NON-GTPASE PROTEINS THAT SHARE COMMON MOTIFS WITH G DOMAINS: CONVERGENT OR DIVERGENT EVOLUTION OR DOMAIN RECOMBINATION?

Por

Jorge Hernández-Torres [a*], Alfonso Pineda-Barbosa [a] and Jacques Chomilier [b,c]

Abstract

Hernández-Torres J., A. Pineda-Barbosa & J. Chomilier: Non-gtpase proteins that share common motifs with g domains: Convergent or divergent evolution or domain recombination? Rev. Acad. Colomb. Cienc. **34** (132): 289-299, 2010. ISSN 0370-3908.

GTPases constitute a super class of proteins with a common fold. Five specific G motifs located in loops are signatures of this super class. Nevertheless, some proteins may share the fold of the small GTPases, although their functions are totally unrelated. To retrieve them, we specifically searched in the BLAST output listings for non GTPases with available 3D structure, starting from a canonical GTPase sequence as query. We then performed both a sequence analysis by means of HCA and a structural comparison with an established GTPase. It results that, although sequence identity is in the twilight zone, i.e. below 25%, one can evidence some conservations of the catalytic motifs. Nevertheless, mutations have occurred that produced a new function while the global fold is maintained. We discuss whether non-GTPases presumably originated from a common ancestor with an ancient G domain. The evolutive mechanisms relating non-GTPases to GTPases that we can advance are sequence divergence, convergence and DNA recombination. We conclude that the most probable evolutive pathway leading to such structural similarities is that all the studied proteins must have evolved by sequence divergence from a primordial GTP-binding domain.

Key words: HCA, GTP-binding domain, GTPase, protein evolution, protein folding, hydrophobic packing, sequence identity.

List of abbreviations: HCA, Hydrophobic Cluster Analysis; aa, amino acid(s); rmsd: root mean square deviation; ETK, *E. coli* protein tyrosine kinase.

[a] Laboratorio de Biología Molecular, Escuela de Biología, Universidad Industrial de Santander, Apartado Aéreo 678, Bucaramanga, Colombia.

[b] Protein Structure Prediction, IMPMC, CNRS UMR 7590, Université Paris 6, Paris, France.

[c] RPBS, Batiment Lamarck, 15 rue Hélène Brion, 750013 Paris, France.

[*] Corresponding author: Tel. (57)-7-6349088 Fax (57)-7-6346149. Correo electrónico: hernanj@uis.edu.co

Resumen

Las GTPasas constituyen una superclase de proteínas con un plegamiento común. Cinco motivos G específicos, situados en los bucles, son característicos de esta superclase. Sin embargo, algunas proteínas adoptan el plegamiento de las GTPasas, aunque sus funciones son totalmente diferentes. Para encontrarlas, hemos analizado los resultados de búsquedas BLAST con secuencias canónicas de GTPasas, con el propósito de identificar proteínas no GTPasas con estructura 3D disponible. Posteriormente, procedimos a analizar las secuencias seleccionadas, mediante HCA y la superposición con estructuras de GTPasas de referencia. Los resultados obtenidos indican que aunque la identidad de secuencia se encuentra en la zona crepuscular (*twilight zone*), i.e., por debajo de 25%, se pueden evidenciar algunas conservaciones de los motivos catalíticos. Sin embargo, las mutaciones que se han producido dieron lugar a nuevas funciones, mientras que el plegamiento global se mantiene. Finalmente, discutimos si aquellas proteínas no GTPasas se originaron presumiblemente de un ancestro común con un dominio G antiguo. En tal caso, proponemos como mecanismos evolutivos que vinculan a las GTPasas con las no GTPasas, la divergencia, la convergencia y la recombinación del DNA. Concluimos que el mecanismo evolutivo más probable que dio lugar a tales similaridades estructurales es la divergencia desde un dominio primordial de unión al GTP.

Palabras clave: HCA, dominio de unión al ADN, GTPasa, evolución proteica, plegamiento de proteínas, plegamiento hidrofóbico, identidad de secuencia

1. Introduction

Small G proteins (also called small GTPases, small GTP binding proteins and Ras protein superfamily) comprise a wide variety of proteins that share the same architecture of their GTP-binding domain. Although the basic fold of this globular domain is structurally the same for all members of the family, its primary structure is extremely variable in its amino acid (aa) composition (**Caldon *et al.***, 2001). Typically, the GTP-binding domain (or G domain) is arranged in 5 α-helices (α1- α5) and six β-strands (β1- β6) (**Bourne *et al.***, 1991). Five loops named G1-G5, connecting adjacent strands and helices, contain most of the residues of the active site. Because the small GTPases are involved in diverse cellular processes (cell proliferation, protein synthesis, signal transduction), consensus sequences for the loops G1-G5 are given for each one (**Paduch *et al.***, 2001). These loops are essential for the interaction with the substrate in association with Mg^{2+} and GTP/GDP exchanging factors (**Valencia *et al.***, 1991). GTP hydrolysis "enables different GTPases to sort and amplify transmembrane signals, direct the synthesis and translocation of proteins, guide vesicular traffic through the cytoplasm, and control proliferation and differentiation of animal cells" (**Bourne *et al.***, 1991). However, various members of the GTPase superclass lack their GTPase activity (**Leipe *et al.***, 2002).

GTPases and ATPases are closely related in structure and, in fact, some ATPases constitute a subfamily of the GTPase superclass. The specificity to GTP is conferred by the NKXD motif in the G4 loop. A mutation in these aa leads to the loss of affinity for GTP, decreasing protein specificity for this substrate, but increasing for other ones. For instance, in the group of myosins and kinesins, the GTPase activity was replaced by an ATPase, because of an evolutive change in the G4 loop. However, myosin and kinesin 3D structures superpose well with that of ras proteins and are classified as a subgroup within the GTPase family (**Leipe *et al.***, 2002). The conserved Asp residue in the NKX***D*** motif, providing specificity for GTP, is absent in most of ATPases. Thus, none of ATPases has been shown to have GTP specificity.

The aim of this paper is to provide evidence that a diverse set of proteins roughly presents the same fold as the GTP-binding domain. Nevertheless, they do not have GTPase activity and on the contrary, they exhibit activities such ribonuclease, methyltransferase, guanine deaminase, lactate dehydrogenase and others.

By means of hydrophobic cluster analysis (HCA) and comparison of structures as evidence, we discuss whether non-GTPases presumably derive from a common ancestor with an initial GTP-binding domain. Besides, the evolutive mechanisms concerning sequence divergence, convergence and DNA recombination are also considered.

2. Materials and methods

2.1. Sequence alignment and protein superposition

In order to retrieve related sequences, BLAST is used in a first step, with four proteins known from the literature to be G domains as input data, namely: c-Ha-*ras1*, $G_{i\alpha}$, RhoA and *ORFX* (predicted G domain-like). The output listing of BLAST provides in these four cases, a set of G domains in the top ranked sequences. We discard these first matches because they are all already known to be highly similar and we were interested in the extension of the family. Therefore, we looked at the bottom of the sorted sequences by BLAST and focused at the first sequences not annotated as G domains in the header annotations, i.e., presenting very low identity scores with the query. The retrieved sequences were aligned on short fragments (20-120 aa). In order to produce an automatic extension, alignment on the total length of the presumably unrelated sequences (typically 180 aa) was done with ClustalW online at the Pole BioInformatique Lyonnais (http://npsa-pbil.ibcp.fr/), using default parameters (**Larkin *et al.*** 2007). When comparing sequences in the twiglight zone (Rost, 1999), it is known that automatic procedures do not provide accurate alignments, therefore they were manually cured by means of HCA.

2.2. Hydrophobic Cluster Analysis

HCA plots, sequence identity and HCA score calculations were performed following the indications of (**Callebaut *et al.*,** 1997). Briefly, HCA designs a sequence on the surface of a cylinder with the connectivity of an alpha helix. The 2D planar surface is then duplicated in order to keep local environment for each amino acid, and hydrophobic neighbor residues (VILFMWY) in this plot are then clustered. The shapes of the clusters are keen indication of the nature of the secondary structure. Besides, it has been statistically demonstrated that centers of the hydrophobic clusters correspond to the centers of regular secondary structures (**Woodcock *et al.*,** 1992). HCA identity is calculated by the number of identical aligned hydrophobic and non hydrophobic aa in both sequences to the number of aa of the longest sequence. For each sequence, the HCA score is the ratio between the number of topologically conserved residues between sequences 1 and 2, to the total number of hydrophobic residues in both segments (**Gaboriaud *et al.*,** 1987). A HCA score $\geq 60\%$ is an indicator of high sequence identity.

2.3. Structural comparisons

As evolution has more conserved structures than sequences (**Grishin,** 2001; **Wang *et al.*,** 2008), in a final step, structure superpositions were performed to validate our previous hypotheses. Tertiary structures were all obtained from the PDB (**Berman *et al.*,** 2008) and their superpositions were carried out with the on line programs MATRAS (**Kawabata,** 2003) and CE (**Shindyalov & Bourne,** 1998). Superpositions were performed without forcing the structural alignment to match the sequence alignment derived from HCA. Matras and CE produced similar results. Thus, Matras has been chosen, because we noticed in previous papers that its results are the closest from a consensus of three methods (**Papandreou *et al.,*** 2004).

Except step 2, which is a specialty of the Paris group, the work flow used here is rather classical: an identification of a hot spot in a short stretch of sequences, followed by a structural validation. This has been done for instance on the short chain oxidoreductase enzymes (**Duax *et al.,*** 2007), SH3 fold (**Theobald & Wuttke,** 2005) or on more general sequences (**Liu *et al.*,** 2008). The scheme is the following: first sequences are compared in order to derive evidence for a common ancestry. In case of absence of similarity, structure similarity is search for.

3. Results

3.1 Protein sequence database search

When performing BLAST searches with a GTP-binding domain, it is usual to find non-GTPase proteins that share 50-100 aa long fragments with G domains, with a level of identities located in the "twilight zone" of sequence alignment (**Rost**, 1999). Subsequent superpositions of their corresponding 3D structures yield rmsd lower than 6 Å (**Reva *et al.***, 1998), that can be considered as a reasonable limit to admit that they share a common fold, and therefore are related through evolution. Thus, we were interested to elucidate the evolutive relationships among these proteins and the molecular mechanisms leading to the gain of a similar architecture.

A database screening was carried out with G domain sequences, in order to retrieve non-GTPase proteins with both low 1D identity and partial 3D similarity, and analyze them by HCA analysis and protein superposition. The purpose of such alignments was to elucidate the most conserved catalytic and structural amino acids with the G domain. HCA is a fine method of 2D structural analysis that allows alignments between very distantly related proteins, with as low as 10% sequence identity. HCA analysis does not pay any attention to the strict conservation of the residues inside the clusters but rather to the conserved shapes of the clusters, keeping in mind the underlying idea that shape is a testimony of the secondary structure. As shown as follows, HCA was an outstanding method to reveal structural similarities of proteins like GTPases and non-GTPases, which at first glance do not appear to be related.

BLAST screening of PDB (**Berman *et al.***, 2008) and Blocks (**Henikoff & Henikoff**, 1994) databases was done with the following G domain primary structures as queries: (c-Ha-*ras1* (HRas precursor*,* P01112), Guanine nucleotide-binding protein $G_{i\alpha}$, alpha-1 subunit ($G_{i\alpha}$, P04898) and *ORFX* (AF261774)). *ORFX* is a protein of unknown function from *Lycopersicon esculentum* that is involved in determination of fruit size and seems to have a predicted fold similar to the one of *ras* (**Frary *et al.***, 2000).

Most of the retrieved sequences were GTP-binding domains, with high identity with the query domains. However, a few sequences of non-GTPase proteins exhibited local identities in the range of 10 to 23% with G domains. Then, we selected the first non-GTPase sequences in the list for primary, secondary and tertiary sequence alignments by means of ClustalW, HCA and 3D structure superposition, respectively. We did not find sequences that meet our criteria in other than PDB and Blocks databases. In Table 1, we list the retrieved sequences. HCA identities were always slightly higher than ClustalW alignments, in the range 18 to 30%. This is actually frequent, because HCA produces a more sensitive alignment than a standard program. HCA scores were in the range 55-60%, a good indicator of a significant homology (**Gaboriaud *et al.***, 1987). Low rmsd values are also listed in Table 1, following structure superposition.

To illustrate the use of HCA alignments in combination with protein superpositions and the qualitative and quantitative information we can derive from such alignments, we show an example in Fig. 1 and Table 1. Two pairs of distantly related proteins with low 1D identity, the human c-Ha-*ras1* (a GTPase) with RhoA (a GTPase) (**Gómez del Pulgar *et al.,*** 2005) and c-Ha-*ras1* (a GTPase) with the CMP kinase (an ATPase) (1Q3T), were aligned by HCA. Previously, a 35% aa identity has been evidenced between *rho* genes from *Aplysia* and c-Ha-*ras1* (Madaule and Axel, 1985). As seen in Fig. 1a GTPases share most of the hydrophobic clusters (ID=29%) and 3D structures superpose along the entire domain (rmsd 1.62 Å). However, the GTPase and the ATPase (Fig. 1b), which aroused from a unique primordial fold (**Leipe *et al.***, 2003), share an identity of 12% (increased to 23% by HCA) and partial superposing subdomains: the N-terminal 32 aa (i.e., G1 motif and some aa of G2) produces a rmsd of 5.4 Å and the C-terminal 45 aa (G4 and G5 motifs), a rmsd of 5.1 Å. Interestingly, the G3 loop (the so-called Walker B box (**Walker *et al.***, 1982) or kinase 2 motif) conserves both hydrophobic clusters and identities in the GTPase and the ATPase proteins, although they are located at the end of a β-sheet in c-Ha-*ras1* and a α-helix in CMP kinase. The G3 loop provides residues for Mg^{2+} and γ-phosphate binding and is found in other nucleotide binding motifs, not homologous to small G proteins, present in sugar kinases, ABC transporters and ATP synthases (**Paduch *et al.***, 2001). Thus, despite the degree of divergence between GTPases and related ATPases, structural relationships can be evidenced by a refined alignment method like HCA and supported by structure superposition. Our data confirm the tight structural relation of the GTPase and ATPase, notwithstanding evolutive divergences.

3.2. BLAST searches of non-GTPase proteins with a similar fold to the GTP-binding domain

Proteins with diverse functions non-responsible of neither GTPase nor ATP activities, mostly related to nucleic acid binding, share 50-100 aa long fragments with GTP-binding domains, with identities located in the "twilight zone" of sequence alignment. We show in the subsequent alignments, non-GTPase proteins with sequence identity

Table 1. Sequence identities among pairs of proteins, after domain alignments performed with ClustalW or HCA. The approximate length (in amino acids) along which the alignment is performed is indicated in the aa column. HCA scores were calculated as stated by **Gaboriaud *et al.***, (1987). Rmsd over superposed length values (number of aa) are shown. In most of the cases the superposition is performed on a whole domain, but in two cases, corresponding to 1b and 3a, a hinge fragment is evidenced from the HCA analysis. Consequently, the superpositions have been calculated on the parts surrounding these hinge regions. Uniprot accession numbers are the following: P01112 for c-Ha-Ras; P61186 for RhoA; Q97PK6 for CMP kinase; P04898 for $G_{iá}$; Q57599 for Rnase HII; Q9GT92 for LDH; P55135 for Rum A; O34598 for Guanine deaminase; af261774 for OrfX.

Figure	Aligned proteins	ClustalW	HCA	HCA Score	aa	RMSD (Å)/aa
1a	c-Ha-*Ras1* - RhoA	29	30	68	160	1.62/ 161
1b	c-Ha-*Ras1* - CMP kinase	12	23	59	170	5,4/32; 5,1/45
2a	$G_{iá}$ - RNase HII	14	20	56	170	4,03/79
2b	RhoA - LDH	18	22	62	170	3,02 /78
2c	c-Ha-*Ras1* - RumA	10	18	49	160	3,46/64
3a	c-Ha-*Ras1* - Guanine deaminase	10	18	50	170	3,84/29; 4,27/47
3b	c-Ha-*Ras1* - OrfX	15	17	53	150	NA

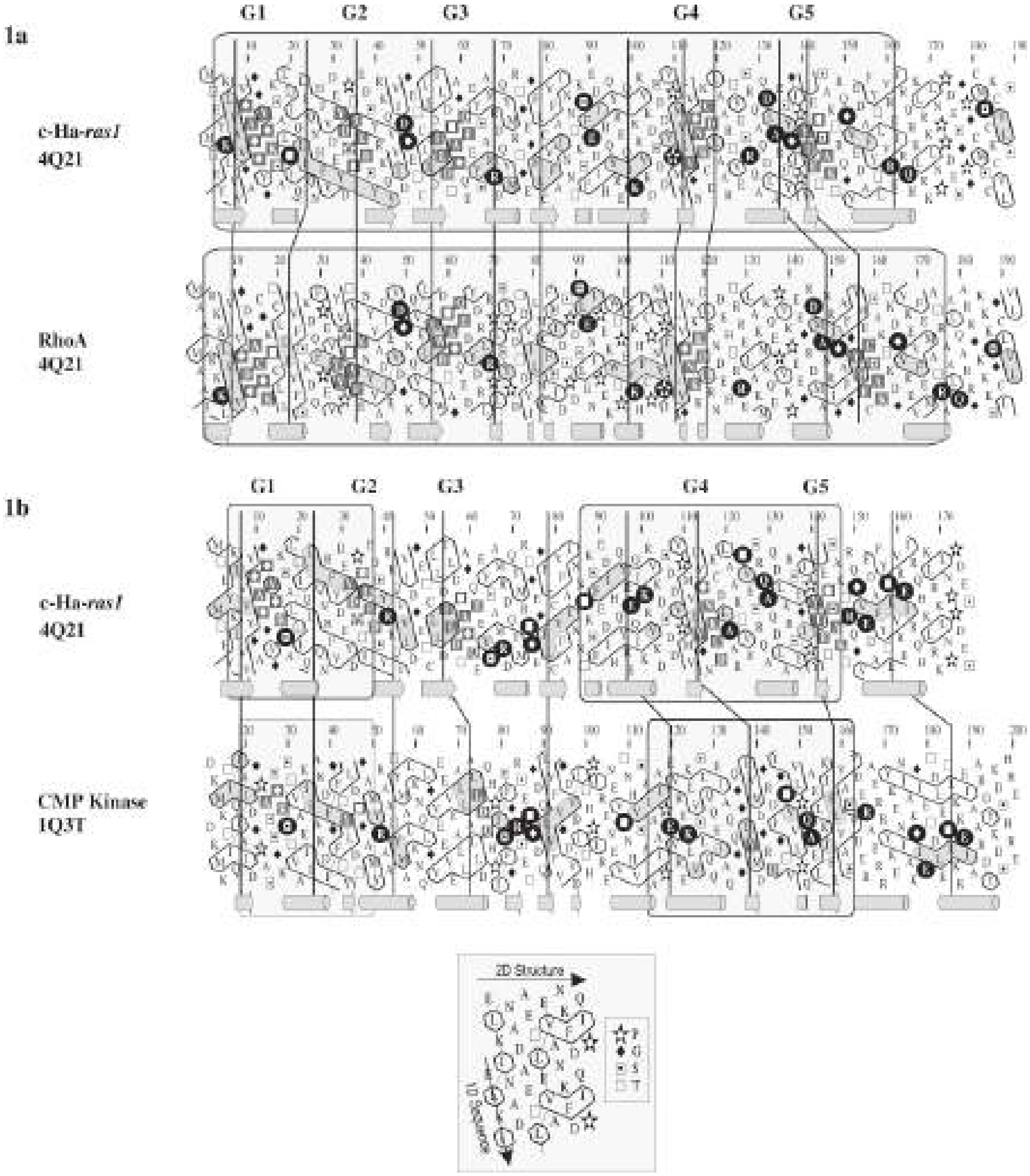

Figure 1. HCA plots of pairs of aligned sequences. a) Two low identity GTPases, c-Ha-*ras1* (PDB code 4Q21) and RhoA (PDB code 1X86_b); b) c-Ha-*ras1* aligned with the CMP kinase (PDB code 1Q3T, ATPase). **Symbols**: Conserved hydrophobic clusters are grey shaded. Non hydrophobic identities are indicated by white letters inside black circles. White letters on a grey square indicate catalytic aa in G motifs (G1-G5 on top) and residue conservation in non-GTPases. Black or white aa in a larger font size (some inside black circles) are residues that interact with a specific ligand as cofactors or substrates. Under each plot, the secondary structures are schemed, based on PDB files. The boxed regions indicate the fragments of the proteins for which the structures can be superimposed. The onset helps interpreting the HCA plots. Because of the duplication (see methods), sequence is read vertically, one line over two, and the secondary structure is read horizontally, a cluster corresponding statistically to a regular secondary structure. Vertical lines connect the occurrences of analogous clusters.

and rmsd of the same order as for G and ATPase domains previously analyzed, whereof their evolutive relation with G domains has not been so far proposed nor explained.

With the G domain of $G_{i\alpha}$ as a query sequence, known to belong to the G-alpha family, we retrieved by BLAST the Archaeal RNase HII of the hypothermophile *Methanococcus jannaschii* (**Lai *et al.***, 2000), homologous to the human major RNase H. ClustalW and HCA identities were 14 and 20%, respectively. In Fig. 2a we can observe that the most relevant conserved hydrophobic clusters are mainly associated to α-helices, which are rather long for RNase HII. Structure superposition gave a rmsd of 4.03 Å over 79 aa at the C-terminal end. A striking observation in Fig. 2a is that the most conserved secondary structures and catalytic residues are in boundary of G4 and G5 loops, and G4 is the motif that determines the interaction with GTP or ATP (see introduction). Because RNase HII interacts with a polymer of nucleotides (*E. coli* RNase HII cleaves the RNA strand of a RNA-DNA hybrid, endonucleolytically at the P-O3' bond), one can speculate that the motifs responsible for nucleotide interaction and hydrolysis have been rearranged to perform RNA degradation.

When using RhoA as a query sequence, we obtained the lactate dehydrogenase (LDH) from *Cryptosporidium parvum*. Sequence identities were 18 and 22% for ClustalW and HCA alignments, respectively (Fig. 2b). After superposition of the two domains, we obtained a rmsd of 3.02 Å over 78 aa in the C-terminal region, including the G4 and G5 motifs. The LDH structure (PDB code 2FM3) is complexed with substrate (pyruvic acid) and cofactor (NADH). The aa involved in the interaction with the ligand are located around G1-G5 motifs. For example, residues QI (positions 14-15) are in the G1 motif of RhoA, D35 in G2, ItN (120-122) (conserved residues in capital letters) in G4 and MagV (145-148) in G5. Thus the similar distribution of hydrophobic clusters, the low rmsd and the coincident positions occupied by catalytic residues once more allow to conclude to the existence of a partial common fold for the two proteins. Hence, the aa pertinent for interaction with GTP have evolved to be able to bind the dinucleotide coenzyme NADH.

With *ORFX* (see below) as BLAST query, we retrieved in Blocks database the Block number IPB010280C (motifs G4-G5) of *Thermococcus kodakarensis* RumA protein (Q5JHF7). In *E. coli,* this enzyme that catalyzes the transfer of a methyl group from S-adenosylmethionine (SAM), specifically to uridine 1939 of 23S rRNA to yield 5-methyluridine (**Lee *et al.***, 2004). We show in Fig. 2c a HCA alignment between *E. coli* RumA and c-Ha-*ras1*, not *ORFX,* because of the absence of PDB structures; however *ORFX* is predicted to superpose well with c-Ha-*ras1* since their Z scores for global and local alignments are high (3.2 and 4, respectively) (**Frary *et al.***, 2000). Identities are 10% and 18% with ClustalW and HCA, respectively. The major cluster and secondary structure conservations are located, like in RNase HII and LDH, at the C-terminal end (Fig. 2c). We were able to superpose these regions, obtaining a rmsd of 3.46 Å over 64 aa. Lee et al. locate the putative site of interaction of RumA with SAM to residues 268-282 (AGV…EWL, Fig. 2c) (**Lee *et al.***, 2004). These positions perfectly match with G1 and G2 motifs in c-Ha-*ras1*.

With c-Ha-*ras1* as a query sequence, we retrieved the *Bacillus subtilis* Guanine Deaminase, an enzyme that catalyzes the hydrolytic deamination of guanine into xanthine (**Liaw *et al.***, 2004) (Fig. 3a). Sequence identities are 10 and 18%. The high conservation of secondary structures and their direct relation with hydrophobic clusters in number and shape are obvious in Fig. 3a. A protein superposition was possible in two fragments (Fig. 3a, Table 1) yielding a rmsd of 3.87 Å over 28 aa (N-terminal) and 4.27Å over 55 aa (middle of the protein). 1-2 aa in each G1-G4 motifs are shared between the two sequences.

The last sequence we analyzed is *ORFX* (AF261774), one of the most intriguing proteins with a probable G-domain fold. It is a protein of unknown function, but involved in determination of fruit size in tomato (**Frary *et al.***, 2000) and classified in Pfam as a member of the PLAC8 family (Placenta-specific gene 8 protein). *ORFX* was predicted in the literature as sharing a similar fold (LOOPP program (**Tobi & Elber,** 2000) than the human oncogene *ras* protein (PDB code 6Q21). We show in Fig. 3b a HCA alignment of c-Ha-*ras1* and *ORFX*. A striking observation is the high correspondence of hydrophobic clusters and aa conservation of the G motifs, excepting G3, with a HCA identity of 17% (Table I). Unfortunately there is no available 3D structure for protein superposition. Because of its assumed regulatory activity, we tried to associate *ORFX* with transcription factors. HCA analysis shows that *ORFX* is enriched in cysteines, a common characteristic of zinc finger proteins; thus, we aligned *ORFX* with members of the zinc finger family and we show an example with Atzfp6 (Swiss-Prot entry Q39265) in Fig. 3c. Interestingly, the numerous cysteines of *ORFX* are in the same domain organization than zinc finger proteins (ZF1 and ZF2 in Fig. 3c). A prediction of zinc-binding cysteines with Predzinc (**Shu *et al.*,** 2008) revealed four candidate residues with probability scores exceptionally high to constitute a putative zinc-binding domain (ZF 2, Fig. 3c): Asp111 (0.576), Cys118 (0.742), Cys121 (0.742) and Cys124 (0.656). Besides,

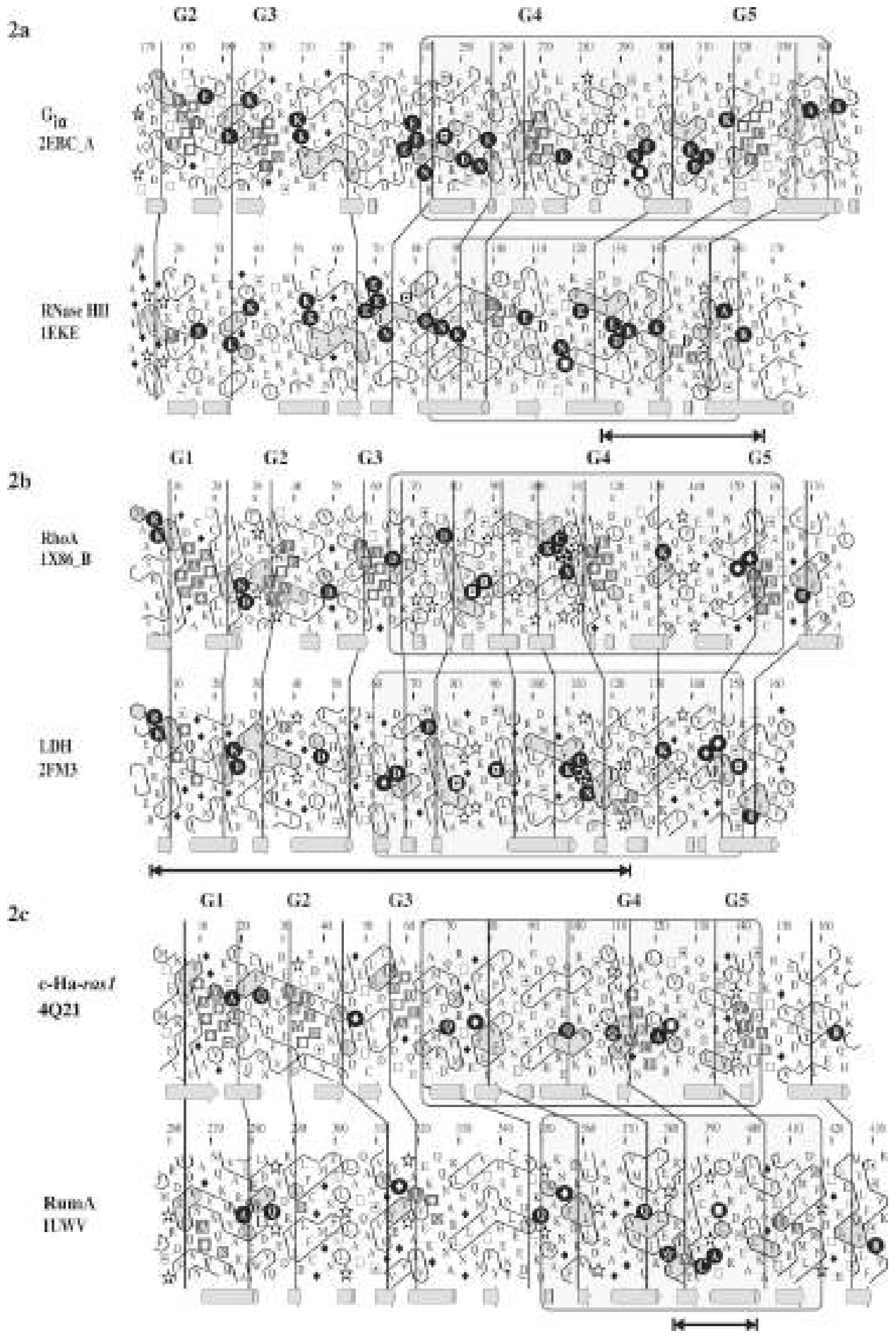

Figure 2. HCA plots of pairs of aligned sequences. a) $G_{i\acute{a}}$ (PDB code 2EBC, GTPase) aligned with RNase HII (PDB code 1EKE, non GTPase); b) RhoA (PDB code 1X86_B, GTPase), aligned with LDH (PDB code 2FM3, non GTPase); c) c-Ha-ras1 (PDB code 4Q21, GTPase), RumA (PDB code 1UWV, non GTPase). The HSSP producing hits with BLAST are marked by a black line under the alignments.

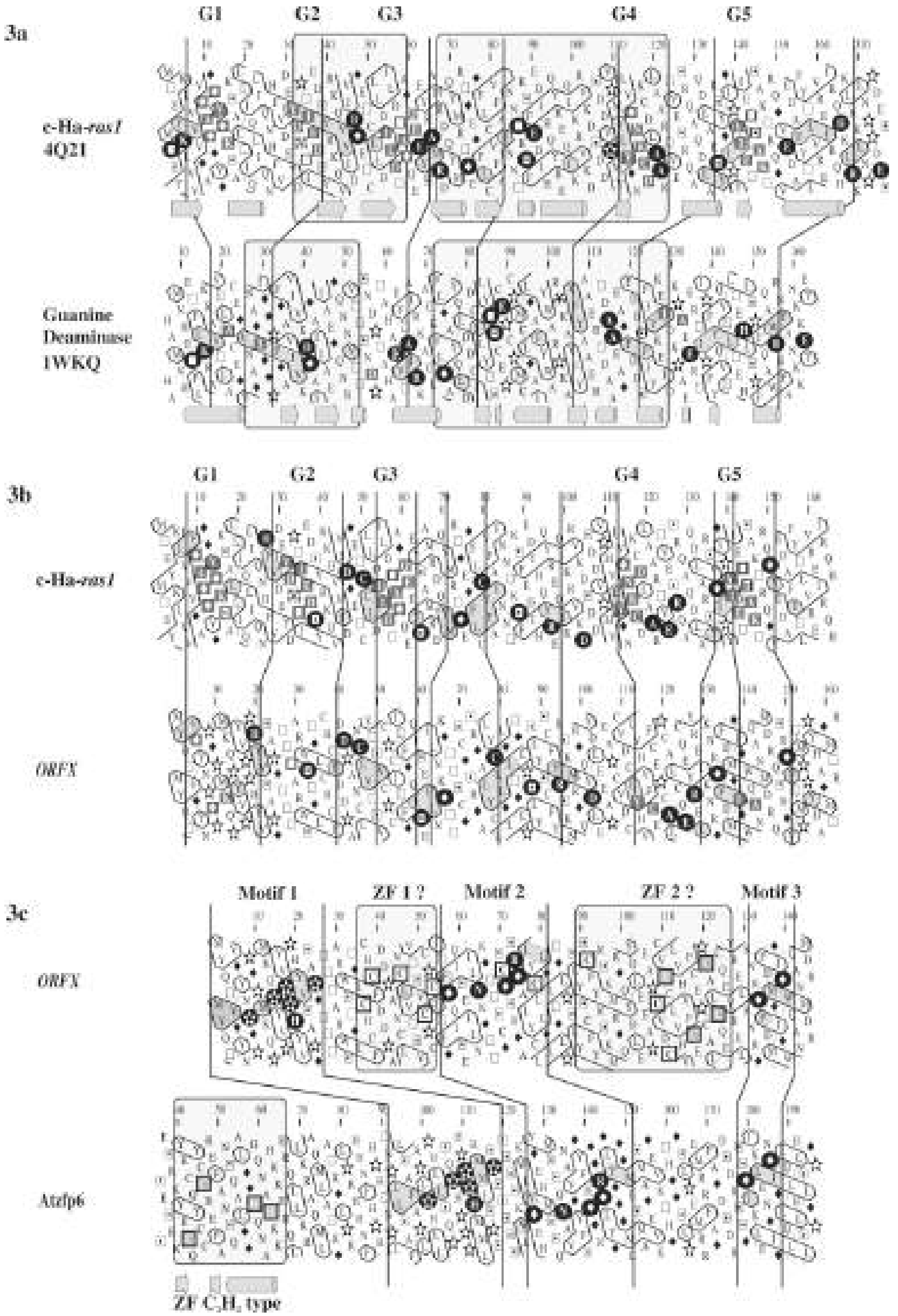

Figure 3. HCA plots of pairs of aligned sequences. a) c-Ha-ras1 (PDB code 4Q21, GTPase) aligned with guanine deaminase (PDB code 1WKQ, non GTPase); b) c-Ha-*ras1* aligned with *ORFX* c) *ORFX* (accession number af261774) aligned with Atzfp6 (Q39265).

we can observe in Fig. 3c the remarkable residue conservation of hydrophobic (clusters) and non-hydrophobic residues for motifs 1 (23% identity), 2 (20% identity) and 3. These results enable us to propose that *ORFX* may originate from a primordial G domain and belong to an unknown class of zinc finger proteins with transcription regulatory functions.

In order to supply a graphic view of our hypothesis of inclusion of new members in the G domain family, we show in Fig. 4 structural superpositions or RhoA with LDH on top, and c-Has-ras1 superposed on RumA at the bottom. Global rmsd in these two cases is 3,02 Å and 3,46 Å, respectively, therefore validating our assumption of two new members in the family.

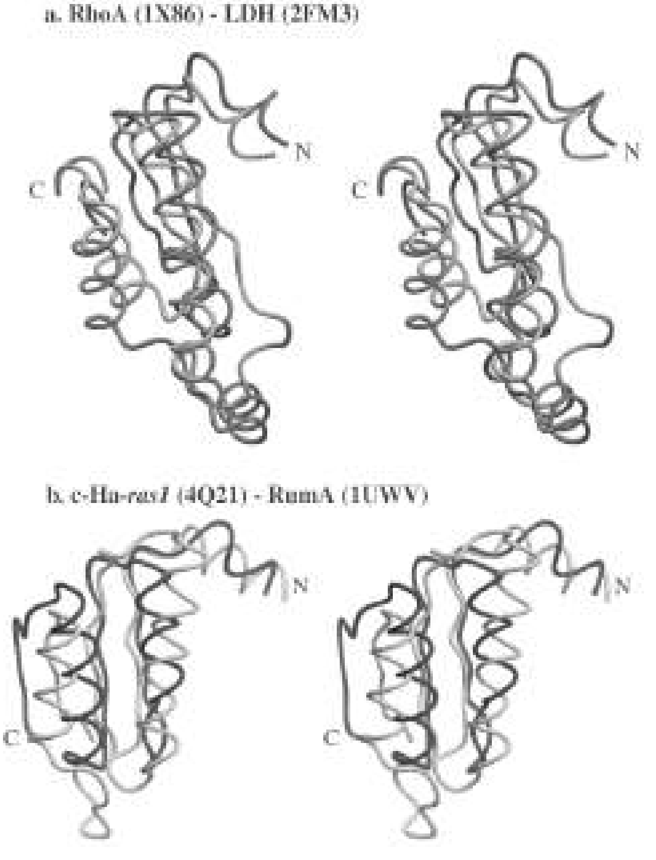
a. RhoA (1X86) - LDH (2FM3)
N
C
N
C
b. c-Ha-*ras1* (4Q21) - RumA (1UWV)
N
C
N
C

Figure 4. Superposed stereo structures of a protein known to belong to the G domain (RhoA (P^{71}-G^{155}) on top and c-Has-ras1 (R^{68}-S^{144}) at the bottom) with a proposed new member of the family (LDH (D^{84}-C^{161}) on top and RumA (Q^{350}-L^{414}) at the bottom.

4. Discussion

The phylogenetic tree of GTPase and ATPase families is now well established and in numerous cases they originate from a common nucleotide hydrolytic domain (**Leipe *et al.***, 2002; **Leipe *et al.***, 2003). However, some proteins share very low identity with the other members of the superclass and are difficult to classify in this family by means of sequence alignments. Structure superposition is of great help to produce a reliable alignment. A decisive criterion to incorporate a putative GTPase in the family is the conservation of the G1 to G5 motifs. Even though G domains may be highly divergent, they are still recognizable by the appropriate methods (Fig. 1).

In this work, we provide some arguments for the existence of unrelated proteins, as far as biological function is concerned, that share a series of secondary structures with the G domain of small GTPases. Following the functional annotation of proteins in databases, all compared polypeptides in Table 1 constitute a metabolically distinct supra family, i.e., "homologous enzymes that catalyze mechanistically distinct reactions in different metabolic pathways and have conserved active-site residues that perform different functions in different members of the supra family" (**Gerlt & Babbitt**, 2000). Three ways by which non-GTPase proteins may have acquired related structures with G domains are convergence, divergence and partial gene recombination. Convergent and divergent evolution may be difficult to distinguish. However, it could be feasible to consider that non-GTPase proteins of Table 1 are related to a common ancestor of the present GTP-binding domain. One can advance the following arguments: i) there are significant similarities in hydrophobic cluster positions and shapes, sequence identities and structural similarities between superimposed subdomains, leading to the conclusion that their analogous fold is not accidental; ii) functional amino acids of non GTPase proteins fit with the G1-G5 motifs. **Blouin *et al.*** (2004) found that "the sequence variability among these homolog proteins (28 GTP-binding domains) is directly linked to the structural variability of surface loops" and that "these regions are self-contained and thus mostly free of the evolutionary constraints imposed by the conserved core of the domain" (**Blouin *et al.***, 2004). Therefore, it is possible that most of the adaptations of G domains to new functions are possible because of the structural flexibility of the G1-G5 loops; the best example of such adaptation is the lactate dehydrogenase (LDH) from *C. parvum* (Fig. 2b). iii) As seen in Fig. 1a, the G domains may be highly divergent within the family (HCA identity of 30%); a structural related ATPase (Fig. 1b) yielded a lower score (23% by HCA). Interestingly, non-GTPase proteins exhibited similar identities ranging from 18% to 23% and significant rmsd. All these data allow to conclude to the existence of a common ancestor for all these proteins; iv) interestingly, the substrates for most of these proteins are nucleotides or polymers of nucleotides; thus, the structural affinity for nucleotides is maintained but with novel activities. v) Previously, we have demonstrated that a GTP-binding domain was recycled to have a receptor activity in the A domain of the chloroplast GTPase receptor Toc159 (**Hernández Torres *et al.***, 2007). This is one example of adaptation of the G domain to a new function, independent of nucleotide binding. A second case of a GTPase domain adaptation is *ORFX*. As seen in Fig. 3b, the alignment between ras and *ORFX* proteins is well consistent in hydrophobic clusters, as well as in non-hydrophobic aa (17% identity). In this work we propose that *ORFX* has a function of a zinc-finger transcription factor and we present strong evidence in Fig. 3c. It is no clear how a nucleotide hydrolyzing domain becomes a transcription factor; however its primary nucleotide binding activity and the flexibility of loop regions must have been the key for this domain transformation (**Blouin *et al.***, 2004).

A second way by which non-GTPase proteins may have appeared is convergent molecular evolution (CME). There are few reports demonstrating this evolutive event, somehow controversial. CME is defined as “the independent evolution of similar nucleotide or amino acid sequences in two unrelated molecules” (**Zakon,** 2002). Structural convergence is one of the four classes of CME and could be the most suitable for non-GTPase proteins. As stated by Zakon (2002), “in some cases, molecules with very different amino acid sequences can assume similar structural motifs, and these may carry out similar functions”. If it was the case for non-GTPase proteins, the convergence would be only in the 3D structure but not in the functional level. The best models of reported CME refer to different structures with the same function (see for example the hexokinase, ribokinase, and galactokinase families of sugar kinases (**Bork *et al.*,** 1993). There are two points that refute a case of CME for non GTPases: i) convergence would occur only for structure but not for function and ii) all the aligned proteins display conserved functional residues with GTPase G1-G5 motifs (Figs. 2 and 3). It would be difficult to accept all those analogous residues as a product of chance, particularly if each protein exhibits a different metabolic function.

The third probable mechanism at the origin of non-GTPases could be the partial recombination between ancient genes and a primordial GTP-binding domain, in order to produce new genes which gained only some of G1-G5 motifs. However, this alternative would be less probable than convergence, because the high conservation of hydrophobic and identity clusters, as well as functional residues.

Taking together all results, we conclude that the most feasible mechanism by which non-GTPases hold reminiscences of a common ancestral GTP-binding domain is divergence. Then it was simpler for nature to assign new functions to an already existing nucleotide-interacting domain, with a particularly flexible architecture in their catalytic loops, than building new ones from scratch.

Acknowledgements

RPBS bioinformatics resources have been used for this project (http://bioserv.rpbs.univparis-diderot.fr/). The authors are grateful to Dr. Nanjiang Shu for the use of Predzinc program.

References

Berman, HM., Z. Westbrook, FG. Gilliland, TN. Bhat, H. Weissig, IN. Shindyalov & PE. Bourne. 2008. The Protein Data Bank. Nucleic Acids Res. 28: 235-242.

Blouin, C., D. Butt & AJ. Roger. 2004. Rapid evolution in conformational space: a study of loop regions in a ubiquitous GTP binding domain. Protein Sci. 13: 608-616.

Bork, P., Sander, C. & A. Valencia. 1993. Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, and galactokinase families of sugar kinases. Protein Sci. 2: 31-40.

Bourne, HR., DA. Sanders & F. McCornick. 1991. The GTPase superfamily: a conserved structure and molecular mechanism. Nature 349: 117-127.

Caldon, CE., P. Yoong & PE. March. 2001. Evolution of a molecular switch: universal bacterial GTPases regulate ribosome function. Mol. Microbiol. 41: 289-297.

Callebaut, I., G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat & JP. Mornon. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell. Mol. Life Sci. 53: 621-645.

Duax, WL., Huether, R., Pletnev. V., Umland, TC., Weeks, CM. 2007. Divergent evolution of a specific protein fold and identification of its oldest surviving ancestor. Biotechnology and Bioinformatics Symposium, Paper ID:50.

Frary, A., TC. Nesbitt, A. Frary, S. Grandillo, E. Van der Knaap, B. Cong, J. Liu, J. Meller, R. Elber, KB. Alpert & SD. Tanksley. 2000. *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. Science 289: 85-88.

Gaboriaud, C., V. Bissery, T. Benchetrit & JP. Mornon. 1987. Hydrophobic cluster analysis: an efficient new way to compare and analyze amino acid sequences. FEBS Lett. 224: 149-155.

Gerlt, JA. & PC. Babbitt. 2000. Can sequence determine function? Genome Biol. 1: 1-10.

Gómez del pulgar, T., SA. Benitah, PF. Valerón, C. Espina & JC. Lacal. 2005. Rho GTPase expression in tumourigenesis: evidence for a significant link. Bioessays 27: 602-613.

Grishin, N. 2001.Fold change in evolution of protein structures. J Struct Biol 134:167-185.

Henikoff, S. & JG. Henikoff. 1994. Protein family classification based on searching a database of blocks. Genomics 19: 97-107.

Hernández Torres, J., MA. Maldonado Arias & J. Chomilier. 2007. Tandem duplications of a degenerated GTP-binding domain at the origin of GTPase receptor Toc159 and thylakoidal SRP. Biochem. Biophys. Res. Commun. 364: 325-331.

Kawabata, T. 2003. MATRAS: a program for protein 3D structure comparison. Nucleic Acids Res. 31: 3367-3369.

Lai, L., H. Yokota, LW. Hung, R. Kim & SH. Kim. 2000. Crystal structure of archaeal RNase HII: a homologue of human major RNase H. Structure 8: 897-904.

Lee, TT., S. Agarwalla & RM. Stroud. 2004. Crystal structure of RumA, an iron-sulfur cluster containing *E. coli* ribosomal RNA 5-methyluridine methyltransferase. Structure 12: 397-407.

Larkin MA., Blackshields G., Brown NP., Chenna R., McGettigan PA., McWilliam H., Valentin F., Wallace IM., Wilm A., Lopez R., Thompson JD., Gibson TJ. & Higgins DG. 2007. ClustalW and ClustalX version 2. Bioinformatics 23: 2947-2948.

Leipe, DD., EV. Koonin & L. Aravind. 2003. Evolution and classification of P-loop kinases and related proteins. J. Mol. Biol. 333: 781-815.

________, Y. I. Wolf, EV. Koonin & L. Aravind. 2002. Classification and evolution of P-loop GTPases and related ATPases. J. Mol. Biol. 317: 41-72.

Liaw, SH., YJ. Chang, CT. Lai, HC. Chang & GG. Chang. 2004. Crystal structure of *Bacillus subtilis* guanine deaminase. J. Biol. Chem. 279: 35479-35485.

Liu, Z-P., Wu, LY., Wang, Y., Zhang, XS., Chen, L. 2008. Bridging protein local structures and protein functions. Amino acids. 35: 627-650.

Madaule, P. & R. Axel. 1985. A novel ras-related gene family. Cell 41: 31-40.

Paduch, M., F. Jelen, & J. Otlewski. 2001. Structure of small G proteins and their regulators. Acta Biochim. Pol. 48: 829-850.

Papandreou, N., Eliopoulos, E., Berezovsky, I., Lopes, A., Chomilier, J. 2004. Universal positions in globular proteins : observation to simulation. Eur. J. Biochem. 271: 4762-4768.

Reva, BA., AV. Finkelstein & J. Skolnick. 1998. What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? Fold. Des. 3: 141-147.

Rost, B. 1999. Twilight zone of protein sequence alignments. Protein Eng. 12: 85-94.

Shindyalov, IN. & PE. Bourne. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11: 739-747.

Shu, M., Zhou, T. & S. Hovmöller. 2008. Prediction of zinc-binding sites in proteins from sequence. Bioinformatics 24: 775-782.

Theobald, D., Wuttke, D. 2005. Divergent evolution within protein superfolds inferred from profile based phylogenetics. J. Mol. Biol. 354: 722-737.

Tobi, D. & R. Elber. 2000. Distance dependent, pair potential for protein folding: results from linear optimization. Proteins 41: 40-46.

Valencia, A., M. Kjeldgaard, EF. Pai & C. Sander. 1991. GTPase domains of Ras p21 oncogene protein and elongation factor Tu: analysis of three-dimensional structures, sequence families, and functional sites. Proc. Natl. Acad. Sci. U.S.A. 88: 5443-5447.

Walker, JE., M. Saraste, MJ. Runswick & NJ. Gay. 1982. Distantly related sequences in the a and b-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. 1: 945-951.

Wang, L., Qiu, Y., Wang, J., Zhang, X. 2008. Recongnition of structure similarities in proteins. Jrl Syst Sci & Complexity. **28**:665-675.

Woodcock, S., JP. Mornon & Henrissat B. 1992. Detection of secondary structure elements in proteins by Hydrophobic Cluster Analysis. Protein Eng. 5: 629-635.

Zakon, HH. (2002) Convergent evolution on the molecular level. Brain Behav. Evol. 59: 261.

Recibido: octubre 20 de 2009

Aceptado para su publicación: septiembre 12 de 2010