DOI: 10.1534/genetics.106.058990

Long-Term Evolution and Functional Diversification in the Members of the Nucleophosmin/Nucleoplasmin Family of Nuclear Chaperones

José M. Eirín-López,*,† Lindsay J. Frehlick* and Juan Ausió*,[1]

*Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia V8W 3P6, Canada and †Departamento de Biología Celular y Molecular, Universidade da Coruña, 15071 A Coruña, Spain

Manuscript received April 12, 2006
Accepted for publication May 24, 2006

ABSTRACT

The proper assembly of basic proteins with nucleic acids is a reaction that must be facilitated to prevent protein aggregation and formation of nonspecific nucleoprotein complexes. The proteins that mediate this orderly protein assembly are generally termed molecular (or nuclear) chaperones. The nucleophosmin/nucleoplasmin (NPM) family of molecular chaperones encompasses members ubiquitously expressed in many somatic tissues (NPM1 and -3) or specific to oocytes and eggs (NPM2). The study of this family of molecular chaperones has experienced a renewed interest in the past few years. However, there is a lack of information regarding the molecular evolution of these proteins. This work represents the first attempt to characterize the long-term evolution followed by the members of this family. Our analysis shows that there is extensive silent divergence at the nucleotide level suggesting that this family has been subject to strong purifying selection at the protein level. In contrast to NPM1 and NPM-like proteins in invertebrates, NPM2 and NPM3 have a polyphyletic origin. Furthermore, the presence of selection for high frequencies of acidic residues as well as the existence of higher levels of codon bias was detected at the C-terminal ends, which can be ascribed to the critical role played by these residues in constituting the acidic tracts and to the preferred codon usage for phosphorylatable amino acids at these regions.

NUCLEOSOMES and ribosomes are highly organized nucleoprotein complexes. Orderly assembly of the protein components onto the nucleic acid component is facilitated in part by assembly proteins that are called molecular (or nuclear) chaperones (Laskey *et al.* 1978; Dingwall and Laskey 1990; Philpott *et al.* 2000). Among the different nuclear chaperones present in metazoan organisms, the nucleophosmin/nucleoplasmin (NPM) family comprises three major functional types: NPM1, NPM2, and NPM3.

NPM1 is also known as nucleophosmin, B23, numatrin in mammals, and NO38 in amphibians. The N-terminal domain of NO38 forms an eight-stranded β-barrel and a set of five monomers fit together to form a stable pentamer (Namboodiri *et al.* 2004). Members of this type are among the most abundant phosphorylated proteins in the nucleolus of somatic cells and have been implicated in ribosome assembly. They have been shown to be involved in nucleic acid binding (Dumbar *et al.* 1989), ribonuclease activity (Herrera *et al.* 1995), and association with maturing preribosomal ribonucleoprotein particles (Olson *et al.* 1986). Their role during cell proliferation was originally believed to be restricted to ribosome maturation (Feuerstein *et al.* 1988; Chan *et al.* 1989). However, NPM1 is a multifunctional protein that can also bind pRb synergistically stimulating DNA polymerase α (Takemura *et al.* 1999). NPM1 also appears to be intimately involved in centrosome duplication, as it associates specifically with unduplicated centrosomes, and phosphorylation of NPM1 by CDK-2/cyclin E is required for duplication to occur (Okuda *et al.* 2000). Furthermore, NPM1 members are involved in several transcription processes through their interaction with transcription factors and histone chaperone activity (Kondo *et al.* 1997; Okuwaki *et al.* 2001; Swaminathan *et al.* 2005; Weng and Yung 2005). NPM1 is induced after genotoxic stress, protecting cells against DNA damage by either increasing their DNA repair capacity or decreasing the apoptotic signal (Wu *et al.* 2002; Maiguel *et al.* 2004). In addition, members of the NPM1 family also may be involved in the transport of ribosomal or other nuclear proteins across the nuclear membrane, as these proteins are known to shuttle between the cytoplasm and nucleus and to stimulate the nuclear importation of proteins (Borer *et al.* 1989).

The NPM2 type (or nucleoplasmin) was first isolated from *Xenopus laevis* egg extracts and this protein represents an archetypal molecular chaperone (Laskey *et al.* 1978; Earnshaw *et al.* 1980; Dingwall and Laskey 1990), which is found natively bound to histones H2A and H2B facilitating their storage in oocytes and eggs (Dingwall and Laskey 1990; Philpott *et al.* 2000; Burns *et al.* 2003). NPM2 proteins are acidic, thermostable, and form oligomeric complexes consisting of five

[1]*Corresponding author:* University of Victoria, Department of Biochemistry and Microbiology, Petch Building, Room 220, Victoria, BC V8W 3P6, Canada. E-mail: jausio@uvic.ca

identical subunits of ~22 kDa that can both bind histones and promote chromatin assembly *in vitro* (Laskey *et al.* 1978; Dingwall *et al.* 1987; Ohsumi and Katagiri 1991b). Nucleoplasmin is the most abundant protein in Xenopus oocytes and participates in the decondensation of sperm chromatin after fertilization by removing the sperm nuclear basic proteins (SNBPs) (Ohsumi and Katagiri 1991a; Philpott *et al.* 1991). In concert with the unrelated protein N1/N2, which binds histones H3 and H4, it participates in the assembly of nucleosomes in the male pronucleus (Dingwall and Laskey 1990; Philpott *et al.* 2000). The chromatin remodeling ability of Xenopus NPM2 at fertilization has been shown to be dependent on its extent of phosphorylation. This post-translational modification enhances the H2A/H2B exchange activity during the decondensation of sperm chromatin, and it increases its ability to promote nucleosome assembly *in vitro* (Leno *et al.* 1996).

The nucleoplasmin monomer folds into two different domains: an N-terminal core domain, which consists of an eight-stranded β-barrel that fits snugly within a stable pentamer, and a solvent-exposed C-terminal acidic domain, which is rich in turn structures (Hierro *et al.* 2001; Namboodiri *et al.* 2004). In addition to the presence of two acidic tracts, the C-terminal region contains a bipartite nuclear localization sequence. Within amphibians, nucleoplasmin is a highly conserved protein (Prado *et al.* 2004; Frehlick *et al.* 2006) that has been shown to bind equally well to the three major types of SNBPs: protamines (Rice *et al.* 1995; Prieto *et al.* 2002), protamine-like type (Rice *et al.* 1995; Prado *et al.* 2004; Ramos *et al.* 2005), and histone type (Ramos *et al.* 2005), both *in vivo* and *in vitro*.

The lack of specificity with which amphibian nucleoplasmin binds to SNBPs contrasts with the higher specificity with which it binds to core histones (Arnan *et al.* 2003), especially to H2A-H2B (Dutta *et al.* 2001).

An intriguing aspect of other NPM2 members in mammals is that their absence has no effect on meiosis, sperm DNA decondensation, or the first S-phase after fertilization. It is possible in this instance that NPM1 and NPM3, which are present in oocytes (such as in mice), could possibly provide a mechanism of compensation (Burns *et al.* 2003).

NPM3 (also known as NO29 in Xenopus) includes a group of less well-characterized proteins. In the case of Xenopus, NO29 migrates in SDS–PAGE gels as a 29-kDa protein and shares many physical characteristics with NPM1 and NPM2 at the structural level, including an acidic domain, multiple potential phosphorylation sites, and a putative nuclear localization signal (Zirwes *et al.* 1997). Proteins of the NPM3 type include a group of molecular chaperones that are ubiquitously expressed in many tissues, showing the highest levels in pancreas and testis and the lowest in lung, in the case of humans (Shackleford *et al.* 2001). Like NPM1, NPM3 is mainly localized to the nucleolus in interphase cells and active rRNA transcription appears to be required for this localization (Huang *et al.* 2005). NPM3 forms a complex with NPM1 and has been implicated in regulation of NPM1 function in ribosome biogenesis (Huang *et al.* 2005). In addition, inhibition of NPM3 expression in mammalian oocytes prohibits paternal chromatin decondensation (McLay and Clarke 2003), suggesting this family member may also serve multiple functions.

The diversification of NPM proteins during animal evolution must have been determined by the presence of different structural and functional constraints acting on these proteins. In this article we take advantage of the molecular data currently available for NPM proteins of different taxonomic groups to analyze their long-term evolution. Special attention is paid to the relative importance of the functional and structural constraints acting at the protein and nucleotide level.

MATERIALS AND METHODS

A total of 106 nucleotide coding sequences belonging to 25 different species of metazoans have been used in our analyses (see supplemental Table 1 at http://www.genetics.org/supplemental/). These include 104 NPM family sequences (54 NPM1, 27 NPM2, 13 NPM3, 2 NPM3 pseudogenes, 8 NPM-like from invertebrates) and 2 outgroup sequences (N1/N2 nuclear histone binding protein from *X. laevis* and nuclear autoantigenic sperm protein, NASP, from human). Sequences retrieved from databases were subsequently corrected for errors in accession numbers and nomenclature, and they were aligned on the basis of their translated amino acid sequences using the CLUSTAL_X (Thompson *et al.* 1997) and BIOEDIT programs (Hall 1999) with the default parameters. The alignment of the complete set of sequences consisted of 3873 nucleotide positions (excluding the start and stop codons) corresponding to 1293 amino acid sites. The independent alignments for each of the four NPM lineages are shown in supplemental alignments 1–4 (http://www.genetics.org/supplemental/) and in all cases were checked for errors by visual inspection. The distinction between the N-terminal and the acidic C-terminal regions of NPM proteins was established in this work on the basis of the information available for tertiary protein structure of the core regions: NPM1 [N terminus, nucleotide position (nt pos.) 1–363; C terminus, nt pos. 364–927] (Namboodiri *et al.* 2003, 2004) and NPM2 (N terminus, nt pos. 1–369; C terminus, nt pos. 370–684) (Dutta *et al.* 2001). For the cases of NPM3 (N terminus, nt pos. 1–435; C terminus, nt pos. 436–630) and NPM-like from invertebrates (N terminus, nt pos. 1–363; C terminus, nt pos. 364–1287) this was deduced from the alignment of these sequences with those of NPM1 and NPM2.

All molecular evolutionary analyses in this work were carried out using the program MEGA v. 3.1 (Kumar *et al.* 2004). The extent of nucleotide and amino acid divergence between sequences was estimated by means of the uncorrected differences (*p*-distance) as this distance is known to give better results than more complicated methods when the number of sequences is large and the number of positions used is relatively small, because of its smaller variance (Nei and Kumar 2000). The numbers of synonymous (p_S) and nonsynonymous (p_N) nucleotide differences per site were computed using the modified Nei–Gojobori method (Zhang *et al.* 1998), providing in both cases the transition/transversion ratio (R). Distances were estimated using the complete-deletion option

(except for the case of the protein and nucleotide phylogenetic tree reconstructions, where the pairwise-option deletion was used) and standard errors were calculated by the bootstrap method with 1000 replicates. The presence and nature of selection was tested in NPM genes by using the codon-based *Z*-test for selection, establishing the alternative hypothesis as H_1: $p_N < p_S$ and the null hypothesis as H_0: $p_N = p_S$ (Nei and Kumar 2000). The *Z*-statistic and the probability that the null hypothesis is rejected were obtained, indicating the significance level as $^{**}P$ ($P < 0.001$) and $^{*}P$ ($P < 0.05$).

The presence of selection in the three main NPM lineages (1–3) was further studied by testing for deviations from neutrality. The GC content at fourfold degenerate sites was assumed to represent the genomic GC content and in addition was considered as an approximation to the neutral expectation. The influence of selection on certain amino acids was analyzed by determining the correlation between the genomic GC content and the proportion of GC-rich (GAPW) and GC-poor (FYMINK) residues. Under the neutral model, GC-rich and GC-poor amino acids will be positively and negatively correlated with genomic GC content, respectively (Kimura 1983). If the frequency of these amino acids is influenced by selection, no correlation between genomic GC content and amino acid frequency would be expected (Rooney 2003). Correlations were computed for complete sequences and for discriminating between the N-terminal and C-terminal segments by using the Spearman rank correlation coefficient. The statistical significance was assessed through regression analysis. Given that all changes at second codon positions are nonsynonymous, whereas changes at fourfold degenerate sites are synonymous, the effects of selection and mutation bias will be apparent by comparing these two types of sites. Under the neutral model, nucleotide frequencies at second codon positions should not be significantly different from the frequencies at fourfold degenerate sites (Rooney *et al.* 2000). If selection is biasing the frequency of concrete amino acids, frequencies of nucleotides at second codon positions will be higher than expected under the neutral model, represented by fourfold degenerate sites. Comparisons were also made for NPM lineages by discriminating between N-terminal and C-terminal domains and the statistical significance was determined by *t*-tests once the homogeneity of standard deviations was confirmed.

The neighbor-joining tree-building method (Saitou and Nei 1987) was used to reconstruct the phylogenetic trees. To assess that our results are not dependent on this choice, phylogenetic inference analyses were completed by the reconstruction of a maximum-parsimony tree (Rzhetsky and Nei 1992) using the close-neighbor-interchange (CNI) search method with search level 1 and with 10 replications for the random addition trees option. We decided to combine the bootstrap (Felsestein 1985) and the interior-branch test methods (Rzhetsky and Nei 1992; Sitnikova 1996) to test the reliability of the obtained topologies, producing the bootstrap probability (BP) and the confidence probability (CP) values for each internal branch, assuming BP $> 80\%$ and CP $\geq 95\%$ as statistically significant (Sitnikova *et al.* 1995). Human NASP (Welch *et al.* 1990) and the nuclear histone binding N1/N2 protein from *X. laevis* (Kleinschmidt *et al.* 1986) were used as outgroups in the reconstruction, given that they are also histone-binding proteins but unrelated to NPM proteins (Shackleford *et al.* 2001).

The analysis of the nucleotide variation across different NPM coding regions was performed using a sliding-window approach, by estimating the total (π) and the synonymous (π_S) nucleotide diversity (average number of nucleotide differences per site between two sequences) with a window length of 20 bp and a step size of 5 bp (for π) and a window length of 10 bp and a step size of 5 bp (for π_S). The codon usage bias in NPM genes was estimated as the effective number of codons (ENC) (Wright 1990), where the highest value (61) indicates that all synonymous codons are used equally (no bias) and the lowest (20) that only a preferred codon is used in each synonymous class (extreme bias). Preferences in the usage of different synonymous triplets were tested by multiple range tests, discriminating among means using the Fisher's least significant difference (LSD) procedure. Both analyses were conducted with the program DnaSP v. 4.10 (Rozas *et al.* 2003).

The three-dimensional structure of *X. laevis* NO29 (NPM3) core monomer was modeled using the coordinates determined for the crystal of the Xenopus NO38-core monomer (Namboodiri *et al.* 2004) as a reference using the SWISS-MODEL server (Schwede *et al.* 2003). Post-translational phosphorylation sites at serines, threonines, and tyrosines in the NPM protein sequences were predicted by using the NetPhos v. 2.0 server (Blom *et al.* 1999), selecting only those positions scoring >0.5 in phosphorylation potential.

RESULTS

Evolution of the NPM protein family: The protein phylogeny was reconstructed from 106 NPM sequences (including 52 nonredundant sequences) of 25 species belonging to different metazoan phyla (Figure 1; see supplemental Table 1 at http://www.genetics.org/supplemental/). The three NPM types and the invertebrate NPM-like type are well defined by the topology and by the confidence levels calculated for each internal node. The different taxonomic groups are also well differentiated with respect to each of the NPM types. Given that many NPM sequences are not yet annotated or are wrongly assigned to their corresponding subtype in the databases, the present topology represents a powerful tool that allowed 17 previously unannotated sequences to be unequivocally ascribed to a given NPM type. Seven of these sequences were identified *in silico* from the draft genomes of rhesus monkey, orangutan, zebrafish, the fish *Tetraodon nigroviridis*, the yellow fever mosquito, and the African malaria mosquito.

While the tree topology shows the presence of a monophyletic origin for NPM1 proteins, the polyphyletic origin observed for NPM2 and NPM3 is the result of differences between mammals and amphibians, giving rise to independent groups in the phylogeny. In NPM2, two differentiation events (nodes 1 and 1′ in the tree) occurred that led first to the nucleoplasmin lineage from mammals and subsequently to the differentiation of the nucleoplasmin lineage from amphibians and fish. It is important to note that, as with NPM2, amphibian and fish NPM3 proteins (closer to NPM1) also differentiate from mammalian NPM3 proteins. This pattern of differentiation, which was also corroborated by reconstructing a maximum-parsimony phylogenetic tree, may bear important functional implications and will be further discussed below.

The lineages corresponding to NPM3 and NPM1 differentiated later than NPM2 (nodes 2 and 3 in the tree, respectively). The relationship between types 1 and

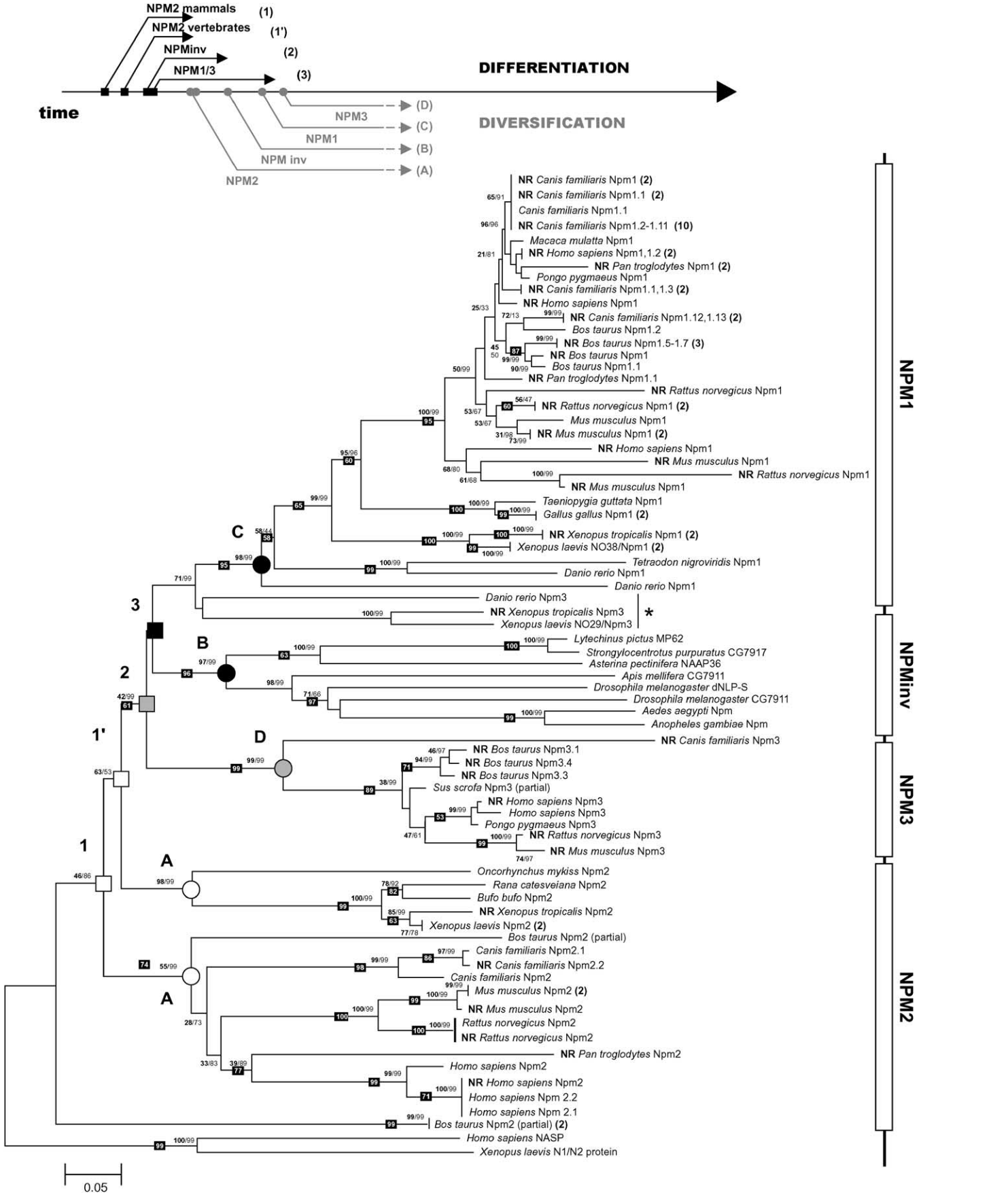
NPM2 mammals
(1)
NPM2 vertebrates
(1')
NPMinv
(2)
NPM1/3
(3)
DIFFERENTIATION
time
(D)
NPM3
(C)
DIVERSIFICATION
NPM1
(B)
NPM inv
(A)
NPM2
NR Canis familiaris Npm1 (2)
NR Canis familiaris Npm1.1 (2)
Canis familiaris Npm1.1
NR Canis familiaris Npm1.2-1.11 (10)
Macaca mulatta Npm1
NR Homo sapiens Npm1,1.2 (2)
NR Pan troglodytes Npm1 (2)
Pongo pygmaeus Npm1
NR Canis familiaris Npm1.1,1.3 (2)
NR Homo sapiens Npm1
NR Canis familiaris Npm1.12,1.13 (2)
Bos taurus Npm1.2
NR Bos taurus Npm1.5-1.7 (3)
NR Bos taurus Npm1
Bos taurus Npm1.1
NR Pan troglodytes Npm1.1
NR Rattus norvegicus Npm1
NR Rattus norvegicus Npm1 (2)
Mus musculus Npm1
NR Mus musculus Npm1 (2)
NR Homo sapiens Npm1
NR Mus musculus Npm1
NR Rattus norvegicus Npm1
NR Mus musculus Npm1
Taeniopygia guttata Npm1
Gallus gallus Npm1 (2)
C
NR Xenopus tropicalis Npm1 (2)
Xenopus laevis NO38/Npm1 (2)
Tetraodon nigroviridis Npm1
Danio rerio Npm1
Danio rerio Npm1
3
Danio rerio Npm3
NR Xenopus tropicalis Npm3
Xenopus laevis NO29/Npm3
*
B
Lytechinus pictus MP62
Strongylocentrotus purpuratus CG7917
Asterina pectinifera NAAP36
2
Apis mellifera CG7911
Drosophila melanogaster dNLP-S
Drosophila melanogaster CG7911
Aedes aegypti Npm
Anopheles gambiae Npm
1'
NR Canis familiaris Npm3
D
NR Bos taurus Npm3.1
NR Bos taurus Npm3.4
NR Bos taurus Npm3.3
Sus scrofa Npm3 (partial)
NR Homo sapiens Npm3
Homo sapiens Npm3
Pongo pygmaeus Npm3
NR Rattus norvegicus Npm3
NR Mus musculus Npm3
1
A
Oncorhynchus mykiss Npm2
Rana catesveiana Npm2
Bufo bufo Npm2
NR Xenopus tropicalis Npm2
Xenopus laevis Npm2 (2)
Bos taurus Npm2 (partial)
Canis familiaris Npm2.1
NR Canis familiaris Npm2.2
Canis familiaris Npm2
A
Mus musculus Npm2 (2)
NR Mus musculus Npm2
Rattus norvegicus Npm2
NR Rattus norvegicus Npm2
NR Pan troglodytes Npm2
Homo sapiens Npm2
NR Homo sapiens Npm2
Homo sapiens Npm 2.2
Homo sapiens Npm 2.1
Bos taurus Npm2 (partial) (2)
Homo sapiens NASP
Xenopus laevis N1/N2 protein
0.05
NPM1
NPMinv
NPM3
NPM2

TABLE 1

Average numbers of amino acid (p_{AA}) and nucleotide (p_{NT}) differences per site, and average synonymous (p_S) and nonsynonymous (p_N) differences per site in the four NPM lineages defined in Figure 1, discriminating among complete coding regions, N-terminal and C-terminal domains

NPM lineage	p_{AA} (SE)	p_{NT} (SE)	p_S (SE)	p_N (SE)	R	P
NPM1 complete	0.218 ± 0.014	0.227 ± 0.009	0.508 ± 0.018	0.139 ± 0.010	0.9	16.898**
NPM1 N terminus	0.110 ± 0.018	0.186 ± 0.013	0.511 ± 0.022	0.067 ± 0.011	1.3	16.924**
NPM1 C terminus	0.294 ± 0.020	0.256 ± 0.011	0.492 ± 0.025	0.188 ± 0.015	0.8	10.005**
NPM2 complete	0.401 ± 0.024	0.335 ± 0.012	0.604 ± 0.018	0.252 ± 0.017	0.8	14.075**
NPM2 N terminus	0.396 ± 0.028	0.342 ± 0.015	0.595 ± 0.022	0.255 ± 0.021	0.9	11416**
NPM2 C terminus	0.408 ± 0.039	0.324 ± 0.020	0.603 ± 0.029	0.250 ± 0.030	0.8	8.095**
NPM3 complete	0.371 ± 0.022	0.336 ± 0.012	0.573 ± 0.019	0.254 ± 0.017	0.9	12.877**
NPM3 N terminus	0.306 ± 0.021	0.297 ± 0.013	0.590 ± 0.023	0.199 ± 0.016	0.7	13.897**
NPM3 C terminus	0.608 ± 0.040	0.472 ± 0.023	0.642 ± 0.038	0.425 ± 0.033	0.6	4.840**
NPM-like complete	0.517 ± 0.019	0.434 ± 0.015	0.682 ± 0.020	0.359 ± 0.023	0.7	9.754**
NPM-like N terminus	0.475 ± 0.036	0.434 ± 0.019	0.854 ± 0.023	0.348 ± 0.029	0.7	8.353**
NPM-like C terminus	0.607 ± 0.046	0.434 ± 0.024	0.671 ± 0.034	0.380 ± 0.041	0.7	5.428**

The average transition/transversion ratio used in the estimation of p_S and p_N is denoted as R. $p_S > p_N$ in all comparisons (**$P < 0.001$). SE indicates standard errors calculated by the bootstrap method with 1000 replicates.

3 ($p = 0.300 \pm 0.033$ substitutions per site) was the closest of all, followed by types 2 and 3 ($p = 0.411 \pm 0.071$), and types 1 and 2 ($p = 0.453 \pm 0.035$). This is most likely the result of their different polyphyletic origin as well as the longer time elapsed since the differentiation of NPM2 (and thus, more accumulated changes). This observation is supported by the protein variation observed within lineages (Table 1), which is higher for NPM2 ($p = 0.401 \pm 0.024$ substitutions per site), followed by NPM3 and NPM1 ($p = 0.371 \pm 0.022$ and $p = 0.218 \pm 0.014$, respectively), in agreement with the temporal differentiation frame of the NPM types. The group corresponding to NPM-like sequences from invertebrates also shows a monophyletic origin and shares the closest common ancestor with the NPM1 group of proteins.

After the differentiation of the three vertebrate NPM lineages, the early diversification of NPM2, which apparently took place at the same point in mammals and amphibians/fishes (node A in the tree), was followed by that of NPM-like from invertebrates (node B) and by NPM1 (node C) and NPM3 (node D). Interestingly, the differentiation process leading to the appearance of the three vertebrate NPM types appears to be absent in invertebrates. Nevertheless, some diversification is also present in the invertebrate NPM-like proteins, as is evidenced from their presence in protostomes (such as arthropods) and deuterostomes (such as echinoderms) (see Figure 1).

Nucleotide variation among NPM genes: To further examine the evolutionary relationships among NPM family members, an additional phylogenetic tree was reconstructed (Figure 2) from the genomic nucleotide regions coding for the different NPM proteins analyzed in Figure 1. The two mouse NPM3 pseudogenes were also included in this analysis (MacArthur and Shackleford 1997). However, because any NPM comparison of sequences within and between species was close to or had even reached the saturation level, the nucleotide-based tree was of low reliability and in what follows we mainly focus our discussion on the protein phylogeny (Figure 1). In the tree that is based on the synonymous nucleotide differences per site, shown in Figure 2, different NPM types intersperse extensively with each other. This, together with the long branch lengths observed by this analysis, reveals that the nature of the nucleotide

Figure 1.—Phylogenetic relationships among NPM proteins. The reconstruction was carried out by calculating the evolutionary amino acid *p*-distances from the NPM sequences of all the organisms analyzed (see supplemental Table 1 at http://www.genetics.org/supplemental/). NPM types are indicated on the right near the species names and the number of sequences analyzed is indicated within parentheses. Numbers for interior nodes indicate BP (boldface type) and CP (lightface type) confidence values. The numbers in black boxes indicate the bootstrap values obtained in the reconstruction of the maximum parsimony trees, carried out using all the informative positions in the alignment, by the close-neighbor-interchange (CNI) search method with search level 1 and with 10 replications for the random addition trees option. Confidence values were based on 1000 replications, and are only shown when the value is >50%. The differentiation and diversification events indicated by squares and circles at the nodes in the phylogeny are summarized in the upper part of the figure. The asterisk (*) denotes the only exception to the functional clustering of NPM proteins exhibited by Xenopus NPM3 and *Danio rerio* NPM3. Nonredundant sequences are indicated in the tree with the prefix NR.

variation in the different NPM lineages is essentially synonymous. The level of silent variation was very similar for each of the three vertebrate lineages (NPM1, $p_S = 0.508 \pm 0.018$; NPM2, $p_S = 0.604 \pm 0.018$; NPM3 $p_S = 0.573 \pm 0.019$, Table 1) and slightly higher in the invertebrate NPM-like genes ($p_S = 0.682 \pm 0.020$). When comparing these values with the nonsynonymous differences, we found that p_S was always significantly greater than p_N [$^{**}P < 0.001$, *Z*-test of selection, (Table 1)].

Although the nucleotide coding sequences of these proteins have diverged extensively through silent substitutions, different NPMs from the same species do not necessarily cluster together in the phylogenies on the basis of their protein sequences (Figure 1) and synonymous nucleotide substitutions (Figure 2). In general, the amount of silent variation was very high between NPM coding regions and the range of p_S values was nearly the same when these regions were compared both within and between related species (supplemental Table 2 at http://www.genetics.org/supplemental/). From comparison of NPMs of different types from different vertebrate species we observed that genes from a species are not more closely related to each other than they are to NPM genes belonging to very different species of vertebrates (supplemental Table 3 at http://www.genetics.org/supplemental/). For instance, the average synonymous divergence between human NPM1 and NPM2 genes is $\sim 0.855 \pm 0.048$ substitutions/site, which is significantly higher than that observed between human NPM1 and any of the other types in either human or any other vertebrates.

Importantly, the relative divergence of mouse NPM3 pseudogenes from functional NPM3 genes was estimated as 0.222 ± 0.012 nucleotide substitutions/site, which is not significantly different from the variation presented within functional NPM3 genes (0.287 ± 0.011 substitutions/site). This observation, together with the short branch lengths exhibited by pseudogenes in the nucleotide phylogeny and their clustering with mouse functional NPM3 genes, suggests that the pseudogenization events are not very ancestral, as there was not enough time to accumulate a high level of divergence.

By discriminating between the N-terminal core region and the disordered C terminus, it is possible to ascribe a significantly lower amino acid variation to the former when compared with the latter (Table 1). Conversely, the nucleotide variation is roughly the same in terms of silent variation in both domains of the molecule and always significantly greater than the nonsilent variation ($^{**}P < 0.001$, *Z*-test of selection). Nevertheless, the nonsilent variation is significantly lower at N-terminal domains of the proteins. This suggests the presence of the strongest functional constraints in this region, which in turn is the main target of the purifying selection acting on NPM proteins. The NPM-like proteins from invertebrates appear to depart from this observation, as similar levels of protein, nucleotide, and silent and nonsilent variation are observed when comparing the complete sequence with that of the different domains of the protein (Table 1).

The nature of the nucleotide variation exhibited by sequences among different species was further analyzed by calculating the nucleotide diversity (π) and the synonymous nucleotide diversity (π_S) across NPM sequences using a sliding-window approach, as shown in Figure 3. The relative contribution of π_S to π is evident, as in most cases the overall amount of nucleotide variation is the result of the underlying synonymous variation. While on average the amount of π_S ranges between 0.55 and 0.70 substitutions per site along the four different types of NPM sequences, a slight increase in the value of π in the case of NPM1 and NPM3 can be observed at C-terminal regions. This is most likely due to a relaxation of the structural and functional constraints in these regions of the molecules. In contrast, the inverse peaks of nucleotide diversity observed at C-terminal regions of the different NPM genes are determined by the presence of acidic tracts exclusively composed by either glutamic or aspartic acid residues, which are both encoded by twofold degenerate codons. The values of π and π_S appear also to be constrained by the presence of the nuclear localization signal (NLS) in the C-terminal region of NPM2, resulting in a reduced nucleotide variation in the segment composing this element. In the case of NPM-like genes from invertebrates, the massive divergence results in a great number of indels when comparing the different sequences, which makes it very difficult to discern between different patterns of variation.

Amino acid frequency and nucleotide composition of NPMs: The different members of the NPM family are characterized by glutamic and aspartic acid-rich acidic tracts located at the C-terminal regions of the molecule, which are involved in interactions with other proteins (Figure 4). The presence of selection for certain biased amino acids in the NPM lineages was first analyzed by determining the correlation coefficients between GC content and the frequency of GC-rich and GC-poor amino acids, schematized in Figure 5. In the case of the NPM1 lineage, neither the frequency of GC-rich nor that of GC-poor amino acids is correlated with the GC content in the complete molecule or discriminating in the N-terminal and the C-terminal domains (Table 2). In the case of NPM2, a significant negative correlation was observed between the GC content and the frequency of GC-rich and GC-poor amino acids when analyzing the complete molecule, as well as in the concrete case of the C-terminal domain, where a significant negative correlation is observed between the GC content and the frequency of GC-rich residues. In the case of NPM3, significant correlations were observed only in the case of the complete molecule, positive between GC content and GC-rich amino acids and negative with respect to GC-poor residues. Thus, the only case in which a

FIGURE 2.—Phylogenetic neighbor-joining tree of NPM complete nucleotide-coding sequences. The reconstruction was performed by using the number of synonymous nucleotide differences per site (p_S) estimated by the modified Nei–Gojobori method (p-distance). The NPM types, number of sequences, redundance, and BP and CP confidence values are indicated as in Figure 1. The NPM3 pseudogenes from mouse are referred as Ψ.

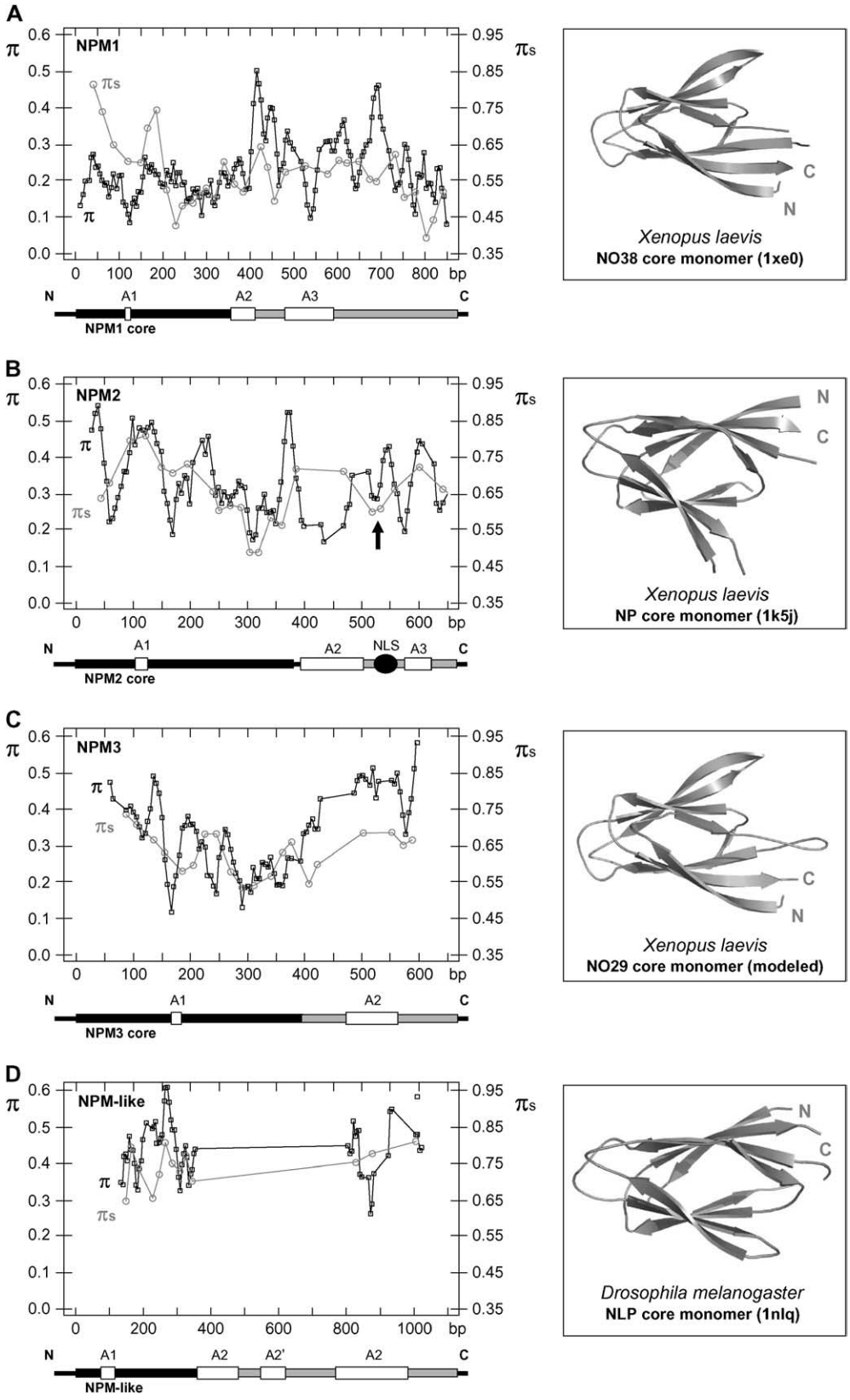

FIGURE 3.—Total (π) and synonymous (π_S) nucleotide diversity (expressed as the average number of nucleotide difference per site) across the coding regions of NPM1 (A), NPM2 (B), NPM3 (C), and NPM-like from invertebrates (D). The diversity values were calculated using a sliding-window approach with a window length of 20 bp and a step size of 5 bp (for π) and a window length of 10 bp and a step size of 5 bp (for π_S). The corresponding secondary structure, as well as the tertiary structure, of each NPM type is represented below and on the right of the corresponding graph, respectively. In the NPM secondary structures, the core regions are indicated by solid boxes, the C-terminal segments by shaded boxes, the acidic tracts by open boxes, and the nuclear localization signal (NLS) by a solid circle.

significant correlation agrees with the predictions of the neutral model is the negative correlation observed between GC content and the frequency of GC-poor residues in the case of the complete NPM2 molecules.

While in the case of GC-rich amino acids there are no representatives found with significantly higher frequencies than the others, lysine shows high frequencies in all the NPM lineages for the case of GC-poor amino acids.

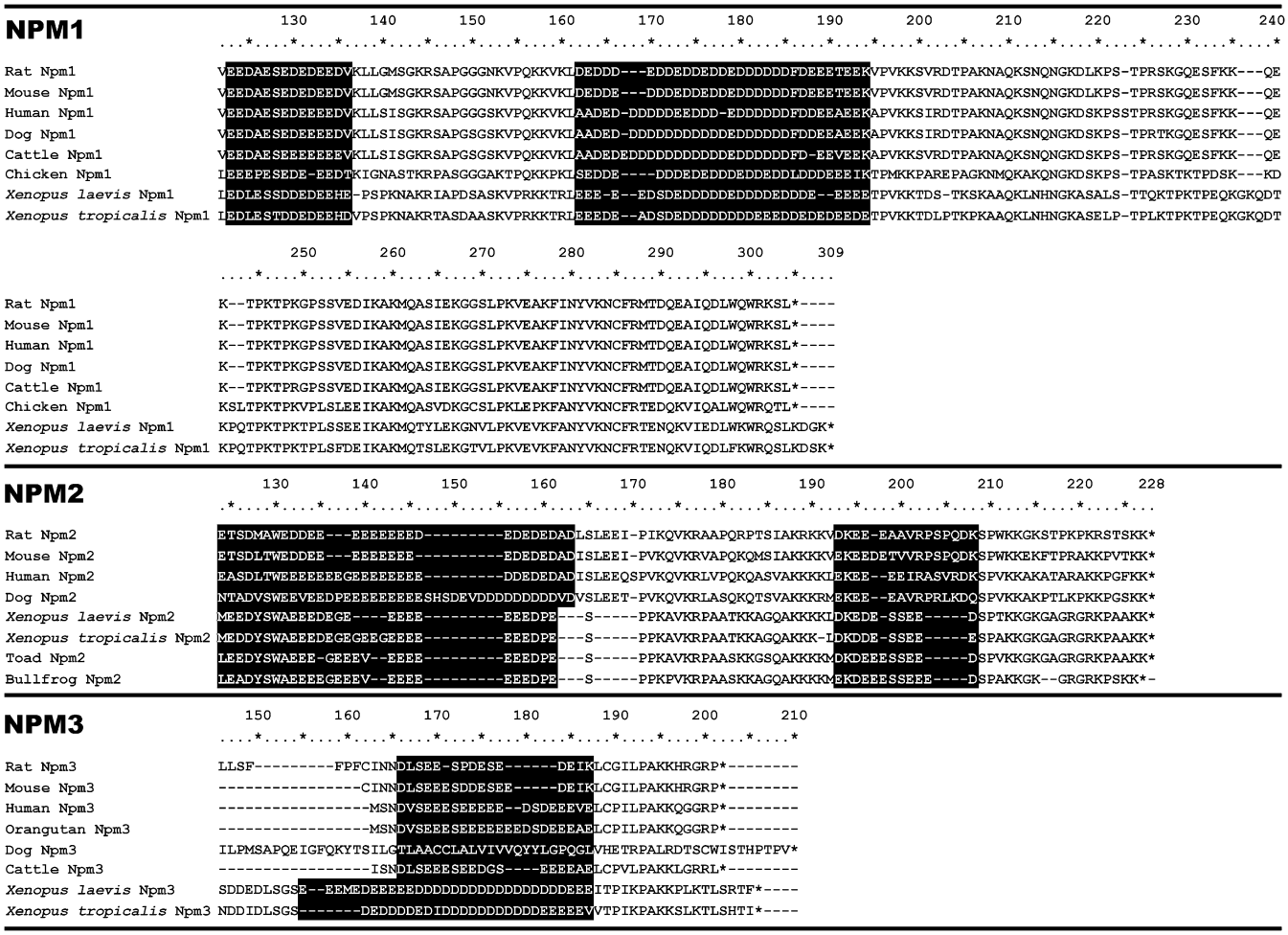

FIGURE 4.—Amino acid alignments corresponding to the C-terminal segment of NPM1, NPM2, and NPM3 in different representative species showing the characteristic acidic tracts (solid boxes).

Codons for lysine possess adenine at the second positions, as do the codons for glutamic acid and aspartic acid, which are the major components of the acidic tracts present in the C-terminal domains of NPM lineages. To further assess the effect of selection on these amino acids, the nucleotide frequencies for adenine at the second codon positions were compared with those presented by fourfold degenerate positions. The adenine content at second codon positions was significantly higher than the observed frequency for this nucleotide at fourfold degenerated positions (Table 3). The differences were most evident in the case of the C-terminal domain of NPMs, where the acidic tracts (rich in glutamic and aspartic acids) are present. These results suggest that selection influences nucleotide composition, especially at the C-terminal regions, maintaining high frequencies of acidic residues.

NPM codon usage bias and phosphorylation: The presence of functional constraints at the protein level allows for a large extent of silent variation in the nucleotide sequences, resulting in a subsequent decrease in the codon bias exhibited by NPM genes. As shown in Figure 6, the overall ENC for NPM genes ranges from 49.415 ± 9.708 (for NPM-like genes of invertebrates) to 54.200 ± 4.222 (for NPM2). When discriminating between the different protein domains, the C-terminal region displays a trend that is slightly more biased than the N-terminal core, with the exception of NPM-like proteins of invertebrates. This quite unexpected observation may be related to the presence of more phosphorylatable residues at the C-terminal segments of NPM proteins, which have been shown to be critical for their correct structure and interaction with histones (DUTTA *et al.* 2001; ARNAN *et al.* 2003; PRADO *et al.* 2004). The predicted phosphorylation sites and the codon usage for each residue of NPM proteins are detailed in Figure 7. Most of the potential phosphorylation targets occur at the N- or C-terminal segments, in agreement with the proposed role for this post-translational modification in unfolding this highly negatively charged segment of the molecule (BAÑUELOS *et al.* 2003; PRADO *et al.* 2004).

Few residues are predicted to be phosphorylated across all taxa at the same position but many positions

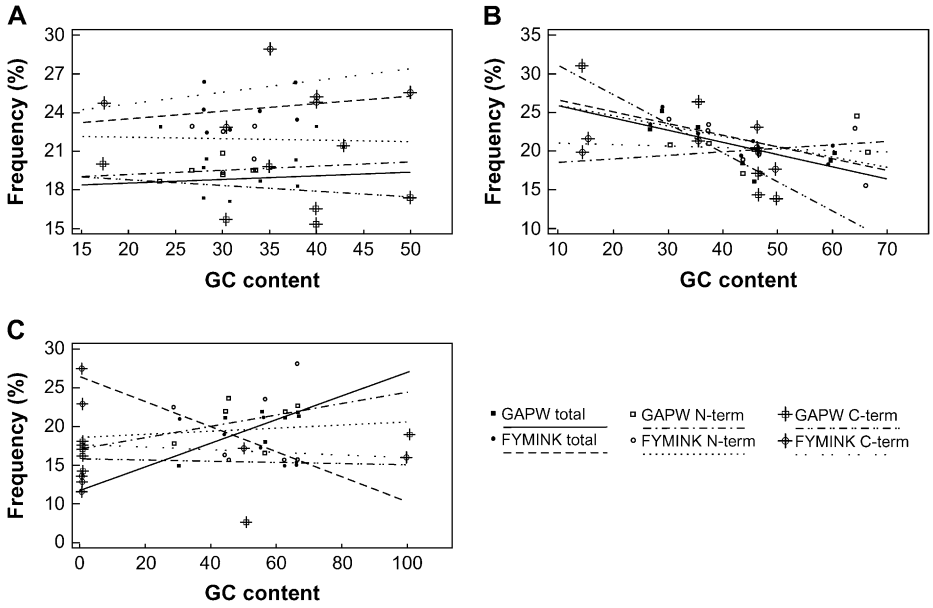

FIGURE 5.—Relationship between GC content and the frequencies of GC-rich (GAPW) and GC-poor (FYMINK) amino acid classes in NPM1 (A), NPM2 (B), and NPM3 (C) discriminating between the complete proteins, the N-terminal domains, and the C-terminal domains.

are conserved potential targets of phosphorylation when the groups of mammals, birds, and amphibians are individually considered (see supplementary alignments 1–4 at http://www.genetics.org/supplemental/). With NPM1, all predicted phosphorylatable tyrosines are encoded by the codon TAT, and codons TCT and ACA are also used to encode most of the serines and threonines amenable to phosphorylation. The analysis of the codon usage for these three residues in NPM1 genes show that they are significantly preferred over the remaining synonymous triplets ($*P < 0.05$, Fisher's LSD; see supplemental Table 4 at http://www.genetics.org/supplemental/).

In the case of NPM2 and NPM3, the differences in codon usage observed between amphibians and mammals extend beyond their amino acid primary structure as shown by the protein phylogeny, resulting in the usage of different preferred codons. In mammals, AGC is the preferred codon of most of the putative phosphorylatable serines of NPM2 and NPM3, whereas TCA and TCT codons are the preferred ones in the case of amphibians. By contrast, the phosphorylatable threonines are primarily encoded by ACC in amphibian NPM2 and NPM3 proteins, while the codons ACC and ACA are preferentially used in mammals. Finally, AGC and ACC are also the preferred codons for serine and threonine, respectively, for the NPM-like proteins from invertebrates (Figure 7; supplemental Tables 5–7 at http://www.genetics.org/supplemental/).

DISCUSSION

Despite the critical functional roles of the members of the nucleophosmin/nucleoplasmin family in cell metabolism and early development of metazoan organisms, there is a lack of information regarding the mechanisms involved in the evolution of these proteins. This is mainly due to the limited amount of sequence data available until very recently. The work presented here represents a first attempt to characterize the mechanisms behind the evolutionary change of the NPM genes.

An evolutionary overview of the NPM family: The protein phylogeny shown in Figure 1 reveals that NPM protein evolution is subject to strong functional and structural constraints leading to the diversification of the family members to accomplish a broad variety of functions in the cell nucleus, as evidenced from the clustering by type of different NPM proteins from different taxa. These types of analyses have proven to be very useful in assessing the paralogous relationship that exists between amphibian NO29 and NO38 proteins and mammalian NPM3 and NPM1 members, respectively, in agreement with what had been previously suggested by SHACKLEFORD *et al.* (2001). Furthermore, as a consequence of the functional diversification exhibited by these proteins along their evolution, the protein phylogenies have allowed us to clearly determine the type (NPM1, -2, or -3) of several previously unassigned sequences.

As shown by the phylogeny in Figure 1, NPM types 1 and 3 are the most closely related. Both of these types, in contrast to NPM2, are ubiquitously expressed in many tissues (SHACKLEFORD *et al.* 2001) and localize to the nucleolus (HUANG *et al.* 2005). In addition, NPM1 and NPM3 form a complex in the cell and through this interaction NPM3 may help regulate NPM1's role in ribosome biogenesis (HUANG *et al.* 2005). This suggests the potential for shared functional constraints between NPM1 and NPM3.

The phylogeny indicates that the NPM1 lineage was the one to most recently appear in the evolution of

TABLE 2

Correlations between genomic GC content and the frequency of GC-rich (GAPW) and GC-poor (FYMINK) amino acids in complete NPM proteins and discriminating between N-terminal and C-terminal domains

NPM lineage	Spearman rank correlation coefficient, r_S	*P*-value
NPM1	Complete	
GC fourfold *vs.* GAPW	0.018	0.839
GC fourfold *vs.* FYMINK	−0.216	0.726
	N terminus	
GC fourfold *vs.* GAPW	0.131	0.569
GC fourfold *vs.* FYMINK	0.000	0.922
	C terminus	
GC fourfold *vs.* GAPW	0.036	0.671
GC fourfold *vs.* FYMINK	0.668	0.454
NPM2	Complete	
GC fourfold *vs.* GAPW	−0.571	0.095*
GC fourfold *vs.* FYMINK	−0.786	0.027**
	N terminus	
GC fourfold *vs.* GAPW	−0.147	0.582
GC fourfold *vs.* FYMINK	−0.720	0.156
	C terminus	
GC fourfold *vs.* GAPW	−0.927	0.009***
GC fourfold *vs.* FYMINK	−0.299	0.688
NPM3	Complete	
GC fourfold *vs.* GAPW	0.559	0.046**
GC fourfold *vs.* FYMINK	−0.667	0.039**
	N terminus	
GC fourfold *vs.* GAPW	0.404	0.396
GC fourfold *vs.* FYMINK	−0.118	0.909
	C terminus	
GC fourfold *vs.* GAPW	0.134	0.875
GC fourfold *vs.* FYMINK	0.134	0.811

The asterisks indicate significance levels as $*P < 0.1$, $**P < 0.05$, $***P < 0.01$.

TABLE 3

Differences between the frequencies of the nucleotide adenine at second codon positions *vs.* the frequencies at fourfold degenerate positions in the functional domains of NPM genes

NPM lineage	*t* statistic	*P*-value
NPM1		
N terminus	−5.343	1.032×10^{-4}
C terminus	3.263	5.658×10^{-3}
NPM2		
N terminus	1.012	3.315×10^{-1}
C terminus	3.566	3.879×10^{-3}
NPM3		
N terminus	2.615	2.259×10^{-2}
C terminus	2.859	1.431×10^{-2}

NPMs. In contrast, NPM2 seems to have been the earliest lineage, a conclusion that is further supported by the higher amino acid variation among taxa within this type. While the role played by NPM2 in histone storage in the oocyte is common for most organisms, the sperm chromatin remodeling function played after fertilization is not as obvious, mainly in the case of amphibians lacking protamines in their sperm (Frehlick *et al.* 2006). A possibility would involve the acquisition of this sperm chromatin remodeling function following the differentiation of specific sperm nuclear basic proteins. Although further studies are needed to clearly trace this process, the functional diversification presented by NPM2 along vertebrate evolution is in good agreement with the notion of its early origin. In contrast to the widespread expression of NPM1 and -3 across different tissue types, NPM2 in both mammals and amphibians is only expressed in oocytes and eggs. Thus, NPM2 may have differentiated early to fulfill the critical role of storing histones in oocytes and eggs.

Invertebrate NPM-like proteins seem to deviate from the differentiation/diversification process observed in vertebrate NPMs as they still retain a monophyletic origin, exhibiting some extent of similarities with the NPM1 lineage. This suggests that they are following a different evolutionary path that is also functionally constrained.

An intriguing feature of the phylogenetic protein trees is the polyphyletic origin of NPM2 and NPM3 proteins, where mammals apparently constitute a well-differentiated group. The polyphyletic origin for NPM2

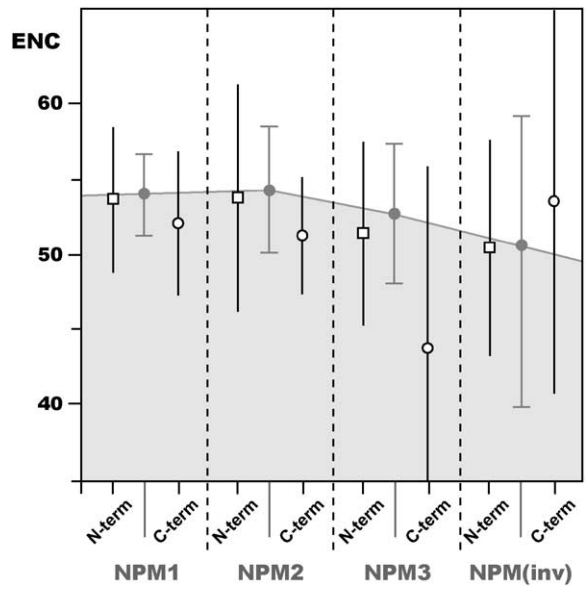

Figure 6.—Codon usage bias in NPM genes. The effective number of codons (ENC) was calculated for complete NPM genes (solid circles) and discriminating between the N-terminal (open boxes) and the C-terminal (open circles) structural domains. Values are given as averages and standard deviations are represented by bars.

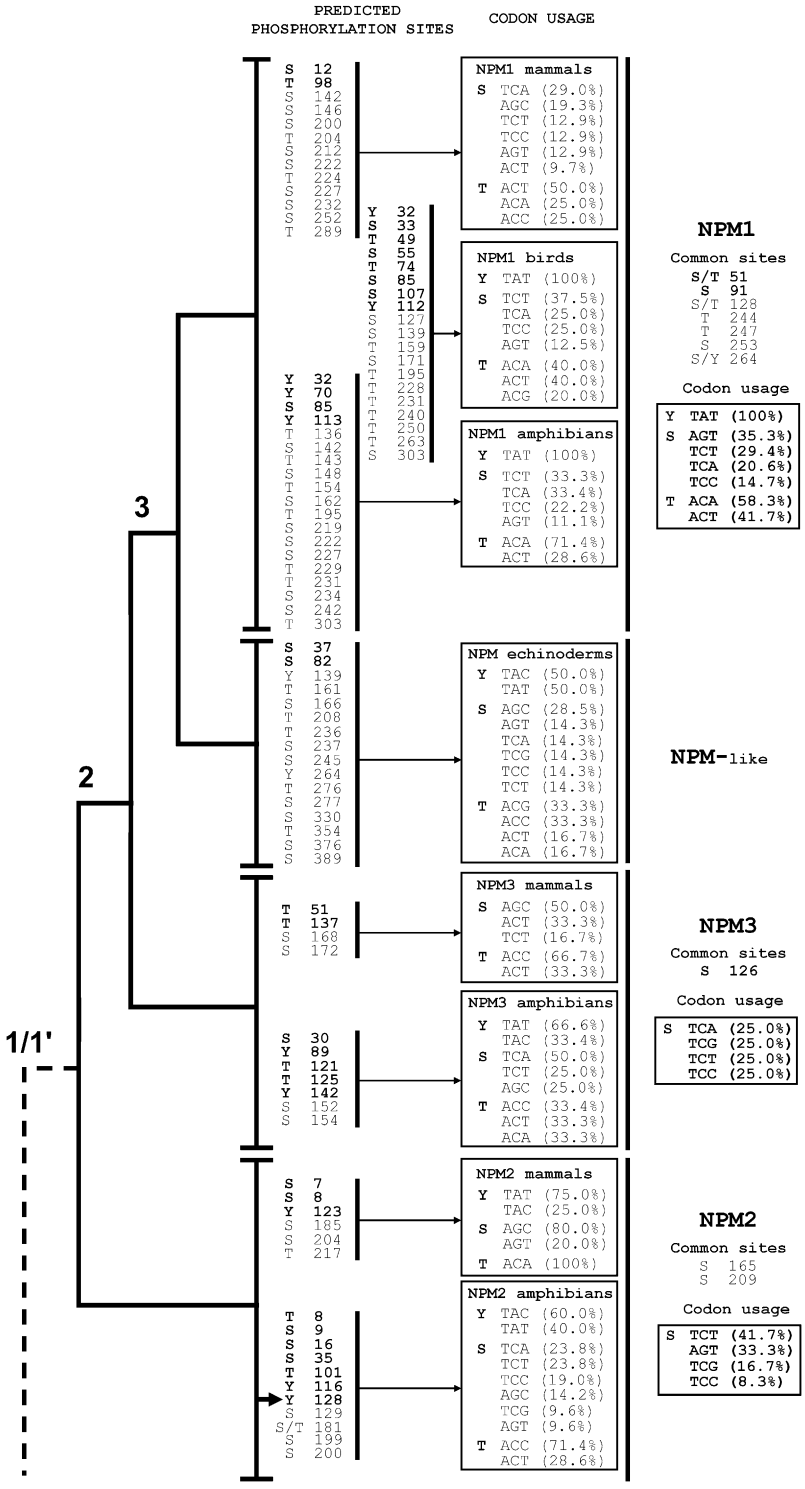

FIGURE 7.—Predicted phosphorylation sites in NPM proteins. The potential phosphorylation sites at serine (S), threonine (T), and tyrosine (Y) residues (threshold >0.5) were predicted and the corresponding codon usage at these positions for NPM proteins in different taxonomic groups residues was calculated (see supplemental alignments 1–4 at http://www.genetics.org/supplemental/). Common sites indicate those positions predicted to be phosphorylated in all species within a given NPM subtype. Sites shown in boldface type highlight positions located at N-terminal regions of the molecules. The Y128 position in NPM2, whose dephosphorylation has recently been linked to chromatin condensation during apoptosis in Xenopus (LU *et al.* 2005), is indicated by an arrow. The branching pattern and the numbering of the nodes at the left denote the evolutionary relationships among NPM proteins described in Figure 1.

and NPM3 is further supported by the maximum-parsimony reconstruction, discarding a possible dependency on the neighbor-joining method in obtaining this result. Furthermore, the codon usage shows different trends between mammals and the rest of organisms in the case of the polyphyletic lineages, which is mainly determined by the differentiation observed between mammals and amphibians. This phenomenon raises very important questions regarding the evolution of NPM2 and NPM3 that, as referred in FREHLICK et al (2006), seems to play different roles in mammals compared to that played in amphibians. More specifically, the sperm chromatin remodeling function of NPM2 is not necessarily critical in mammals, as NPM2-null mice still show normal progression of the paternal chromatin remodeling immediately after fertilization (BURNS *et al.*

2003). Otherwise, in amphibians such as *X. laevis*, NPM2 is required for this initial step. In addition, the down-regulation of NPM3 expression in mammalian oocytes prohibits paternal chromatin condensation, suggesting that NPM3 and not NPM2 may be the family member required for decondensation of the paternal chromatin in mammals (McLay and Clarke 2003). These observations hint at a potential difference between the roles fulfilled by mammalian and amphibian NPM2 and NPM3 proteins, at least during early embryonic development, which may account for the observed polyphyly. Further studies on NPM2 and NPM3 structure and function in mammals and other organisms will be needed to better understand the reason behind the different observed functions and separate polyphyletic origins.

Although less time appears to have elapsed since the appearance of the NPM3 lineage, it shows a similar vertebrate class-dependent differentiation to that observed for NPM2. NPM3 may have differentiated following the loss of the nucleic acid binding domain found in the C terminus of NPM1, which may allow NPM3 to regulate ribosome biogenesis by regulating the binding of NPM1 to RNA (Huang *et al.* 2005).

Functional constraints and NPM structure: NPM genes diverge extensively and significantly through silent substitutions (Table 1; supplemental Tables 2 and 3 at http://www.genetics.org/supplemental/), indicating that their long-term evolution is subject to a strong purifying selection process that operates at the protein level. Such a mode of evolution is also common for somatic histones and SNBPs, with the exception of protamines, which experienced a shift toward an adaptive selection process (Eirín-López *et al.* 2004; Nei and Rooney 2005; Eirín-López *et al.* 2006a,b). The NPM N-terminal core region represents the most conserved domain of the protein and contributes importantly to the structural identity of each NPM type. While the synonymous variation in the N terminus of the molecule is as high as in the less-constrained C-terminal region, the nonsynonymous variation is reduced in the N-terminal regions. The higher extent of nonsynonymous variation at the C-terminal end is probably due to the low selectivity for the occurrence of either aspartic or glutamic acid in the acidic tracts at C-terminal regions, as either residue will equally maintain the overall negative charge of the molecule, which is required for the function of this domain. The selective pressure on the maintenance of a correct protein structure can be further ascertained from the highly conserved pentameric organization of NPM1 and NPM2, which is the structural conformation required for binding to core histones (Dutta *et al.* 2001; Arnan *et al.* 2003; Namboodiri *et al.* 2003, 2004; Prado *et al.* 2004). Such quaternary structural arrangement is ultimately determined by the tertiary organization (folding) of residues constituting the N-terminal part of the NPM monomers, which contrasts with the intrinsically disordered (Hansen *et al.* 2005) C-terminal region (Hierro *et al.* 2001; Prieto *et al.* 2002).

While the NPM2 protein N-terminal folded regions are subject to critical structural constraints, the post-translational phosphorylation of serine, threonine, and tyrosine residues at the N-terminal exposed regions of the molecule and at the C-terminal tail are critical for the interaction of this protein with other chromosomal proteins and for chromatin remodeling (Dutta *et al.* 2001; Arnan *et al.* 2003; Prado *et al.* 2004). The predicted phosphorylation sites shown in Figure 7 and in the supplemental alignments 1–4 (http://www.genetics.org/supplemental/) reveal that only a few residues are potentially phosphorylated in all species at a given position for all NPM types. It appears that the phosphorylation sites in each of the three NPM lineages are specific either for mammals or for all the other vertebrate species, a fact that reflects the divergence of these molecules during the transition to mammals. The functional constraint imposed by the phosphorylation of these residues is evidenced not only by the conservation of the positions they occupy at the protein level, but also by the usage of preferred triplets, which ultimately resulted in the observed codon bias. Interestingly, the preferred codons used by mammals are also different from those preferentially used in other vertebrate classes. Hence, the nucleotide combination used to encode preferentially phosphorylatable residues appears to be taxa specific, which contrasts with the functional clustering of NPM types independent of the species to which they belong.

Furthermore, there is a codon usage heterogeneity within amphibians in organisms with different types of SNBPs. In amphibians, the nature of NPM2 appears to be independent of the SNBP type, as amphibians with SNBPs of the protamine (P) type (Bufo), protamine-like (PL) type (Xenopus), and histone (H) type (Rana) all have NPM2 proteins with similar primary structures (Frehlick *et al.* 2006). However, at the nucleotide level each of these amphibians use different codons in most of their potentially phosphorylatable NPM2 residues. An important exception is that of tyrosine 128 (see supplemental alignment 2 at http://www.genetics.org/supplemental/), which is encoded in all four amphibian sequences by a TAC codon. The biased codon usage conservation at this position must underlie a critically important functional constraint in the NPM2 molecule. Indeed, dephosphorylation of this residue has been shown to occur during apoptosis in Xenopus egg extracts and has been shown to regulate chromatin condensation mediated by changes in the interaction of NPM2 with chromatin and the loss of its chromatin decondensation activity (Lu *et al.* 2005).

While the massive synonymous nucleotide variation among NPM genes may result in a decrease of the codon usage bias, the overall effect of a preferred codon usage

by phosphorylatable amino acids results in the higher codon bias that is observed at the C-terminal regions (where most of this post-translational modification takes place). Although this explanation is valid for the case of NPM1 and NPM-like lineages, further analyses are necessary in the case of NPM2 and NPM3, since the number of predicted phosphorylation sites is similar in both (N-terminal and C-terminal) protein regions. In addition, the presence of acidic tracts composed of either aspartic acid or glutamic acid residues within these regions could contribute to enhance the codon bias displayed by NPM1 and NPM3, where these two amino acids are encoded by preferred codons.

Selection for biased amino acid frequency at C-terminal regions of NPM proteins: Under the neutral model, a protein's amino acid content is influenced by the nucleotide composition of its corresponding gene (Kimura 1983). However, the effect of selection at the protein level could alter nucleotide composition bias if it was strong enough. This was shown to occur in the case of mammalian protamine 1, in which the maintenance of arginine (essential for the DNA-binding function) at high levels can be related to the primary function of the protein, attesting to its importance (Rooney *et al.* 2000). A similar effect of selection for high frequencies of alanine and lysine has been described in the *tolA* gene from proteobacteria, in which a bias in the C + A nucleotide composition of the gene has resulted due to selection for high levels of alanine and lysine (Rooney 2003).

In the case of NPM proteins, comparisons between genomic GC content and the frequency of GC-rich and GC-poor amino acids reveal a deviation from neutrality, as no significant correlations are detected in most cases (Figure 5, Table 2). As can be ascertained from the compositional analysis of NPM proteins, it seems that high levels of lysine and, in addition, glutamic and aspartic acid, are required for the correct structure and function of NPM members. Given that codons for lysine, glutamic acid, and aspartic acid invariably show the presence of an adenine at second positions (Table 3), the frequency of this base was compared between second codon positions and fourfold degenerated positions. Our results reveal significantly higher frequencies in the case of the second codon positions, which run counter to the expectations of the neutral model (Rooney *et al.* 2000). The greatest differences were detected in the case of C-terminal regions, suggesting the presence of selection maintaining acidic residues at high frequencies at these domains, which are critical for the interactions of NPM proteins with other molecules through their acidic tracts. In contrast to protamines, where the overall arginine content is maintained at high levels regardless of position in the proteins (Rooney *et al.* 2000), the acidic tracts of NPMs are located at specific positions in C-terminal regions, indicating that in this case the underlying selection mechanisms operate under constraints determined by both the length and the position of these regions in the NPM molecules.

Long-term evolution of the NPM family: The long-term evolution of the different members of the NPM family follows a process involving a strong purifying selection acting at the protein level, evidenced from the functional diversification of the different NPM proteins along their evolution and from the extensive silent divergence observed at the nucleotide level. The functional and structural constraints exhibited by the N-terminal core region of these proteins are determined by a primary structure that contributes to the structural identity of each NPM type. In the case of the C-terminal domains of these proteins, the constraints transcend from the protein level to the nucleotide level, as revealed by the codon usage bias observed for the acidic residues and potentially phosphorylatable residues of these regions and by the high level of adenine at second codon positions determined by the selection for high frequencies of acidic amino acids. Thus, the amino acid code involved in the post-translational modifications (phosphorylation) of NPMs represents one of the main selective constraints in these proteins, maintained not only at the protein level but also by a preferred codon usage at these positions.

Notably, the long-term evolution of NPM proteins and histones (for which at least NPM1 and NPM2 operate as chaperones) exhibit some common trends. Both groups of proteins are under a strong purifying selection at the protein primary structure level, with an extensive silent variation at the nucleotide level and the presence of a code in the post-translational modifications (yet to be established in NPMs) referred to as histone code in the case of histones (Strahl and Allis 2000). A very exciting possibility would involve a process of parallel evolution between these two groups of interacting proteins resulting in the functional diversification in some of the NPM lineages. However, more functional and structural data are still needed before this hypothesis can be tested.

We thank two anonymous reviewers and Shozo Yokoyama for useful comments on the early version of this manuscript. This work was supported by grants from Natural Sciences and Engineering Research Council of Canada (NSERC), grant number OGP 0046399 (to J.A.), by a Postdoctoral Marie Curie International Fellowship within the 6th European Community Framework Programme (to J.M.E.-L.), and by NSERC PGS-D fellowship (to L.F.).

LITERATURE CITED

Arnan, C., N. Saperas, C. Prieto, M. Chiva and J. Ausió, 2003 Interaction of nucleoplasmin with core histones. J. Biol. Chem. **278:** 31319–31324.

Bañuelos, S., A. Hierro, J. M. Arizmendi, G. Montoya, A. Prado *et al.*, 2003 Activation mechanism of the nuclear chaperone nucleoplasmin: role of the core domain. J. Mol. Biol. **334:** 585–593.

Blom, N., S. Gammeltoft and S. Brunak, 1999 Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. **294:** 1351–1362.

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

Burns, K. H., M. M. Viveiros, Y. Ren, P. Wang, F. J. DeMayo *et al.*, 2003 Roles of NPM2 in chromatin and nucleolar organization in oocytes and embryos. Science **300:** 633–636.

Chan, W. Y., Q. R. Liu, J. Borjigin, H. Busch, O. M. Rennert *et al.*, 1989 Characterization of the cDNA encoding human nucleophosmin and studies of its role in normal and abnormal growth. Biochemistry **28:** 1033–1039.

Dingwall, C., and R. A. Laskey, 1990 Nucleoplasmin: The archetypal molecular chaperone. Sem. Cell Biol. **1:** 11–17.

Dingwall, C., S. M. Dilworth, S. J. Black, S. E. Kearsey, L. S. Cox *et al.*, 1987 Nucleoplasmin cDNA sequence reveals polyglutamic acid tracts and a cluster of sequences homologous to putative nuclear localization signals. EMBO J. **6:** 69–74.

Dumbar, T. S., G. A. Gentry and M. O. Olson, 1989 Interaction of nucleolar phosphoprotein B23 with nucleic acids. Biochemistry **28:** 9495–9501.

Dutta, S., I. V. Akey, C. Dingwall, K. L. Hartman, T. Laue *et al.*, 2001 The crystal structure of the nucleoplasmin-core: implications for histone binding and nucleosome assembly. Mol. Cell **8:** 841–853.

Earnshaw, W. C., B. M. Honda, R. A. Laskey and J. O. Thomas, 1980 Assembly of nucleosomes: the reaction involving *X. laevis* nucleoplasmin. Cell **21:** 373–383.

Eirín-López, J. M., A. M. González-Tizón, A. Martínez and J. Méndez, 2004 Birth-and-death evolution with strong purifying selection in the histone H1 multigene family and the origin of *orphon* H1 genes. Mol. Biol. Evol. **21:** 1992–2003.

Eirín-López, J. M., L. J. Frehlick and J. Ausió, 2006a Protamines, in the footsteps of linker histone evolution. J. Biol. Chem. **281:** 1–4.

Eirín-López, J. M., J. D. Lewis, L. Howe and J. Ausió, 2006b Common phylogenetic origin of protamine-like (PL) proteins and histone H1: evidence from bivalve PL genes. Mol. Biol. Evol. **23:** 1304–1317.

Felsestein, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution Int. J. Org. Evolution **39:** 783–791.

Feuerstein, N., P. K. Chan and J. J. Mond, 1988 Identification of numatrin, the nuclear matrix protein associated with induction of mitogenesis, as the nucleolar protein B23. Implication for the role of the nucleolus in early transduction of mitogenic signals. J. Biol. Chem. **263:** 10608–10612.

Frehlick, L. J., J. M. Eirín-López, E. D. Field, D. F. Hunt and J. Ausió, 2006 The characterization of amphibian nucleoplasmins yields new insight into their role in sperm chromatin remodeling. BMC Genomics **7:** 99.

Hall, T. A., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids Symp. Ser. **41:** 95–98.

Hansen, J., X. Lu, E. Ross and R. Woody, 2005 Intrinsic protein disorder, amino acid composition, and the histone terminal domains. J. Biol. Chem. **281:** 1853–1856.

Herrera, J. E., R. Savkur and M. O. Olson, 1995 The ribonuclease activity of nucleolar protein B23. Nucleic Acids Res. **23:** 3974–3979.

Hierro, A., J. M. Arizmendi, J. De las Rivas, M. A. Urbaneja, A. Prado *et al.*, 2001 Structural and functional properties of *Escherichia coli*-derived nucleoplasmin. Eur. J. Biochem. **268:** 1739–1748.

Huang, N., S. Negi, A. Szebeni and M. O. Olson, 2005 Protein NPM3 interacts with the multifunctional nucleolar protein B23/nucleophosmin and inhibits ribosome biogenesis. J. Biol. Chem. **280:** 5496–5502.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kleinschmidt, J. A., C. Dingwall, G. Maier and W. W. Franke, 1986 Molecular characterization of a karyophilic, histone-binding protein: cDNA cloning, amino acid sequence and expression of nuclear protein N1/N2 of *Xenopus laevis*. EMBO J. **5:** 3547–3552.

Kondo, T., N. Minamino, T. Nagamura-Inoue, M. Matsumoto, T. Taniguchi *et al.*, 1997 Identification and characterization of nucleophosmin/B23/numatrin which binds the anti-oncogenic transcription factor IRF-1 and manifests oncogenic activity. Oncogene **15:** 1275–1281.

Kumar, S., K. Tamura and M. Nei, 2004 MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinform. **5:** 150–163.

Laskey, R. A., B. M. Honda, A. D. Mills and J. T. Finch, 1978 Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. Nature **275:** 416–420.

Leno, G. H., A. D. Mills, A. Philpott and R. A. Laskey, 1996 Hyperphosphorylation of nucleoplasmin facilitates *Xenopus* sperm decondensation at fertilization. J. Biol. Chem. **271:** 7253–7256.

Lu, Z., C. Zhang and Z. Zhai, 2005 Nucleoplasmin regulates chromatin condensation during apoptosis. Proc. Natl. Acad. Sci. USA **102:** 2778–2783.

MacArthur, C. A., and G. M. Shackleford, 1997 *Npm3*: a novel, widely expressed gene encoding a protein related to the molecular chaperones nucleoplasmin and nucleophosmin. Genomics **42:** 137–140.

Maiguel, D. A., L. Jones, D. Chakravarty, C. Yang and F. Carrier, 2004 Nucleophosmin sets a threshold for p53 response to UV radiation. Mol. Cell. Biol. **24:** 3703–3711.

McLay, D. W., and H. J. Clarke, 2003 Remodelling the paternal chromatin at fertilization in mammals. Reproduction **125:** 625–633.

Namboodiri, V. M., S. Dutta, I. V. Akey, J. F. Head and C. W. Akey, 2003 The crystal structure of Drosophila NLP-core provides insight into pentamer formation and histone binding. Structure **11:** 175–186.

Namboodiri, V. M., I. V. Akey, M. S. Schmidt-Zachmann, J. F. Head and C. W. Akey, 2004 The structure and function of Xenopus NO38-core, a histone chaperone in the nucleolus. Structure **12:** 2149–2160.

Nei, M., and S. Kumar, 2000 *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nei, M., and A. P. Rooney, 2005 Concerted and birth-and-death evolution in multigene families. Annu. Rev. Genet. **39:** 121–152.

Ohsumi, K., and C. Katagiri, 1991a Characterization of the ooplasmic factor inducing decondensation of and protamine removal from toad sperm nuclei: involvement of nucleoplasmin. Dev. Biol. **148:** 295–305.

Ohsumi, K., and C. Katagiri, 1991b Occurrence of H1 subtypes specific to pronuclei and cleavage-stage cell nuclei of anuran amphibians. Dev. Biol. **147:** 110–120.

Okuda, M., H. F. Horn, P. Tarapore, Y. Tokuyama, A. G. Smulian *et al.*, 2000 Nucleophosmin/B23 is a target of CDK2/cyclin E in centrosome duplication. Cell **103:** 127–140.

Okuwaki, M., K. Matsumoto, M. Tsujimoto and K. Nagata, 2001 Function of nucleophosmin/B23, a nucleolar acidic protein, as a histone chaperone. FEBS Lett. **506:** 272–276.

Olson, M. O., M. O. Wallace, A. H. Herrera, L. Marshall-Carlson and R. C. Hunt, 1986 Preribosomal ribonucleoprotein partcles are a major component of a nucleolar matrix fraction. Biochemistry **25:** 484–491.

Philpott, A., G. H. Leno and R. A. Laskey, 1991 Sperm decondensation in *Xenopus* egg cytoplasm is mediated by nucleoplasmin. Cell **65:** 569–578.

Philpott, A., T. Krude and R. A. Laskey, 2000 Nuclear chaperones. Semin. Cell Dev. Biol. **11:** 7–14.

Prado, A., I. Ramos, L. J. Frehlick, A. Muga and J. Ausió, 2004 Nucleoplasmin: a nuclear chaperone. Biochem. Cell Biol. **82:** 437–445.

Prieto, C., N. Saperas, C. Arnan, M. H. Hills, X. Wang *et al.*, 2002 Nucleoplasmin interaction with protamines. Involvement of the polyglutamic tract. Biochemistry **41:** 7802–7810.

Ramos, I., A. Prado, R. M. Finn, A. Muga and J. Ausió, 2005 Nucleoplasmin-mediated unfolding of chromatin involves the displacement of linker-associated chromatin proteins. Biochemistry **44:** 8274–8281.

Rice, P., R. Garduno, T. Itoh, C. Katagiri and J. Ausió, 1995 Nucleoplasmin-mediated decondensation of *Mytilus* sperm chromatin. Identification and partial characterization of a nucleoplasmin-like protein with sperm-nuclei decondensing activity in *Mytilus californianus*. Biochemistry **34:** 7563–7568.

Rooney, A. P., 2003 Selection for highly biased amino acid frequency in the TolA cell envelope protein of proteobacteria. J. Mol. Evol. **57:** 731–736.

Rooney, A. P., J. Zhang and M. Nei, 2000 An unusual form of purifying selection in a sperm protein. Mol. Biol. Evol. **17:** 278–283.

Rozas, J., J. C. Sánchez-del Barrio, X. Messeguer and P. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Rzhetsky, A., and M. Nei, 1992 A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. **9:** 945–967.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Schwede, T., J. Kopp, N. Guex and M. C. Peitsch, 2003 SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. **31:** 3381–3385.

Shackleford, G. M., A. Ganguly and C. A. MacArthur, 2001 Cloning, expression and nuclear localization of human *NPM3*, a member of the nucleophosmin/nucleoplasmin family of nuclear chaperones. BMC Genomics **2:** 8.

Sitnikova, T., 1996 Bootstrap method of interior-branch test for phylogenetic trees. Mol. Biol. Evol. **13:** 605–611.

Sitnikova, T., A. Rzhetsky and M. Nei, 1995 Interior-branch and bootstrap tests of phylogenetic trees. Mol. Biol. Evol. **12:** 319–333.

Strahl, B., and C. D. Allis, 2000 The language of covalent histone modifications. Nature **403:** 41–45.

Swaminathan, V., A. H. Kishore, K. K. Febitha and T. K. Kundu, 2005 Human histone chaperone nucleophosmin enhances acetylation-dependent chromatin transcription. Mol. Cell. Biol. **25:** 7534–7545.

Takemura, M., K. Sato, M. Nishio, T. Akiyama, H. Umekawa *et al.*, 1999 Nucleolar protein B23.1 binds to retinoblastoma protein and synergistically stimulates DNA polymerase alpha activity. J. Biochem. **125:** 904–909.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.

Welch, J. E., L. J. Zimmerman, D. R. Joseph and M. G. O'Rand, 1990 Characterization of a sperm-specific nuclear autoantigenic protein. I. Complete sequence and homology with the *Xenopus* protein. Biol. Reprod. **43:** 559–568.

Weng, J. J., and B. Y. Yung, 2005 Nucleophosmin/B23 regulates PCNA promoter through YY1. Biochem. Biophys. Res. Commun. **335:** 826–831.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Wu, M. H., J. H. Chang and B. Y. Yung, 2002 Resistance to UV-induced cell-killing in nucleophosmin/B23 over-expressed NIH 3T3 fibroblasts: enhancement of DNA repair and up-regulation of PCNA in association with nucleophosmin/B23 over-expression. Carcinogenesis **23:** 93–100.

Zhang, J., H. F. Rosenberg and M. Nei, 1998 Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95:** 3708–3713.

Zirwes, R. F., M. S. Schmidt-Zachmann and W. W. Franke, 1997 Identification of a small, very acidic constitutive nucleolar protein (NO29) as a member of the nucleoplasmin family. Proc Natl Acad Sci USA **94:** 11387–11392.

Communicating editor: S. Yokoyama