NON-GTPase PROTEINS WITH DIVERSE FUNCTIONS SHARE A COMMON FOLD WITH THE GTP-BINDING DOMAIN, AS REVEALED BY HYDROPHOBIC CLUSTER ANALYSIS

ALFONSO PINEDA BARBOSA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGIA
2008

NON-GTPase PROTEINS WITH DIVERSE FUNCTIONS SHARE A COMMON FOLD WITH THE GTP-BINDING DOMAIN, AS REVEALED BY HYDROPHOBIC CLUSTER ANALYSIS

ALFONSO PINEDA BARBOSA

Trabajo de investigación para optar el titulo de Biólogo

DIRECTOR:
Jorge Hernández Torres
PhD Ciencias

CODIRECTOR:
Jacques Chomilier
PhD Física

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS ESCUELA DE BIOLOGIA
2008

CONTENT

LIST OF FIGURES

Pág.

LIST OF TABLES

Pág.

RESUMEN

TITULO: RELACIÓN FUNCIONAL Y ESTRUCTURAL DE PROTEÍNAS ORIGINADAS A PARTIR DE LA DEGENERACIÓN DE UN PRIMIGENIO DOMINIO G, MEDIANTE EL ANÁLISIS DE AGREGADOS HIDROFÓBICOS.*

AUTOR: ALFONSO PINEDA BARBOSA.**

PALABRAS CLAVE: HCA, dominio de unión a GTP, GTPasa, Evolución de proteínas, estructura proteínica, empaquetamiento hidrofóbico, similitud de secuencia.

RESUMEN:
Las pequeñas proteínas G (también llamadas pequeñas GTPasas, pequeñas proteínas de unión a GTP y superfamilia de proteínas Ras) comprende una amplia variedad de proteínas que comparten la misma arquitectura que el dominio de unión a GTP. Aunque la estructura básica de este dominio globular es estructuralmente la misma para todos los miembros de la familia, su primaria estructura es extremadamente variable en su composición de aminoácidos (aa). Debido a que las pequeñas GTPasas están involucradas en diversos procesos celulares (proliferación celular, síntesis de proteínas, transducción de señales), existen secuencias consenso para los bucles G1-G5 para cada uno de los miembros. Estos bucles son esenciales para la interacción con el sustrato en asociación con Mg2+ y factores de intercambio GTP/GDP. La hidrólisis de GTP permite que diferentes GTPAsas ordenen y amplifiquen señales transmembranales, direccionen la síntesis y traslocación de proteínas, guíen el tráfico vesicular a través del citoplasma, y controlen la proliferación y diferenciación de células animales. Sin embargo, miembros de la superclase GTPasa carecen de la actividad GTPasa. La especificidad por el GTP es concedida por el motivo NKXD del bucle G4. Una mutación en aquellos aminoácidos conlleva a la pérdida de la afinidad por GTP, afectando la especificidad por este sustrato, pero ganando especificidad por otros. Este documento se centra en proteínas con funciones no relacionadas, pero con la estructura de las pequeñas GTPasas. Utilizando, como resultado de búsquedas en BLAST, secuencias de proteínas, especialmente no GTPasas con estructura 3D disponible, este trabajo presenta análisis de secuencia hechos por HCA y comparación estructural con GTPasas caracterizadas. Resultó que, aunque la identidad de secuencia está en la zona umbral, i.e. inferior al 25%, se puede evidenciar conservación de la localización de los motivos catalíticos. Es así que, mutaciones ocurridas produjeron nuevas funciones mientras la estructura global se mantiene.

---

* Trabajo de Investigación.

** Facultad de Ciencias. Escuela de Biología. Director: Jorge Hernández Torres. Codirector: Jacques Chomilier.

ABSTRACT

TITLE: NON-GTPase PROTEINS WITH DIVERSE FUNCTIONS SHARE A COMMON FOLD WITH THE GTP-BINDING DOMAIN, AS REVEALED BY HYDROPHOBIC CLUSTER ANALYSIS.*

AUTOR: ALFONSO PINEDA BARBOSA.**

KEYWORDS*:* HCA, GTP-binding domain, GTPase, protein evolution, protein folding, hydrophobic packing, sequence similarity

ABSTRACT:
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 in the 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. 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 of their members. These loops are essential for the interaction with the substrate in association with Mg2+ and GTP/GDP exchanging factors. 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". However, members of the GTPase superclass lack their GTPase activity. 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, affecting protein specificity for this substrate, but gaining for other ones. This paper focus on proteins with unrelated functions, but with the fold of the small GTPases. By searching in the BLAST output specifically non GTPases with a 3D structure available, this work performed both a sequence analysis by means of HCA and structural comparisons with established GTPases. It results that, although sequence identity is in the twilight zone, i.e. below 25%, one can evidence conservations of the catalytic motif location. Nevertheless, mutations occurred that produce a new function while the global fold is maintained.

---

* Trabajo de Investigación.

** Facultad de Ciencias. Escuela de Biología. Director: Jorge Hernández Torres. Codirector: Jacques Chomilier.

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 in the 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 [1]. Typically, the GTP-binding domain (or G domain) is arranged in 5 α-helices (α1- α5) and six β-strands (β1- β6) [2]. 5 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 of their members [3]. These loops are essential for the interaction with the substrate in association with Mg^{2+} and GTP/GDP exchanging factors [4]. 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" [2]. However, members of the GTPase superclass lack their GTPase activity [5]. 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, affecting protein specificity for this substrate, but gaining 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 [5]. The conserved Asp residue in the NKXD 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 superposition of PDB structures, we demonstrate that all the analyzed proteins presumably derive from a common ancestor and we propose that an initial GTP-binding domain assumed new roles in the cell through evolution, by interacting with nucleotide-containing coenzymes or polymers of nucleotides. To our knowledge, this is the first report of structural relation among GTPases and non-GTPase proteins.

2. Materials and methods

2.1 Hydrophobic Cluster Analysis

HCA plots, sequence identity and HCA score calculations were performed following the indications of Callebaut et al. (1997) [6]. Briefly, HCA designs a sequence on a 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 [7]. 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 [8]. A HCA score ≥60% is an indicator of high sequence identity.

2.2 Sequence alignment and protein superposition

Sequences were retrieved by BLAST from a list of five proteins known from the literature to be G domains, namely: c-Ha-ras, Arf-1A, $G_{i\alpha}$, RhoA and OrfX (predicted G domain-like). Alignment of protein sequences was done with the on line service of ClustalW of the Pole BioInformatique Lyonnais (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html), using default parameters. Tertiary structures were obtained from the RCSB PDB [9]. Structural superpositions were carried out with the on line programs MATRAS [10] and CE [11].

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 fragments of 50-100 aa long with G domains, with identities located in the “twilight zone” of sequence alignment [12]. Protein superpositions of the corresponding 3D structures yield rmsd lower than 6 Å. Therefore, our assumption was that homologous fragments could share a similar architecture, as a result of divergent evolution from an ancestral G domain. To demonstrate this, we started to search for a GTPase sequence with an identity as low as possible with a well characterized GTP-binding domain, and with an available 3D structure. Then, we aligned and superposed the query GTP-binding domains with related ATPases of the GTPase superclass. Our results evidenced that G domains may be highly divergent, but still recognizable by appropriate methods.

A database screening was carried out with G domain sequences, in order to retrieve non-GTPase proteins with similar secondary and tertiary structures, evidenced by HCA analysis and protein superposition, respectively. Sequence capture was done with G domain primary structures, by BLAST searches against PDB [13] and Blocks [14] databases. We do not found sequences that accomplish our criteria in other available databases. 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 alignment by means of ClustalW,

HCA and 3D structure superposition, respectively. Our conclusion is that proteins sharing secondary and local 3D structures with GTP-binding domains might share the same ancestor, instead of structural convergence during protein evolution, option that is subsequently discussed.

The sequences of GTP-binding domains representative of the small GTPases, such as HRas precursor (c-Ha-*ras1,* P01112), ADP-ribosylation factor 1 (Arf-1A, P32889), Guanine nucleotide-binding protein G_i, alpha-1 subunit ($G_{i\alpha}$, P04898) and OrfX (AF261774) were used as query sequences to BLAST searches. 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* protein [15].

In Table 1, we list the retrieved sequences. HCA identities were always slightly higher than ClustalW alignments, with identities from 18 to 30%. This is actually because HCA produces a different alignment than a standard program. HCA scores were in the range 55-60%, a good indicator of homology [8]. Low rmsd values are also listed in Table 1, following structure superposition.

3.2 Sequence identity below 30% with the G domains of two small GTPases

Some GTPase proteins share very low identity with the other members of the superclass. The decisive criterion to incorporate them in GTPase family is the conservation of the G1 to G5 motifs. To demonstrate the high sequence variability among G domains within the members of the GTPase family, we started this work with c-Ha-*ras1*, a human oncogene protein [16], as a query sequence. We

retrieved RhoA (P06749), a tumor associated protein [17], among sequences with very low identity. Previously, it has been evidenced a 35% aa homology of *rho* genes from *Aplysia* with H-ras [18]. We can observe in Fig. 1a that these two proteins conserve the G1 to G5 motifs together with the number and shape of hydrophobic clusters. Sequence identities with ClustalW and HCA were about 30%. The superposition of their 3D structures produces a rmsd of 1.62 Å over 161 aa, indicating a very good match between both domains. We show with this example that the fold of two G domains, with sequence identity located in the twilight zone, may be more conserved than primary structure and that HCA and protein superposition may evidence their structural relationship.

3.3 Sequence identity below 30% within a GTP-binding domain and two ATPases

It has been suggested that protein kinases, such as ETK and small kinases, such as adenosylcobinamide kinase, evolved within the GTPase family [5]. Nevertheless, "structural studies suggest that the best-characterized P-loop kinases, namely nucleotide kinases form a monophyletic lineage distinct from all other P-loop NTPases" [19]. Despite the degree of divergence between GTPases and related ATPases, their structural relationship can be evidenced by a refined alignment method like HCA. BLAST searches in Blocks with c-Ha-*ras1* as input retrieved the CMP Kinase of *Streptococcus pneumoniae.* CMP kinase catalyzes the phosphoryl transfer from ATP to CMP and dCMP, resulting in the formation of nucleotide diphosphates [20]. In Fig. 1b we show a HCA alignment of both proteins and in Table 1 we find that identity increased from 12% to 23% with HCA

alignment. 1 to 4 aa of each G1-G5 motifs are conserved at most, in spite of the fact that tertiary structure superposition of 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. Similar results were obtained by comparing the nucleoside kinase of *Methanocaldococcus jannaschii* (2C49 in the PDB), an Archaeal ATP-dependent ribokinase that phosphorylates inosine, cytidine, guanosine and adenosine [21]. It shares 12% identity (ClustalW) and 17% (HCA) with c-Ha-*ras1*. Only 3 aa of each G1-G5 motif are conserved at most in spite of the fact that tertiary structure superposition of the C-terminal 66 aa (covering both G4 and G5 motifs) produces a rmsd of 3.64 Å, which is a reasonable value for fold comparison (data not shown). Our data confirm the tight structural relation of the GTPase and ATPase from a unique primordial fold, notwithstanding their evolutive divergence at the central or N-terminal end.

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

We show, in the subsequent examples, non-GTPase proteins with low identity, but rmsd of the same order as for G and ATPase domains. With Arf-1A as a query sequence, we found a 22-kDa subunit of human C8, one component of the cytolytic membrane attack complex of complement (MAC) [22]. C8γ belongs to the lipocalin family of small secreted proteins which bind small hydrophobic ligands [23]. As we can see in Table 1, a ClustalW alignment yields an identity of 15% and the HCA analysis increased it to 23%. Protein superposition gave a rmsd of 3.67 Å

over 49 aa, covering the G1 and G2 motifs. Although the rest of the proteins does not superpose well, our HCA analysis reveals that isolated clusters and secondary structures are conserved. As shown in Fig. 2a C8γ shares with the G domain of Arf-1A some aa associated to the G1 (3), G2 (3), G3 (4) and G4 (1) motifs. Interestingly, most of the secondary structures are well conserved between the proteins, especially those close to the motifs involved in GTPase activity and located in catalytic loops. This result is a good indicator of an ancestral relation between the two proteins, in spite of the fact that they exhibit non related activities. With the G domain of $G_{i\alpha}$ as a query sequence, which belongs to the G-alpha family, we found by BLAST an RNase HII from the hypothermophile *Methanococcus jannaschii* [24], the Archaeal RNase HII homologous to human major RNase H. *E. coli* RNase HII cleaves the RNA strand of a RNA-DNA hybrid endonucleolytically at the P-O3′ bond. ClustalW and HCA identities were 14 and 20%, respectively. In Fig. 2b 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. 2b is that the best 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, one can speculate that the motifs responsible for nucleotide interaction and hydrolysis have been rearranged to assume RNA degradation.

When using RhoA as a query sequence, we obtained the lactate dehydrogenase (LDH) from *Cryptosporidium parvum* (Senkovich and Chattopadyay, unpublished). Sequence identities were 18 and 22% for ClustalW and HCA alignments, respectively (Fig. 2c for the HCA alignment). 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 2FM3 structure 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 same position as G1 of RhoA,

D of G2 (35), ItN (conserved residues in capital letters) of G4 (120, 122) and MagV of G5 (145, 148). Hence, the close distribution of hydrophobic clusters, the low rmsd and the coincident position occupied by catalytic residues once more allows to conclude the existence of a unique ancestor for the two proteins. Thus, the aa pertinent for interaction with GTP have evolved to be able to bind the dinucleotide coenzyme NAD^+.

With OrfX (see below) as query in BLAST, we found in Blocks database the RumA protein, an *E. coli* enzyme that catalyzes transfer of a methyl group from S-adenosylmethionine (SAM), specifically to uridine 1939 of 23S rRNA to yield 5-methyluridine [25]. We show in Fig. 3a a HCA alignment between c-Ha-*ras1* (not OrfX because of the absence of a PDB structure, but their HCA alignments will be presented below) and RumA. Identities are 10% and 18% with ClustalW and HCA, respectively. As seen in Fig. 3a, the major cluster and secondary structure

conservations are located, like in RNase HII, and LDH, at the C-terminal end. We were able to superpose these regions, obtaining a rmsd of 3.46 Å over 64 aa. Lee et al., (2004) locate the putative site of interaction of RumA with SAM to residues 268-282 (AGV…EWL, Fig. 3a) [25]. These positions perfectly match with G1 motif in c-Ha-*ras1*, and 3 catalytic aa are shared.

With c-H-*ras1* as a query sequence in a BLAST search, we found the *Bacillus subtilis* Guanine Deaminase, an enzyme that catalyzes the hydrolytic deamination of guanine into xanthine [26] (Fig. 3b). ClustalW and HCA identities were 10% and 18% respectively. The high conservation of secondary structures and their direct relation with hydrophobic clusters in number and shape are obvious in Fig. 3b. A protein superposition was possible in two fragments (Fig. 3b, 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 of 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 [15] and classified in Pfam as a member of the PLAC8 family (Placenta-specific gene 8 protein). OrfX was found in the literature as sharing a very similar fold that the human oncogene *ras* protein (PDB code 6Q21) [15]. Frary et al. found that “the Z scores for global and local alignments of ORFX are high (3.2 and 4, respectively). Such scores were never observed in false positives and suggest an overall shape similar to that of heterotrimeric guanosine triphosphate–binding proteins”. We show in Fig. 3c 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. Unfortunately there is no available 3D structure, but HCA analysis shows that OrfX is enriched in cysteines, in an identical domain organization than zinc finger proteins (data not shown).

4. Discussion

The phylogenetic tree of GTPase and ATPase families is now well established and in numerous cases they originate from a nucleotide hydrolytic domain [5,19] . In this work, we provide some arguments of the existence of unrelated proteins as far as biological function is concerned, which 3D structures have the same fold as 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” [27]. Two ways by which non-GTPase proteins may have acquired related structures with G domains are structural convergence and divergence. Convergent and divergent evolution may be difficult to distinguish. However, it is 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 statistically significant similarities in hydrophobic cluster positions and shapes, non hydrophobic aa, sequence identities and the high structural similarities as derived

from the rmsd between superimposed subdomains, leading to the conclusion that their analogous fold is not accidental; ii) functional amino acids of non GTPase proteins are located in 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” [28]. 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. 2c). 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); however, non-GTPase proteins exhibited higher 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 and iv) interestingly, the substrate for most of these proteins are nucleotides or polymers of nucleotides. Thus, the structural affinity for nucleotides is maintained but with novel activities. 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.

5. Figures and tables

Table 1. Pairs of structurally related proteins with the sequence identity according to the alignment by ClustalW or HCA, and the rmsd and length of aligned structures.

Figure	Aligned proteins	ClustalW	HCA	HCA Score	RMSD/AA
1a	c-Ha-*Ras1* - RhoA	29	30	68	1.62/ 161
1b	c-Ha-*Ras1* - CMP kinase	12	23	59	5,4/32; 5,1/45
2a	Arf-1A - C8γ	15	23	56	3,67 /49
2b	$G_{i\alpha}$ - RNase HII	14	19	56	4,03/79
2c	RhoA - LDH	18	22	62	3,02 /78
3a	c-Ha-*Ras1* - RumA	10	18	49	3,46/64
3b	c-Ha-*Ras1* - Guanine deaminase	10	18	50	3,84/29; 4,27/47
3c	c-Ha-*Ras1* - OrfX	15	19	53	NA

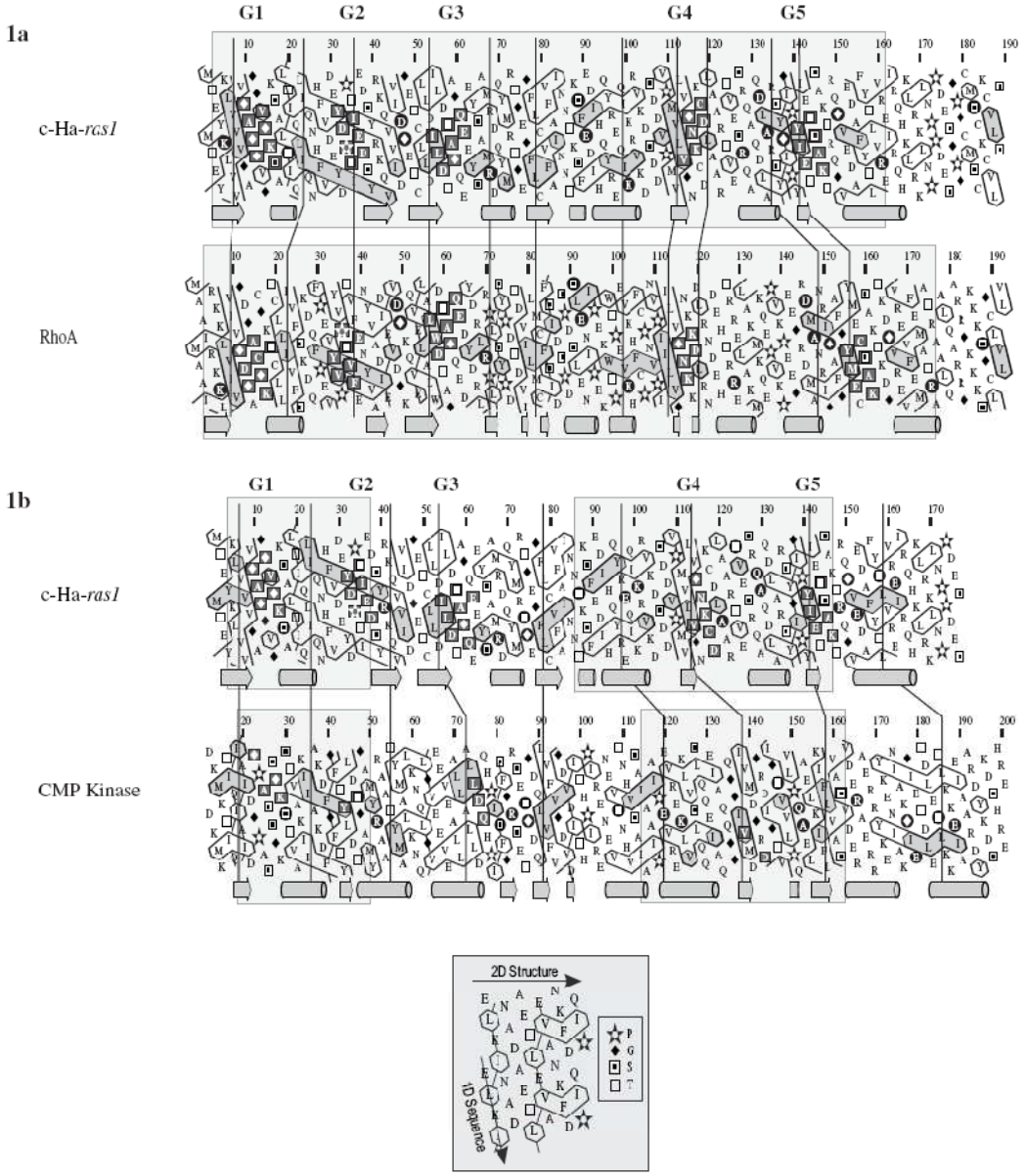

Fig. 1. HCA plots of pairs of aligned sequences. a) c-Ha-*ras1* (PDB code 4Q21) is aligned with RhoA (1X86_b), both from the small GTPases; b) c-Ha-*ras1* aligned with the CMP kinase (1Q3T), ATPase. Conserved hydrophobic clusters are grey shaded. Strict identities are indicated by white letters on a black background. White letters on a grey square indicate catalytic aa in G motifs (G1-G5 on top) and residue conservation in non-GTPases. White letters in a larger font size on a black background are residues that interact with a specific ligand as cofactors or substrates. Under each plot, the secondary structures are schemed. 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.

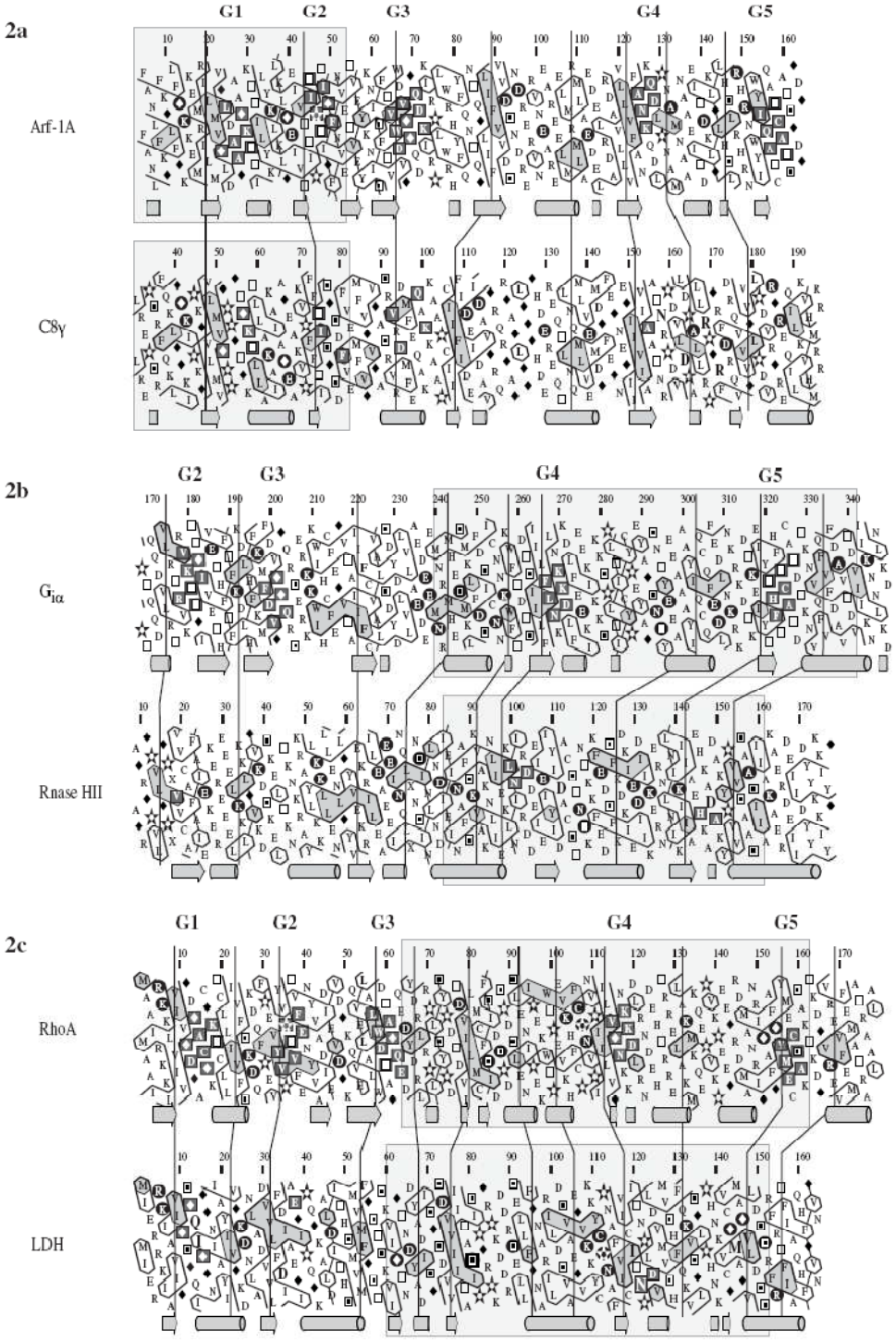

Fig. 2. HCA plots of pairs of aligned sequences. a) Arf-1A (1HUR), from the small GTPase family, aligned with C8 (1LF7), non GTPase; b) Giα (2EBC), GTPase, aligned with RNase HII (1EKE), non GTPase; c) RhoA, GTPase, aligned with LDH (2FM3), non GTPase.

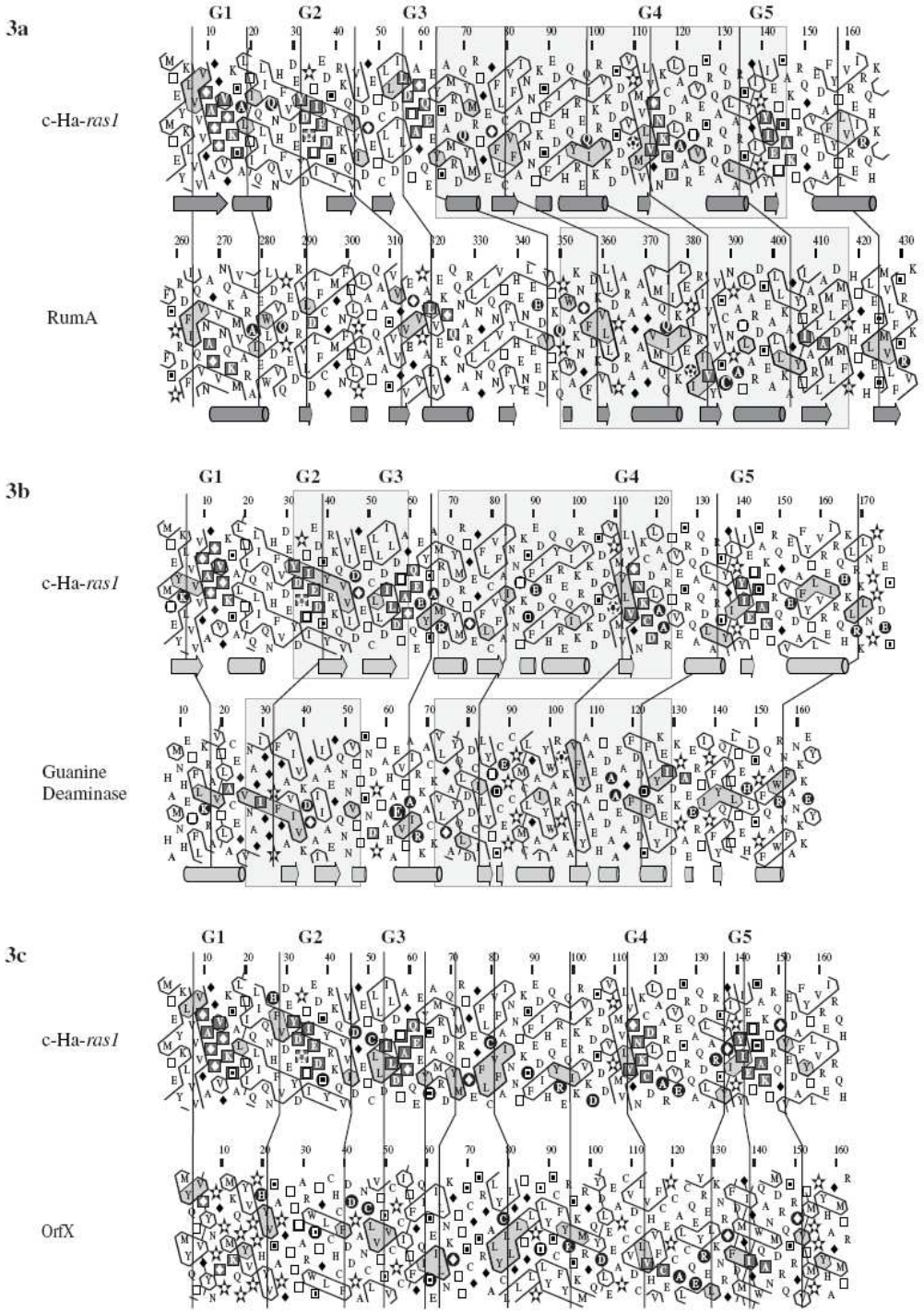

Fig. 3. HCA plots of pairs of aligned sequences. c-Ha-*ras1*, GTPase is aligned with three non GTPases a) RumA (1UWV); b) guanine deaminase (1WKQ); c) OrfX.

References

[1] Caldon, C.E., Yoong, P. and March, P.E. (2001) Evolution of a molecular switch: universal bacterial GTPases regulate ribosome function. Mol. Microbiol. 41, 289-297.

[2] Bourne, H.R., Sanders, D.A. and McCornick, F. (1991) The GTPase superfamily: A conserved structure and molecular mechanism. Nature 349, 117-127.

[3] Paduch, M., Jelen, F. and Otlewski, J. (2001) Structure of small G proteins and their regulators. Acta Biochim. Pol. 48, 829-850.

[4] Valencia, A., Kjeldgaard, M., Pai, E.F. and Sander, C. (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. USA 88, 5443-5447.

[5] Leipe, D.D., Wolf, Y.I., Koonin, E.V. and Aravind, L. (2002) Classification and evolution of P-loop GTPases and related ATPases J. Mol. Biol. 317, 41-72.

[6] Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. and Mornon, J.P. (1997) Deciphering protein sequence

information through hydrophobic cluster analysis (HCA): current status and perspectives. CMLS 53, 621-645.

[7] Woodcock, S., Mornon, J.-P. and Henrissat, B. (1992) Detection of secondary structure elements in proteins by Hydrophobic Cluster Analysis. Protein Eng. 5, 629-635.

[8] Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.P. (1987) Hydrophobic cluster analysis: an efficient new way to compare and analyze amino acid sequences. FEBS Lett. 224, 149-155.

[9] Berman, H.M., Westbrook, Z., Gilliland, F.G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2008) The Protein Data Bank. Nucleic Acids Res. 28, 235-242.

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

[11] Shindyalov, I.N., Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11, 739-747.

[12] Rost, B. (1999) Twilight zone of protein sequence alignments. Protein Eng. 12, 85-94.

[13] Henikoff, S., Henikoff, J.G. (1994) Protein family classification based on searching a database of Blocks. Genomics 19, 97-107.

[14] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235-242.

[15] Frary, A., Nesbitt, T.C., Frary, A., Grandillo, S., Van der Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K. and Tanksley, S.D. (2000) *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. Science 289, 85-88.

[16] Sakai, E., Rikimaru, K., Ueda, M., Matsumoto, Y., Ishii, N., Enomoto, S., Yamamoto, H. and Tsuchida, N. (1992) The p53 tumor-suppressor gene and *ras* oncogene mutations in oral squamous-cell carcinoma. Int. J. Cancer 52.

[17] Gómez del Pulgar, T., Benitah, S.A., Valerón, P.F., Espina, C. and Lacal, J.C. (2005) Rho GTPase expression in tumourigenesis: evidence for a significant link. Bioessays 27, 602-613.

[18] Madaule, P. and Axel, R. (1985) A novel ras-related gene family. Cell 41, 31-40.

[19] Leipe, D.D., Koonin, E.V. and Aravind, L. (2003) Evolution and classification of P-loop kinases and related proteins. J. Mol. Biol. 333, 781-815.

[20] Yu, L., Mack, J., Hajduk, P.J., Kakavas, S.J., Saiki, A.Y. and Lerner, C.G. (2003) Solution structure and function of an essential CMP kinase of *Streptococcus pneumoniae.* Protein Sci. 12, 2613-2621.

[21] Arnfors, L., Hansen, T., Schönheit, P., Ladenstein, R. and Meining, W. (2006) Structure of *Methanocaldococcus jannaschii* nucleoside kinase: an archaeal member of the ribokinase family. Acta Crystallogr. D Biol Crystallogr. 1085-1097.

[22] Ortlund, E., Parker, C.L., Schreck, S.F., Ginell, S., Minor, W., Sodetz, J.M. and Lebioda, L. (2002) Crystal structure of human complement protein C8gamma at 1.2 Å resolution reveals a lipocalin fold and a distinct ligand binding site. Biochemistry 41, 7030-7037.

[23] Schreck, S.F., Parker, C., Plumb, M.E. and Sodetz, J.M. (2000) Human complement protein C8gamma. BBA-Protein Struct M 1482, 199-208.

[24] Lai, L., Yokota, H., Hung, L.W., Kim, R. and Kim, S.H. (2000) Crystal structure of Archaeal RNase HII: a homologue of human major RNase H. Structure 8, 897-904.

[25] Lee, T.T., Agarwalla, S. and Stroud, R.M. (2004) Crystal structure of RumA, an iron-sulfur cluster containing *E. coli* ribosomal RNA 5-methyluridine methyltransferase. Structure 12, 397-407.

[26] Liaw, S.H., Chang, Y.J., Lai, C.T., Chang, H.C. and Chang, G.G. (2004) Crystal structure of *Bacillus subtilis* guanine deaminase. J. Biol. Chem. 279, 35479-35485.

[27] Gerlt, J.A. and Babbitt, P.C. (2000) Can sequence determine function? Genome Biol. 1, 1-10.

[28] Blouin, C., Butt, D. and Roger, A.J. (2004) Rapid evolution in conformational space: a study of loop regions in a ubiquitous GTP binding domain. Protein Sci. 13, 608-616.