs and the NLS practically do not vary as to their situation within the primary structure of the protein among vertebrates. The most variable motif is the amino-terminal domain. While chicken keeps the same organization as in mammals, carp and *Xenopus* contain important variations. In carp nucleolin, six acidic/serin (Ac/Ser) motifs were detected: four of them were identical to those of other vertebrates and the two new ones intercalate between the first and the second corresponding to the mammalian model.
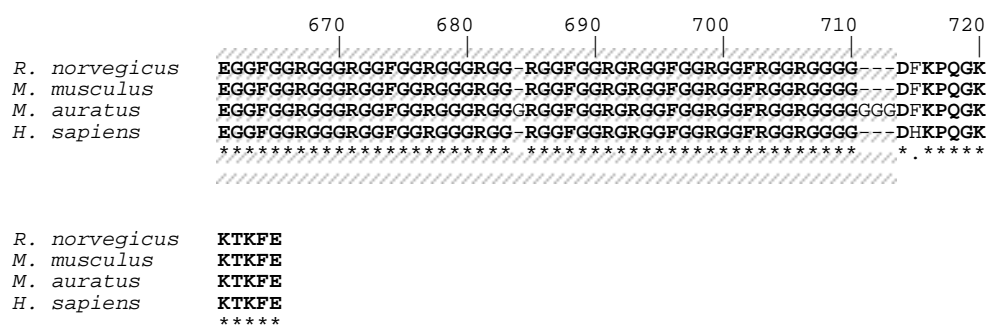
Table 1. **Available data on nucleolin and nucleolin-like proteins in different organisms.** In each case, the table shows the Access Number to either the Swiss-Prot or TrEMBL protein database, the bibliographic reference of the first sequence description, the particular name given, the number of amino acids and the molecular mass expressed in kilodaltons. The URL addresses of data bases are: http://www.ebi.ac.uk/swissprot/access.html, http://us.expasy.org/sprot/, http://srs.embl-heidelberg.de:8000/

Accession No.	Reference	Organism	Name	No. of amino acids	Molecular mass (kDa)
P19338	(Srivastava et al. 1990)	*Homo sapiens* (Human)	Nucleolin (C23 protein)	706	76,2
P08199	(Lapeyre et al. 1987)	*Mesocricetus auratus* (Hamster)	Nucleolin	713	76,9
P09405	(Bourbon et al. 1988)	*Mus musculus* (Mouse)	Nucleolin	706	76,6
P13383	(Bourbon and Amalric 1990)	*Rattus norvegicus* (Rat)	Nucleolin	712	77
P15771	(Maridor et al. 1990)	*Gallus gallus* (Chicken)	Nucleolin	694	75,6
Q804J2	(Alvarez et al. 2003)	*Cyprinus carpio* (Carp)	Nucleolin	693	73,8
P20397	(Caizergues-Ferrer et al. 1989, Rankin et al. 1993)	*Xenopus laevis*	Nucleolin	650	70
Q9U8P6	(Tanaka et al. 2004)	*Ciona intestinalis* (Ascidian)	Nucleolin-like	529	58
P27476	(Lee et al. 1991)	*Saccharomyces cerevisiae*	Nsr1	414	44,5
P41891	(Gulli et al. 1995)	*Schizosaccharomyces pombe*	Gar2	500	52,9
Q9LIH8	(Kaneko et al. 2000)	*Arabidopsis thaliana*	Nucleolin-like	610	66,2
Q40363	(Bögre et al. 1996)	*Medicago sativa* (Alfalfa)	NucMs1	635	67,1
Q41042	(Tong et al. 1997)	*Pisum sativum* (Pea)	Nucleolin-like	611	64,7
Q8LNZ4	(Maeshima and Matsuda 2002)	*Nicotiana tabacum* (Tobacco)	Nucleolin	620	65,1
Q27199	(McGrath et al. 1997)	*Tetrahymena thermophila* (Protozoan)	Nucleolar phosphoprotein (Nopp52)	476	51,7

a

```
                   10        20        30        40        50        60
                   |         |         |         |         |         |
R. norvegicus  VKLAKAGKTHGESKKMAPPPKEVEEDSEDEEMSEDE-DDSSGEEEVVIPQKKGKKATTTP
M. musculus    VKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDE-DDSSGEEEVVIPQKKGKKATTTP
M. auratus     VKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEEE-DDSSGEE-VVIPQKKGKKATATP
H. sapiens     VKLAKAGKNQGDPKKMAPPPKEVEEDSEDEEMSEDEEDDSSGEE-VVIPQKKGKKAAATS
               ********:.:*:.*********************:* ****** ***********:.:*.

                   70        80        90       100       110       120
                   |         |         |         |         |         |
R. norvegicus  AKKVVVSQTKKAAVPTPAKKAAVTPGKKAAATPAKKAVTPAKVVPTPGKKGAAQAKALVP
M. musculus    AKKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVP
M. auratus     AKKVVVSQTKKVAVPTPAKKAAVTPGKKAAATPAKKAVTPAKAVATPGKKGATQAKALVA
H. sapiens     AKKVVVSPTKKVAVATPAKKAAVTPGKKAAATPAKKTVTPAKAVTTPGKKGATPGKALVA
               *******.***.**..**************.****** ***** .:.*******:.****.

                  130       140       150       160       170       180
                   |         |         |         |         |         |
R. norvegicus  TPGKKGAVTPAKGAKNGKNAKKEDSDEDEDEED-EDDSDEDED--EEDEFEPPVVKGVKP
M. musculus    TPGKKGAATPAKGAKNGKNAKKEDSDEDEDEED-EDDSDEDEDDEEEDEFEPPIVKGVKP
M. auratus     TPGKKGAVTPAKGAKNGKNAKKEDSDEDEDDDDDEDDSDEDEEDEEEDEFEPPVVKG-KQ
H. sapiens     TPGKKGAAIPAKGAKNGKNAKKEDSDEEEDDDSEEDEEDDEDEDEDEDEIEPAAMKA---
               *******.*****************.***:.:  ::**::.:  :.***:** . :*.

                  190       200       210       220       230       240
                   |         |         |         |         |         |
R. norvegicus  AKAAPAAPASEDEDEEDDDDEDDDDDEEEEEEDDSEEEVMEITPAKGKKTPAKVVPVKA
M. musculus    AKAAPAAPASEDE--EDDEDDDEEDDEE-DDSEEVMEITTAKGKKTPAKVVPMKA
M. auratus     GKVAAAAPASEDE--DEEEDEEEEEEDEEEED-DSEEEEAMEITPAKGKKAPAKVVPVKA
H. sapiens     ---AAAAPASEDE--DDEDDEDDEDDDDDEED--DSEEEAMETTPAKGKK-AAKVVPVKA
               *.********  :::::**::::*::**:  ..***.** *.***** .*****:**

                  250       260       270       280       290       300
                   |         |         |         |         |         |
R. norvegicus  KSVAEEEDDEDDEDEEEDEDEEDEEDD----EDEDEEEEEEEPVKAAPGKRKKEMTKQKE
M. musculus    KSVAEEEDDEEEDEDDE-DEDDEEEDDE----DDDEEEEEEEEPVKAAPGKRKKEMTKQKE
M. auratus     KNVAEEDDDEEEDEDE-EEDEEEEEDE----EEEEEEEEEEPVKPAPGKRKKEMTKQKE
H. sapiens     KNVAEDEEEDDDEDDDDDEDDEDDDEDDEEDDEE----EEPVKEAPGKRKKEMAKQKA
               *.***:::::*::::: ::*:::*:*:   :::::********* *********:.***

                  310       320       330       340       350       360
                   |         |         |         |         |         |
R. norvegicus  APEAKKQKIEGSEPTTPFNLFIGNLNPNKSVAELKVAISELFAKNDLAAVDVRTGTNRKF
M. musculus    APEAKKQKVEGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVDVRTGTNRKF
M. auratus     VPEAKKQKVEGSESTTPFNLFIGNLNPNKSVAELKVAISEPFAKNDLAVVDVRTGTNRKF
H. sapiens     APEAKKQKVEGTEPTTAFNLFVGNLNFNKSAPELKTGISDVFAKNDLAVVDVRIGMTRKF
               .*******:**:* ** .****:****:.**.. .***  .:*.:*.*****.****.*. :*

                  370       380       390       400       410       420
                   |         |         |         |         |         |
R. norvegicus  GYVDFESAEDLEKALELTGLKVFGNEIKLEKPKGRDSKKVRAARTLLAKNLSFNITEDEL
M. musculus    GYVDFESAEDLEKALELTGLKVFGNEIKLEKPKGRDSKKVRAARTLLAKNLSFNITEDEL
M. auratus     GYVDFESAEDLEKALELTGLKVFGNEIKLEKPKGRDSKKVRAARTLLAKNLSFNITEDEL
H. sapiens     GYVDFESAEDLEKALELTGLKVFGNEIKLEKPKGKDSKKERDARTLLAKNLPYKVTQDEL
               **********************************:**** * ********::*:*:::**

                  430       440       450       460       470       480
                   |         |         |         |         |         |
R. norvegicus  KEVFEDAVEIRLVSQDGRSKGIAYIEFKSEADAEKNLEEKQGAEIDGRSVSLYYTGEKGQ
M. musculus    KEVFEDAMEIRLVSQDGKSKGIAYIEFKSEADAEKNLEEKQGAEIDGRSVSLYYTGEKGQ
M. auratus     KEVFEDALEIRLVSQDGKSKGIAYIEFKSEADAEKNLEEKQGAEIDGRSVSLYYTGEKGQ
H. sapiens     KEVFEDAAEIRLVSKDGKSKGIAYIEFKTEADAEKTFEEKQGTEIDGRSISLYYTGEKGQ
               *******:*****:**:***********:******* ****:******:*:**********

                  490       500       510       520       530       540
                   |         |         |         |         |         |
R. norvegicus  RQE-RTGKNSTWSGESKTLVLSNLSYSATEETLQEVFEKATFIKVPQNPHGKSKGYAFIE
M. musculus    RQE-RTGKTSTWSGESKTLVLSNLSYSATKETLEEVFEKATFIKVPQNPHGKPKGYAFIE
M. auratus     RQE-RTGKNSTWSGESKTLVLSNLSYSATEETLQEVFEKATFIKVPQNQQGKSKGYAFIE
H. sapiens     NQDYRGGKNSTWSGESKTLVLSNLSYSATEETLQEVFEKATFIKVPQNQNGKSKGYAFIE
               .*: * **.*****************:***.***:.*:*************.*.**.*******

                  550       560       570       580       590       600
                   |         |         |         |         |         |
R. norvegicus  FASFEDAKEALNSCNKMEIEGRTIRLELQGPRGSPNARSQPSKTLFVKGLSEDTTEETLK
M. musculus    FASFEDAKEALNSCNKMEIEGRTIRLELQGS----NSRSQPSKTLFVKGLSEDTTEETLK
M. auratus     FASFEDAKEALNSCNKMEIEGRTIRLELQGPRGSPNARSQPSKTLFVKGLSEDTTEETLK
H. sapiens     FASFEDAKEALNSCNKREIEGRAIRLELQGPRGSPNARSQPSKTLFVKGLSEDTTEETLK
               ****************.***** *******..    *.**********************

                  610       620       630       640       650       660
                   |         |         |         |         |         |
R. norvegicus  ESFEGSVRARIVTDRETGSSKGFGFVDFNSEEDAKAAKEAMEDGEIDGNKVTLDWAKPKG
M. musculus    ESFEGSVRARIVTDRETGSSKGFGFVDFNSEEDAKAAKEAMEDGEIDGNKVTLDWAKPKG
M. auratus     ESFEGSVRARIVTDRETGSSKGFGFVDFNSEEDAKAAKEAMEDGEIDGNKVTLDWAKPKG
H. sapiens     ESFDGSVRARIVTDRETGSSKGFGFVDFNSEED---AKEAMEDGEIDGNKVTLDWAKPKG
               ***:.*****************************   *********************
```

```
                    670         680         690         700         710         720
                     |           |           |           |           |           |
R. norvegicus   EGGFGGRGGGRGGFGGRGGGRGG-RGGFGGRGRGGFGGRGGFRGGRGGGG---DFKPQGK
M. musculus     EGGFGGRGGGRGGFGGRGGGRGG-RGGFGGRGRGGFGGRGGFRGGRGGGG---DFKPQGK
M. auratus      EGGFGGRGGGRGGFGGRGGGRGGGRGGFGGRGRGGFGGRGGFRGGRGGGGGGGDFKPQGK
H. sapiens      EGGFGGRGGGRGGFGGRGGGRGG-RGGFGGRGRGGFGGRGGFRGGRGGGG---DHKPQGK
                ************************ *************************  * .*****


R. norvegicus   KTKFE
M. musculus     KTKFE
M. auratus      KTKFE
H. sapiens      KTKFE
                *****
```

b

Fig. 1 (*a*) **Representation of the alignment of amino acid sequences of nucleolin in some mammalian species:** rat *(Rattus norvegicus),* mouse *(Mus musculus),* hamster *(Mesocricetus auratus)* and human *(Homo sapiens).* Asterisks represent identity between amino acids (also marked in bold), double dots indicate conserved amino acids, and a single dot means semiconserved amino acids. The different domains are marked by boxes: lighter grey corresponds to the acidic sequences (Ac/Ser), darker grey, to the bipartite nuclear localization sequence (NLS), dotted pattern, to RNA recognition motifs (RRMs), and dashed pattern, to the glycin-arginin rich (GAR) domain. The alignment was performed with the software Clustal-W (Thompson et al. 1994), which is publicly available in the server NPS@ (Network protein Sequence @nalysis) (Combet et al. 2000), whose URL address is http://npsa-pbil.ibcp.fr. This server belongs to the Pôle Bio-Informatique Lyonnais, from the University of Lyon (France)

(*b*) **Schematic representation of the domains present in the amino acid sequence of mammalian nucleolin.** Patterns are the same as indicated above. The calibrated bar represents the number of amino acids

In the case of *Xenopus*, there is only one of these sequences added to the mammalian model, also located between the first and the second Ac/Ser motifs of this reference model. This specific sequence is aligned with one of the two sequences added in fishes (*C. carpio*) to the mammalian model, in particular with the second sequence (Fig. 3).

Regarding prochordates, the structure and the functional characterization of nucleolin of Ascidians was described by Tanaka et al. (2004), after cloning and sequencing the gene CiRGG1. They propose the existence of a third RRM (in amino acids 163–231), and do not mention even the possibility of existence of NLS. However, in the alignment performed by us, we have found some matches in tracts which could correspond to NLS in all analyzed vertebrates, including *Xenopus* and *C. carpio* (Fig. 4).

The NLS described for chicken nucleolin (Schmidt-Zachmann and Nigg, 1993) matches with that of the human nucleolin and nucleolins from other mammals. This alignment allows us to predict these sequences for other vertebrates, such as *Xenopus* and carp. In both cases it is a bipartite structure in which the first four amino acids in the amphibian are identical to those of the rest of vertebrates, the spacer region contains two amino acids less than in mammals and, from the last four amino acids, the mammalian glutamine is changed by a threonine. In the carp, the total number of amino acids in the structure is kept, although the composition is also slightly changed: from the first four, the fourth changes a lysin by an alanin, and, from the last four, glutamine is changed by alanin (Fig. 4).

```
R. norvegicus              282 KRKKEMTKQKEAPEAKKQK 300
M. musculus                280 KRKKEMTKQKEAPEAKKQK 298
M. auratus                 279 KRKKEMTKQKEVPEAKKQK 297
H. sapiens (Srivastava et al. 1990)  278 KRKKEMAKQKAAPEAKKQK 296

                               ******:***  .*******
```

Fig. 2. Sequences of the bipartite structure of nuclear localization (NLS) present in nucleolins from four mammalian species. Numbers indicate the position of the first and the last amino acid in each sequence. Asterisks show amino acids with identity, double dot, conserved amino acids and the single dot semiconserved amino acids. Underlined amino acids constitute the bipartite structure and non-underlined amino acids form the spacer sequence

If we follow the same strategy for the ascidian and the protozoan, we observe deeper changes. The protozoan nucleolin does not contain either GAR domain or NLS, and the central domain is formed by only two RRMs. The amino-terminal domain includes the same four Ac/Ser regions as the mammalian nucleolin, plus three additional ones. The presence of seven Ac/Ser regions and two RRMs was described by McGrath et al. (1997) in a paper in which the existence of a GAR domain is also postulated. Certainly, this putative GAR sequence would be formed by only four glycins, one prolin and one arginin (amino acids 449–454), which does not appear to constitute a true domain. Otherwise, the ascidian contains two GAR regions, one of them located between two of the RRMs which form the central domain. The NLS is located between the amino-terminal and the central domains, and exhibits the bipartite structure common to all cases analyzed up to now (Fig. 3). From the first four amino acids of NLS, a lysin is changed by an alanin in the ascidian nucleolin with respect to the vertebrate model; the same as in the fish. Regarding the second element of the bipartite structure, the third amino acid is changed, appearing glutamin instead of alanin (Fig. 4). Finally, the amino-terminal domain is formed by two Ac/Ser regions, from which the first is aligned with the second of the vertebrates, and the second contains the sum of the third and fourth regions of vertebrates (Fig. 3).
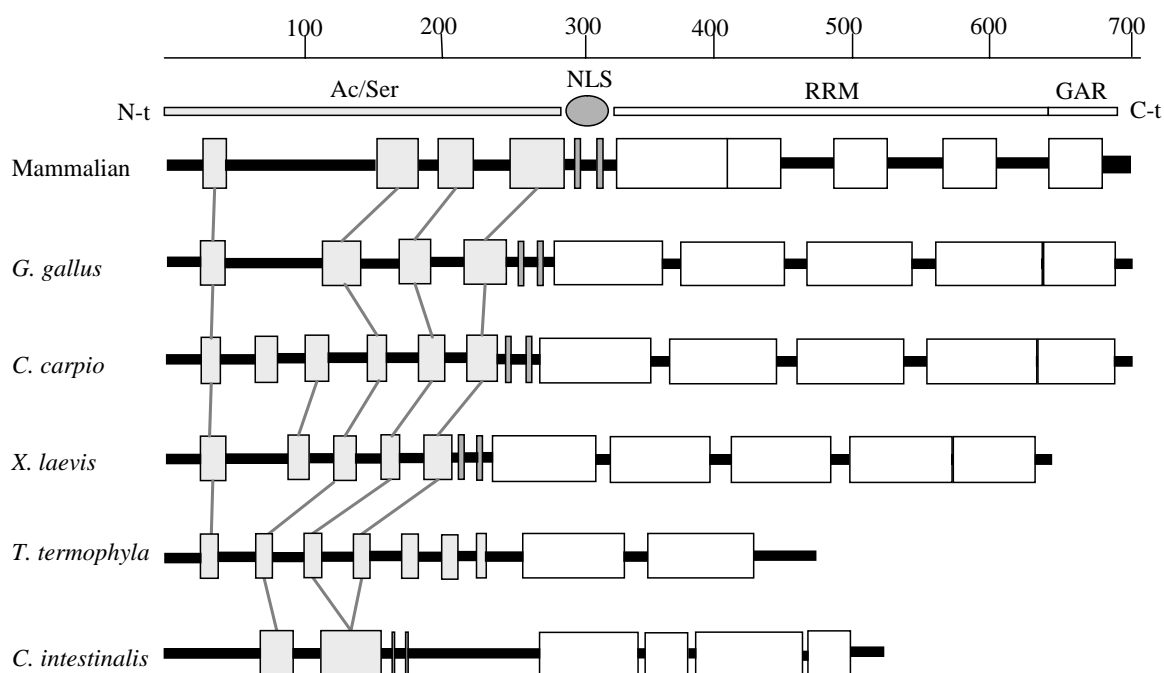
Fig. 3. Schematic representation of the structural domains of nucleolins from Chordata, Prochordata and Protozoa. Correspondence between patterns and domains is the same as in Fig. 1. Vertical lines connect the aligned Ac/Ser sequences. N-t: amino-terminal end. C-t: carboxi-terminal end. The horizontal bar expresses the distance in number of amino acids

```
C. carpio                                    245  KRKAEA K KE KG T PPAKKAK 263

X. laevis                                    205  KRKKEM P KT I- - PEAKKTK 221

G. gallus (Schmidt-Zachmann and Nigg 1993)  253  KRKKEM A N- KS A PEAKKKK 270
                                                  ****        * *** *


C. intestinalis                             166  KRKATETEEVAT KKVK 181
```

Fig. 4. Sequences of the bipartite structure of nuclear localization belonging to three Chordates and one Prochordate. Numbers indicate the position of the first and the last amino acid in each sequence. Asterisks mean absolute identity. Underlined amino acids are those forming the bipartite structure and non-underlined correspond to the spacer sequence

```
P. sativum (Tong et al. 1997)     26  KKGKRQAEEEIKKVS-AKKQK 45

M. sativa                         26  KKGKRQAEEEVKAVS-AKKQK 45

N. tabacum                        24  KKGKREAGEEIEKIMSAKKQK 44

A. thaliana                        6  KKGKRDAEEDLDMQV-TKKQK 26
                                      *****:* *::.    :****
```

Fig. 5. Sequences of the bipartite structure of nuclear localization belonging to four plant species. The meaning of numbers and signs is the same as in Figs. 1, 2 and 4

NUCLEOLIN-LIKE PROTEINS IN PLANTS

In a way similar to the previous cases, we have analyzed the structure of nucleolin-like proteins in plants. For this purpose, based on the domain prediction, we have designed a scheme for each protein, in which each domain has been identified, and on which the comparison of data has been performed. Two basic strategies were utilized for prediction: the first was founded on the existing identity among the various proteins, and the second, on the localization of domain sequences from known domains described in other proteins and available in public databases (Prosite, http://www.expasy.org/ prosite; PDB, http://www.rcsb.org/pdb).

Therefore, we carried out the sequence alignment of known proteins. The result between pea, alfalfa and tobacco was an absolute identity of 39%, 29% of similarity and 32% of differences. When *Arabidopsis* was introduced in the alignment, differences increased up to 47.7%, whereas, if only pea and alfalfa (both belonging to the same family) are aligned, the differences are reduced to only 17%.

The present knowledge of nucleolin NLS in plants is restricted to the hypothesis postulated by Tong et al. (1997) in peas. When they studied the amino acid sequence, they observed a short sequence, similar to the prototype signal of SV40 T-antigen, which was postulated as "potential NLS". If we study in detail this sequence (KKGK), we will find that it also appears in the amino acid sequences of alfalfa, tobacco and *Arabidopsis*, and in all cases they are perfectly aligned. Therefore, the proposal of this sequence as NLS in peas and, by analogy, in the other three plant model species, could be correct. Furthermore, if we extend the sequence analysis, we observe that, after 12–13 amino acids from the first sequence, there is another short sequence, KKQK, which is located and aligned in the same position for the four species. This fact, together with the identification of this sequence as the second part of the bipartite NLS in mammals, strongly suggests that plant NLS is formed by a bipartite structure. The two sequences forming this bipartite structure are perfectly conserved among plants, whereas the spacer sequence is highly variable (Fig. 5).

The localization of NLS in plants is different from that in animals. This sequence is not located between the amino-terminal and the carboxi-terminal domains, but it lies at the beginning of the protein sequence, near the amino-terminal end, in a position prior to the repeated Ac/Ser sequences. Regarding the GAR domain, it is located in the carboxi-terminal end, like in vertebrates, and is rather conserved in plants, except in the case of *Arabidopsis*, in which it appears highly modified. The central domain is formed by two RRMs in all cases, unlike the four RRMs found in animals. The most variable domain among plants, as in other

models, is the amino-terminal region, in which the number of Ac/Ser units is always higher in plants than in animals, ranging from seven in peas (Tong et al. 1997) to nine in alfalfa (Bögre et al. 1996) and *Arabidopsis* (Fig. 6).
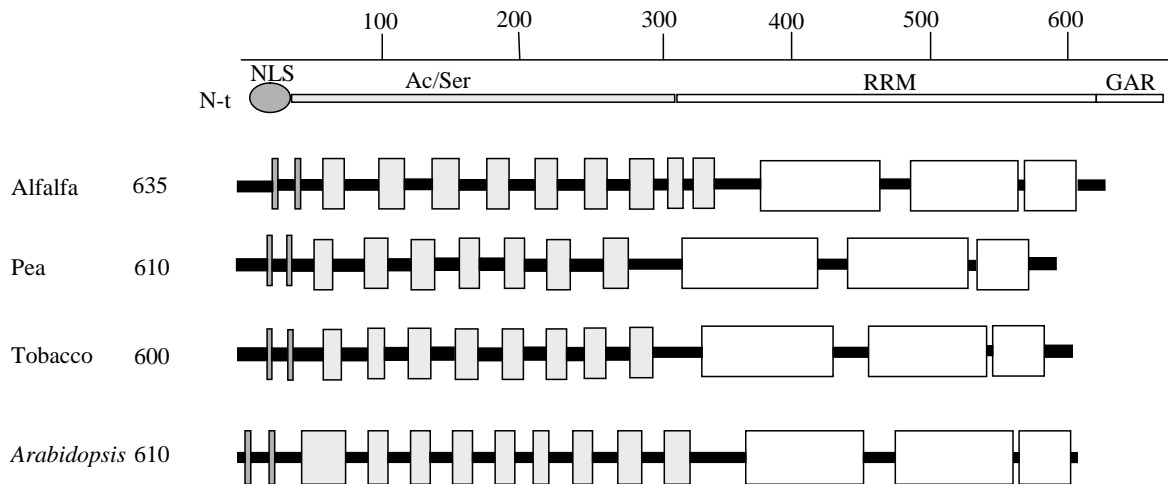

Fig. 6. **Schematic representation of the structural domains of nucleolin-like proteins from plants.** Patterns and abbreviations are the same as in Figs. 1 and 3. The number following the name of the plant indicates the number of amino acids of the nucleolin-like protein in this particular species

NUCLEOLIN-LIKE PROTEINS IN YEAST

Alignment of known yeast nucleolin-like proteins resulted in a total identity of 44% and differences of 32.3%. The nucleolin-like protein of *S. cerevisiae* was called NSR1 (Lee et al. 1991), and contains three main domains corresponding to the general model. This protein resembles the plant scheme in that it contains only two RRMs in the central domain. With regard to NLS, it was proposed that the sequence KKRKS could correspond to this function, in a similar way as with peas (see above). Also in this case, after a spacer sequence of 13 amino acids, the sequence KKQK appears, exactly as in vertebrates and plants, which indicates that yeast nucleolin-like proteins also contain bipartite NLS (Fig. 7).

The primary structure of the protein gar2, the nucleolin-like protein of *S. pombe*, has the typical nucleolin organization (Gulli et al. 1995), but these authors postulate an additional basic domain in the amino-terminal end, which is also proposed by them for NSR1. Otherwise, no clear data are provided for NLS. According to Léger-Silvestre et al. (1997), a bipartite signal is located between the last Ac/Ser sequence and the first RRM. A fusion protein which was constructed containing only the amino-terminal region of gar2, lacking RRMs, is considered sufficient for nucleolar localization (Léger-Silvestre et al. 1997). Nevertheless, the alignment of NSR1 *(S. cerevisiae)* with gar2 *(S. pombe)* shows that the first four amino acids of the NLS of NSR1 are identical to those of gar2, but the spacer sequence and the second part of the bipartite structure are not present in gar2, in which the sequence is highly modified and there is not any putative NLS beyond this point (Fig. 7). This could indicate, if this sequence is indeed responsible for nuclear localization, the existence of a monopartite NLS for *S. pombe*. In the absence of more experimental data, the detailed analysis of the amino acid sequence reveals some fragments in the amino-terminal domain which could be responsible for the nuclear localization. In four cases (positions 3, 52, 72 and 75), the sequence KKXK is detected, and the positions 52 and 72 are separated by 16 amino acids, the last sequence being exactly KKQK; that is, the second component of the bipartite NLS in vertebrates, plants, and also NSR1. Any of these sequences is a candidate for being the NLS, and all of them are present in the mutant constructed by Léger-Silvestre et al. (1997).

PHYLOGENETIC ANALYSIS OF NUCLEOLIN AND NUCLEOLIN-LIKE PROTEINS

The high identity of the distinct proteins as to their primary structure, the presence of the same domains in all of them, and the same location of these in the

primary structure of each group of organisms, lead us to think of the existence of common ancestor for these proteins, which, then, could be defined as homologous protein. In order to test this hypothesis, we performed a phylogenetic analysis of the fifteen species in which the whole sequences of their nucleolins (or nucleolin-like proteins) are known.

For this purpose, the sequences of all of them were aligned using the software "Clustal-W",followed by the construction of the phylogenetic tree and the correction of distances by the algorithm "Jukes-Cantor Distance". The tree generated in this way establishes relationships among the organisms in the same way as they are currently known to be related by means of the bulk of biological data. It is remarkable that the study of a single protein produces such an exact representation of the phylogeny of fifteen organisms as diverse as those

with which we have been dealing. This fact evidences that the amino acid sequences of proteins store considerable evolutionary information.

In conclusion, all proteins studied in this work originated from the descent of an ancestor gene, which was already present in the ancestor species from which all concerned species have evolved. This means that nucleolin is a protein which has evolved slowly, most probably because the function carried out by this protein is a key function that was well established and well fixed in the ancestor species, even at the level of its molecular mechanism. Highly different organisms, such as the human being and yeast have in common a large proportion of their molecules. The evolutionary conservation of this protein makes possible the study of analogies and differences between organisms which are only distantly related.
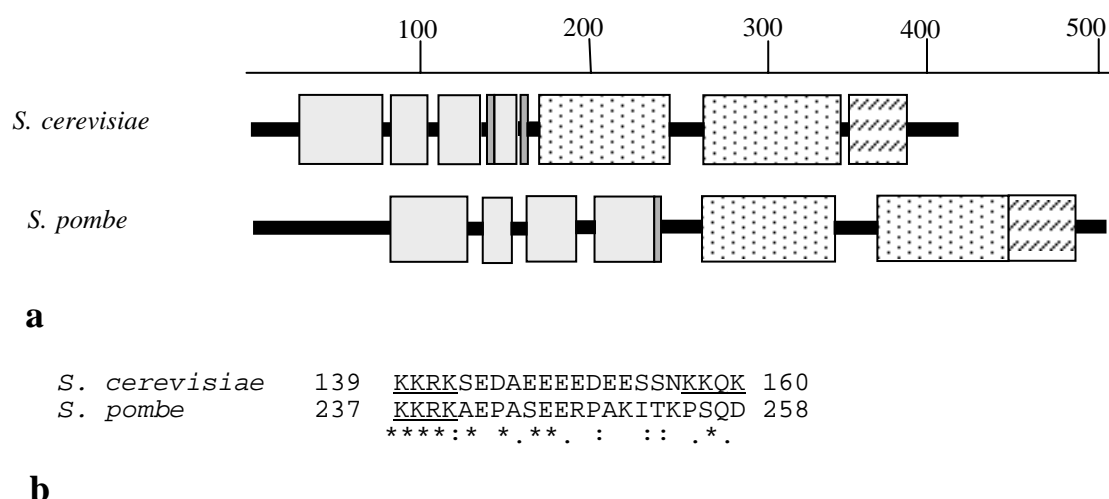

a

```
S. cerevisiae   139   KKRKSEDAEEEEDEESSNKKQK  160
S. pombe        237   KKRKAEPASEERPAKITKPSQD  258
                      ****:* *.**. :  :: .*.
```

b

Fig. 7. **Results obtained from sequences of nucleolin-like proteins in two yeast.** (a) Domain organization in the primary structure of the proteins. (b) Aligned sequences of the nuclear localization domain. Patterns, numbers and signs have the same meaning as in preceding figures
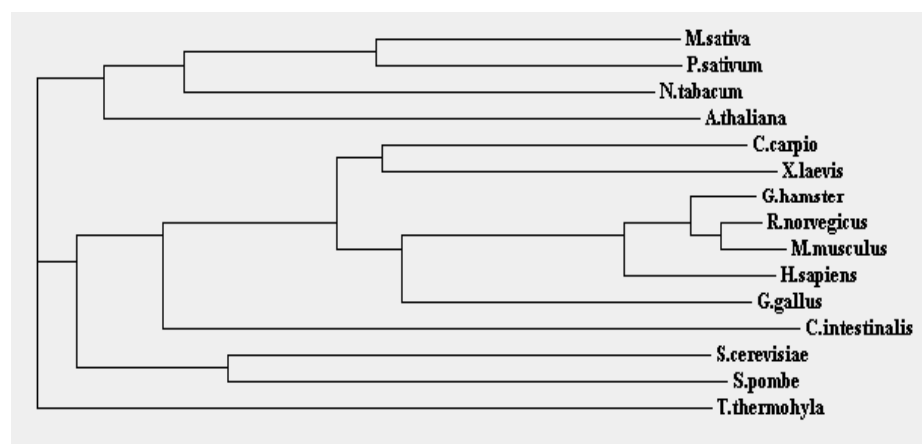

Fig. 8. **Phylogenetic tree generated from the amino acid sequences of nucleolins and nucleolin-like proteins taken from 15 different organisms.** The organization of the tree, obtained exclusively from the data of one protein, closely resembles the phylogenetic organization which is presently known from many biological data of different origin

REFERENCES

Alvarez M., C. Quezada, C. Navarro, A. Molina, P. Bouvet, M. Krauskopf, M.I. Vera: An increased expression of nucleolin is associated with a physiological nucleolar segregation. Biochem. Biophys. Res. Commun. 301: 152–158, 2003.

Belenguer P., M. Caizergues-Ferrer, J.C. Labbé, M. Dorée, F. Amalric: Mitosis-specific phosphorylation of nucleolin by p34^{cdc2} protein kinase. Mol. Cell. Biol. 10: 3607–3618, 1990.

Bögre L., C. Jonak, M. Mink, I. Meskiene, J. Traas, D.T.C. Ha, I. Swoboda, C. Plank, E. Wagner, E. Heberle-Bors, H. Hirt: Developmental and cell cycle regulation of alfalfa *nucMs1,* a plant homolog of the yeast Nsr1 and mammalian nucleolin. Plant Cell 8: 417–428, 1996.

Borer R.A., C.F. Lehner, H.M. Eppenberger, E.A. Nigg: Major nucleolar proteins shuttle between nucleus and cytoplasm. Cell 56: 379–390, 1989.

Bourbon H.M. and F. Amalric: Nucleolin gene organization in rodents: highly conserved sequences within three of the 13 introns. Gene 88: 187–196, 1990.

Bourbon H.M., B. Lapeyre, F. Amalric: Structure of the mouse nucleolin gene: The complete sequence reveals that each RNA binding domain is encoded by two independent exons. J. Mol. Biol. 200: 627–638, 1988.

Caizergues-Ferrer M., P. Belenguer, B. Lapeyre, F. Amalric, M.O. Wallace, M.O.J. Olson: Phosphorylation of nucleolin by a nuclear type NII protein kinase. Biochemistry 26: 7876–7883, 1987.

Caizergues-Ferrer M., P. Mariottini, C. Curie, B. Lapeyre, N. Gas, F. Amalric, F. Amaldi: Nucleolin from Xenopus laevis: cDNA cloning and expression during development. Genes Dev. 3: 324–333, 1989.

Combet C., C. Blanchet, C. Geourjon, G. Deleage: NPS@: network protein sequence analysis. Trends Biochem. Sci. 25: 147–150, 2000.

De Cárcer G., A. Cerdido, F.J. Medina: NopA64, a novel nucleolar phosphoprotein from proliferating onion cells, sharing immunological determinants with mammalian nucleolin. Planta 201: 487–495, 1997.

Ghisolfi L., A. Kharrat, G. Joseph, F. Amalric, M. Erard: Concerted activities of the RNA recognition and the glycine- rich C-terminal domains of nucleolin are required for efficient complex formation with pre-ribosomal RNA. Eur. J. Biochem. 209: 541–548, 1992.

Ginisty H., H. Sicard, B. Roger, P. Bouvet: Structure and functions of nucleolin. J. Cell Sci. 112: 761–772, 1999.

González-Camacho F. and F.J. Medina: Identification of specific plant nucleolar phosphoproteins in a functional proteomic analysis. Proteomics 4: 407–417, 2004.

Gulli M.P., J.-P. Girard, D. Zabetakis, B. Lapeyre, T. Mélèse, M. Caizergues-Ferrer: gar2 is a nucleolar protein from *Schizosaccharomyces pombe* required for 18S rRNA and 40S ribosomal subunit accumulation. Nucl. Acids Res. 23: 1912–1918, 1995.

Kaneko T., T. Kato, S. Sato, Y. Nakamura, E. Asamizu, S. Tabata: Sequence of Arabidopsis thaliana nucleolin-like protein. Protein Database EMBL/GenBank/DDBJ, 2000.

Lapeyre B., H.M. Bourbon, F. Amalric: Nucleolin, the major nucleolar protein of growing eukaryotic cells: an unusual protein structure revealed by the nucleotide sequence. Proc. Nat. Acad. Sci. USA 84: 1472–1476, 1987.

Lee W.C., Z. Xue, T. Mélèse: The NSR1 gene encodes a protein that specifically binds nuclear localization sequences and has two RNA recognition motifs. J. Cell Biol. 113: 1–12, 1991.

Léger-Silvestre I., M.P. Gulli, J. Noaillac-Depeyre, M. Faubladier, H. Sicard, M. Caizergues-Ferrer, N. Gas: Ultrastructural changes in the *Schizosaccharomyces pombe* nucleolus following the disruption of the *gar*2+ gene, which encodes a nucleolar protein structurally related to nucleolin. Chromosoma 105: 542–552, 1997.

Maeshima M. and N. Matsuda: Sequence of Nicotiana tabacum nucleolin-like protein. Protein Database EMBL/GenBank/DDBJ, 2002.

Maridor G., W. Krek, E.A. Nigg: Structure and developmental expression of chicken nucleolin and NO38: coordinate expression of two abundant non-ribosomal nucleolar proteins. Biochim. Biophys. Acta 1049: 126–133, 1990.

Martín M., L.F. García-Fernández, S. Moreno Díaz de la Espina, J. Noaillac-Depeyre, N. Gas, F.J. Medina: Identification and localization of a nucleolin homologue in onion nucleoli. Exp. Cell Res. 199: 74–84, 1992.

McGrath K.E., J.F. Smothers, C.A. Dadd, M.T. Madireddi, M.A. Gorovsky, C.D. Allis: An abundant nucleolar phosphoprotein is associated

with ribosomal DNA in *Tetrahymena* macronuclei. Mol. Biol. Cell 8: 97–108, 1997.

Medina F.J., F. González-Camacho, A. Cerdido, G. De Cárcer: In situ localization of the onion nucleolar protein NopA64 is dependent on cell proliferation mechanisms and cell cycle phases. In Dini,L. and M.Catalano (eds.): Proc. 5th Multinational Cong. Electron Microscopy. Rinton Press Inc., Princeton, New Jersey. 2001, pp. 197–207.

Olson M.O.J.: The role of proteins in nucleolar structure and function. In Strauss P.R. and S.H. Wilson (eds.): The eukaryotic nucleus. Molecular biochemistry and macromolecular assemblies. The Telford Press, Caldwell, New Jersey 1991, pp. 519–559.

Orrick L., M.O.J. Olson, H. Busch: Comparison of nucleolar proteins of normal rat liver and Novikoff hepatoma ascites cells by two dimensional polyacrylamide gel electrophoresis. Proc. Nat. Acad. Sci. USA 70: 1316–1320, 1973.

Peter M., J. Nakagawa, M. Dorée, J.C. Labbé, E.A. Nigg: Identification of major nucleolar proteins as candidate mitotic substrates of cdc2 kinase. Cell 60: 791–801, 1990.

Prestayko A.W., G.R. Klomp, D.J. Schmoll, H. Busch: Comparison of proteins of ribosomal subunits and nucleolar preribosomal particles from Novikoff hepatoma ascites cells by two-dimensional polyacrylamide gel electrophoresis. Biochemistry 13: 1945–1951, 1974.

Rankin M.L., M.A. Heine, S. Xiao, M.D. LeBlanc, J.W. Nelson, P.J. DiMario: A complete nucleolin cDNA sequence from *Xenopus laevis*. Nucl. Acids Res. 21: 169, 1993.

Rappsilber J., U. Ryder, A. Lamond, M. Mann: Large-Scale proteomic analysis of the human spliceosome. Genome Res. 12: 1231–1245, 2002.

Schmidt-Zachmann M.S. and E.A. Nigg: Protein localization to the nucleolus: A search for targeting domains in nucleolin. J. Cell Sci. 105: 799–806, 1993.

Srivastava M., O.W. McBride, P.J. Fleming, H.B. Pollard, A.L. Burns: Genomic organization and chromosomal localization of the human nucleolin gene. J. Biol. Chem. 265: 14922–14935, 1990.

Tanaka K.J., H. Kawamuea, T. Nishikata: The transcript coding for an RNA-binding protein is localized in the anterior side of the ascidian 2-cell stage embryo. Dev. Genes Evol. 210: 464–466, 2004.

Thompson J.D., D.G. Higgins, T.J. Gibson: Clustal-W, improving the sensitivity of progressive multiple alignment through sequence weighting. Nucl. Acids Res. 22: 4673–4680, 1994.

Tong C.G., S. Reichler, S. Blumenthal, J. Balk, H.L. Hsieh, S.J. Roux: Light regulation of the abundance of mRNA encoding a nucleolin-like protein localized in the nucleoli of pea nuclei. Plant Physiol. 114: 643–652, 1997.

Tuteja R. and N. Tuteja: Nucleolin: a multifunctional major nucleolar protein. CRC Crit. Rev. Biochem. Mol. Biol. 33: 407–436, 1998.

✉ **Address:**

Franscisco Javier Medina, Centro de Investigaciones Biológicas (CSIC), Ramiro de Maeztu 9, E-28040 Madrid, Spain; fjmedina@cib.csic.es