PAP IB, a new member of the Reg gene family: cloning, expression, structural properties, and evolution by gene duplication.

Emmanuelle Laurine, Xavier Manival, Claudine Montgelard, Chantal Bideau, Jean-Louis Bergé-Lefranc, Monique Erard, Jean-Michel Verdier

PAP IB, a new member of the Reg gene family: cloning, expression, structural properties, and evolution by gene duplication

Emmanuelle Laurine[a1], Xavier Manival[b2], Claudine Montgelard[c], Chantal Bideau[d], Jean-Louis Bergé-Lefranc[d], Monique Érard[e], Jean-Michel Verdier[a3]

[a] INSERM U710, 34095 Montpellier, France; Université Montpellier II, 34095 Montpellier, France; École Pratique des Hautes Études, 34095 Montpellier, France.

[b] Institut de Cancérologie et d'Immunologie, 13009 Marseille, France.

[c] École Pratique des Hautes Études (EPHE)-UMR CNRS 5554, Université Montpellier II, 34095 Montpellier cedex 05, France.

[d] Laboratoire de Biogénotoxicologie EA1784-IFR PMSE 112, Université d'Aix-Marseille II, 13385 Marseille cedex 05, France.

[e] Institut de Pharmacologie et de Biologie Structurale (IPBS), 31077 Toulouse cedex 04, France.

[1] Current address: The Parkinson's Institute, Sunnyvale, California 94089.

[2] Current address: UMR 7567 CNRS, Faculté des Sciences, 54506 Vandoeuvre-lés-Nancy, France.

[3] Corresponding author. INSERM U710, Université Montpellier II, Place Eugène Bataillon, CC105, 34095 Montpellier cedex 05, France.

Tel/Fax: (33) 4 67 14 32 91. E-mail: verdier@univ-montp2.fr

Keywords: Reg, phylogeny, PAP IB, homology modeling, charge distribution, structure-function relationships

Abstract

Reg proteins are expressed in various organs and are involved in cancers and neurodegenerative diseases. They display a typical C-type lectin-like domain but possess additional highly conserved amino acids. By studying human databases and Expressed Sequence Tags library, we identified a new member called PAP IB. Using probabilistic approaches, we established a phylogenetic tree of eighteen Reg proteins. The dendogram showed that they constitute a superfamily composed of three distinct families (FI to FIII) of paralogues that resulted from duplication. We therefore focused on two proteins, REG Iα and PAP IB, belonging to the more closely related FI and FII families, respectively. REG Iα and PAP IB share 50% sequence identity. After cloning PAP IB, however, we found it was expressed almost only in pancreas, unlike REG Iα, whose expression is ubiquitous. In addition, by building a model of the structure of PAP IB based on the X-ray structure of REG Iα, we observed that the two proteins displayed distinctive surface charge distribution, which may lead to different ligands binding. In spite of their common fold that should result in closely related functions, REG Iα and PAP IB are a good example of duplication and divergence, probably with acquisition of new functions, thus participating in the evolution of the protein repertoire.

1. Introduction

Reg genes belong to a multigene superfamily with the typical C-type lectin-like domain (CTLD) of C-type lectins. Members of the Reg genes have been described in mammals, and their expression was studied mainly in relation to pancreatic functions. Analysis of the Expressed Sequence Tag (EST) collection of sequences and Northern blot experiments in the literature showed that the expression of Reg genes is not restricted to pancreas; but that it is also present in various tissues like stomach, colon, testis, ovary, kidney, and brain. As a consequence, many different functions have been ascribed to Reg proteins. They have been described as inhibitors of calcium carbonate crystal growth *in vitro* ([1] and [2]), as growth factors ([3] and [4]), as mitogens for pancreatic β-cells ([5] and [6]), and as a motoneuron neurotrophic factor [7]. They have also been implicated in carcinogenesis ([8]; [9] and [10]), in the improvement of diabetes in murine models ([11] and [12]), and in neurodegenerative diseases ([13] and [14]).

We thus sought to determine why Reg proteins display such different functions. Comparative genomics in eukaryotes has revealed that proteins accrete new domains and concomitantly acquire new functions [15]. “Domain accretion”, inside the same families or with different families, accounts for the incomparable diversity of eukaryote protein function. However, this diversity should not be found in Reg proteins. Indeed, because they are small proteins (around 150 aminoacids), they consist of only one domain, *i.e*. one evolutionary unit [16]. As a consequence, more subtle differences in the architectures of Reg proteins should explain their pleiotropic roles.

The shared chromosomal localization of Reg genes in humans (2p12 [17], [18]) and mice (6C [19], their tandem order, and their common intron-exon organization [17] suggest they likely derived by duplication from a common ancestor. In this report, we describe the

identification of a new member of the Reg family. By leading detailed phylogenetic analyses of the evolution within this family, and fine structural analysis, we then provide a framework for further functional research on Reg proteins.

2. Material & Methods

2.1. Construction of the data file

The near completion of the human genome sequencing project ([20] and [21]) allowed us to derive the quasi-complete set of Reg genes in a vertebrate genome. To obtain the most exhaustive analysis, we screened all the genomic and structural databases available, *i.e*. the non-redundant GenBank/European Molecular Biology Laboratory database, the DNA data bank of Japan, the human Expressed Sequence Tag library (dbEST), the human genome project-derived sequences (draft genome, High Throughput Genomic Sequences), the Swissprot database, the Protein family (Pfam), and the Structural Classification of Proteins (SCOP). During our work, four successive drafts of the human genome became available. The data presented in this paper were obtained from the latest version.

2.2. Sequence alignment

Eighteen complete sequences of the Reg superfamily were retrieved and were manually aligned using MUST software [22]. The alignment, including small gaps, consisted of 160 amino acids. We removed both the PTP_Pig sequence (gi:3024090) because it was incomplete, and the INGAP sequence of mouse (gi:7023941), because it is probably a Reg IIIδ polymorphic sequence (gi:6633974). The INGAP sequences of hamsters

(gb:AAB16754.1) and humans (gb:AAB86497.1) proposed by Rafaeloff et al. [23] were surprisingly identical. Exhaustive searches in databases showed that this latter sequence has never been identified in any of the genomic databases. We attributed this INGAP sequence to the hamster. Finally, the Reg-related sequence (RS, gi:893382) was not included because of its premature ending due to an in-frame stop codon.

2.3. Phylogenetic analysis

Phylogenetic analysis. Phylogenetic analyses were conducted on the amino acid dataset including 21 taxa (19 Reg sequences and 2 outgroups) for 153 sites, after removal of the seven first amino acids of Reg II_Mouse. Trees were reconstructed using two different probabilistic approaches, the maximum likelihood method and the Bayesian inference. In both cases, we used the empirical substitution matrix between amino acids (JTT substitution model, [24]), with rate heterogeneity between sites being estimated by a gamma distribution with four discrete categories and a proportion of invariable sites. The maximum likelihood approach was performed with the program PhyML [25] for which nodal support was estimated with 1000 bootstrap replications. The Bayesian tree was constructed with the MrBayes 3.0b4 program [26] using 4 chains that were run for 1,000,000 generations. Posterior probabilities were determined from the 50% majority rule consensus of trees sampled every 50 generations and from burn-in of the first 1,500 trees (stationarity of likelihood values was checked empirically).

2.4. PCR experiments

PCR tests were done on human Multiple Tissue cDNA (MTC Human I, Clontech Lab.

Inc., Palo Alto, CA) and on small intestine cDNA synthesized from total RNA (Clontech) with the recombinant Taq DNA polymerase (Gibco, Invitrogen Corp., Carlsbad, CA) according to supplier instructions. Primers were designed for specifically amplifying Reg or PAP genes: REG Iα 5' (AGTGTTAATCCTGGCTACTGTGTG) and 3' (AGAGTGTCCAGGTTGAGTTGAGTT), REG Iβ 5' (GTGCTAATGCTGGCTACTGTGCA) and 3' (GGTGAAGGTACTGAAGATCAGCG), HIP/PAP 5' (GTGGAAAGATTATAACTGTAATGTG) and 3' (CTTTAAAGCCTTAGGCCGTATGA); PAP IB 5' (GTGGAAAGATTATAACTGTGATGCA) and 3' (CTCTGAGATCTCAAGACATGGAA). These primers amplified a 212 bp fragment for REG Iα, a 249 bp fragment for REG Iβ, a 256 bp fragment for HIP/PAP, and a 257 bp fragment for PAP IB. For REG Iα and REG Iβ, the cycling parameters were the following: initial denaturation 3 min 94°C, 30 sec 94°C, 30 sec 65°C, 1 min 72°C (30-35 cycles), final extension 5 min 72°C. The conditions were identical for HIP/PAP and PAP Iβ except that annealing was done at 60°C instead of 65°C. The sequence of each PCR product was checked by DNA sequencing.

2.5. PAP IB subcloning

The PAP clone (ID 2237170) was obtained from the I.M.A.G.E. Consortium (HGMP Resource Centre, Hinxton, UK). PAP IB cDNA was prepared by PCR using the 5' (CCAGACAAAGCTTACCAGATC) and 3' (CCCACCTGCCGAATTCCTTG) primers, then digested by HindIII and EcoRI, and ligated with T4 DNA ligase (Gibco Invitro Corp.) for 1h at room temperature into the HindIII/EcoRI-digested pcDNA 3.1/V5-His mammalian expression vector (Invitrogen). This construct generates a fusion of PAP IB with a C-terminal V5 tag. Then, DH5α^{TM} competent cells (Gibco Invitro Corp.) were transformed and plated

according to manufacturer instructions. Six positive clones were chosen for transfection experiments. The PAP IB nucleotide sequence has been submitted to GenBank (accession number **AY428734**).

2.6. CHO cell culture and transfection

CHO cell line (ECCAC number 85050302) was cultured in Ham's F12 medium supplemented with 10% fetal bovine serum (FBS) (Roche, Basel, Switzerland), penicillin (100 IU/ml), and streptomycin (100 µg/ml). Cells were transfected with lipofectAMINE™ (Gibco, Invitrogen Corp., CA). Forty-eight hours after transfection, the medium was removed and supplemented by a concentrated Laemmli sample buffer to reach the final concentration: 62 mM Tris.Cl pH 6.8, 2% SDS, 5% glycerol, and 5% β-2-mercaptoethanol. The samples were then heated at 100°C for 3 min and subjected to 15% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Proteins were then electrotransferred overnight to a PVDF sheet (Immobilon P, Millipore, Bedford, MA). Non specific sites were blocked with 5% non-fat milk powder in 20 mM Tris.Cl pH 7.4, and 50 mM NaCl (TBS) containing 0.1% Tween 20. The blot was incubated 2h at room temperature with anti-V5 antibody (Gibco Invitrogen Corp.) diluted at 1/1000 in TBS supplemented with 0.1% Tween 20 and 0.3% non-fat milk powder. The membrane was washed five times with TBS, 0.1% Tween 20 and then incubated 2h at room temperature with antiperoxidase conjugated goat anti-mouse IgG diluted at 1/20,000 in TBS, 0.1% Tween 20, and 0.3% non-fat milk. The membrane was washed five times with TBS, 0.1% tween 20. The revelation was done by chemiluminescence (Pierce, Rockford, IL).

2.7. Molecular modeling

Homology modeling of the human PAP IB was performed according to the main principles outlined by Greer [27]. We used the *Accelrys* software modules *InsightII*, *Homology,* and *Discover* (San Diego, CA), run on a Silicon Graphics O2 workstation (SGI, Mountain View, CA). The human PAP IB protein sequence was aligned against the protein sequences from the PDB (Protein Data Bank) by using the SSEARCH program (http://pbil.univ-lyon1.fr) and we identified REG Iα/lithostathine (1QDD) as the best reference to build the PAP IB three-dimensional model. (SCRs) and loops. The main modeling steps involved first the transfer of coordinates between the "structurally conserved regions" (SCRs), the building of loops, and the optimization of the side-chain orientations. Then, the final structural model was refined by energy minimization, using the consistent valence forcefield (cvff*)* and the steepest descent and conjugate gradient algorithms, down to a maximum derivative of 0.01 kcal/Å/mol. The final energy of the resulting model was –1416 kcal. The validity of the model was assessed by both 'structural check' and 'folding consistency verification' using respectively *Prostat* ([28] and [29]) and Profiles_3D [30] programs within the *Homology* module. No spurious angle, bond length, or misfolded region was detected. The percentage of the so-called 'most favored regions' in the Ramachandran plot was 81% for PAP IB (relatively to 83.8% for the template 1QDD).

3. Results

3.1. A novel Reg-like gene, PAP IB, maps at 2p12

Because all Reg proteins had been localized on chromosome 2 at the 2p12 locus, we looked at the existence of new genes in this family through the successive genome drafts. By

analyzing the sequences at locus 2p12, we identified a novel PAP-like sequence that we named PAP IB according to international rules [31]. The PAP IB_Human nucleotidic sequence was predicted by assemblies of EST through the CAP software [32]: BF056837, AI621017, AI027597, AA621263, AA725500, AI025191, AA399061, AW962460, AW196366, AW002913, AA397543, C17743, AA367364, AI198647. The sequence was then translated into the three ORF and the longest one was selected. At the same time, Okamoto group also identified this gene and named it REG III [18].

3.2. Human PAP IB gene cloning and expression

To establish that PAP IB is not only transcribed but also exists as a true protein, we transfected CHO cells with a PAP IB construct as described in "Material & Methods". Figure 1 shows the protein is secreted in the extracellular milieu as expected, and that the size of the band (15.5 kDa) is in good agreement with the expected theoretical size (16.5 kDa).

3.3. Reg and Reg-related genes display specific sequences and specific structural elements

Members of the Reg multigene family showed a strong similarity with the CTLD domain found in the calcium-dependent animal lectin superfamily [33] (Fig. 2). This motif includes four conserved cysteines forming two disulfide bridges (S1 and S2), and additional amino acids that are structurally essential ("♦" in Fig. 2) [34]. There are, however, four major differences. First, the Reg multigene family displays an additional disulfide bridge (S3) very close to the N-terminus. This disulfide bridge directs the separation of Reg proteins into two parts: a very short N-terminal undecapeptide, except those whose lengths vary (Reg II_Mouse, REG IV_Human, and Putative_Mouse), and a globular C-terminal domain.

Second, conversely to true C-type lectins, despite numerous attempts, no carbohydrate ligands have been identified so far for any Reg protein. Third, in true C-type lectins, the CTLD contains two calcium-binding sites. Reg proteins do not bind calcium through these sites. Indeed, site 1 involves an aspartic acid. This acidic residue is systematically lost in Reg proteins and replaced by aromatic amino acids, most often Y, but also F and H (***** in Fig. 2). In addition, site 1 of true C-type lectins is not topologically equivalent to that of REG Iα/lithostathine. Therefore, both topology and amino acid substitution prevent calcium binding ([35], and unpublished observations). Site 2 is topologically conserved through α-carbon in Reg proteins and C-type lectins such as rat-mannose-binding protein, but this site displays different amino acids that are expected to prevent calcium binding [35] (not shown). However, REG Iα/lithostathine binds calcium via a cluster of acidic amino acids located in the surface of the protein [36]. Fourth, there are several other extremely conserved amino acids specific to the whole Reg family compared to true C-type lectins (highlighted in Fig. 2). All in all, these observations showed that Reg proteins formed a distinct subclass of the family of proteins that contain CTLDs with characteristic features. This alignment of Reg genes was then used for phylogenetic studies as explained in the next section.

3.4. Hierarchical organization of the Reg genes

To have an exhaustive view of the Reg genes organization, we first built an IBT tree on 774 members of the lectin superfamily from the Pfam alignment. For this purpose, we used the neighbor-joining algorithm implemented in the MEGA software [Kumar, 1994 #429]. The results revealed the presence of a monophyletic Reg family containing 19 members found in mammals only (data not shown). No related genes have been characterized in other eukaryotic kingdoms such as fungi or plants. On the basis of the IBT tree, we were able to choose two non-mammal lectin sequences as outgroups, the closest to the Reg family: the BAC54022.1

eel sequence and the sp P81017 viper sequence. Results of phylogenetic analyses conducted with the maximum likelihood method and Bayesian inference strongly supported the existence of three different families called FI, FII and FIII (Fig. 3). FI family included 11 sequences among which the mean percent sequence divergence is 39 ± 10 % (mean ± SD). In FI, both reconstruction methods provided a strong support for the existence of 5 subgroups (a, b, c, d, and e in Fig. 3). Among them, subgroups b (Posterior Probabilities = 1, Bootstrap Proportion = 100), c (PP = 1, BP = 100), d (PP = 1, BP = 99), and e (PP = 1, BP = 100) most probably arose from four specific duplications in the rodent lineage because corresponding genes do not exist in the human lineage. By contrast, relationships between the clades b to e were not well supported, preventing to know in which order these duplications arose. The human HIP/PAP and PAP IB (16.1 % of divergence, subgroup a, PP = 0.97, BP = 82) are likely to be paralogous genes. The PTP_bovin gene appeared as the first emergence in the FI family. Although not supported in the maximum likelihood method, this position is coherent with molecular studies on mammalian relationships showing that rodent and primates belong to the same clade (Euarchontoglires ; [Murphy, 2001 #856]). Five Reg sequences were included in the FII family and the mean percent sequence divergence is 24.4 ± 7.5 % whereas the mean divergence between FI and FII families is 57.3 ± 2.5 %. An independent human duplication for the REG I genes was strongly supported (PP=0.9, BP=76). These observations indicated that an ancestor common to rodent and primate lineages possessed the PAP-like and REG1-like genes. Finally, the third family (FIII) only included three members: REG IV_Human, Reg IV_Rat and Putative_Mouse. The mean sequence divergence between the three sequences was 26.4 % and the mean divergence between FIII and FI, and FIII and FII was 69.7 ± 2.6 % and 69.8 ± 1.8 %, respectively. In addition, REG IV_Human is localized on chromosome 1, whereas all other members are located in the same cluster of

chromosome 2 (Fig. 4). These results are in agreement with human Reg genes hierarchical organization of obtained by Nata and collaborators [18].

3.5. Human REG I and PAP gene expression

Because REG I and PAP genes belonged to two families, we explored their expression pattern in a human tissue panel. Unlike the REG Iα gene, expressed in all tested tissues, the REG Iβ, HIP/PAP, and PAP IB genes were expressed almost only in the pancreas (Fig. 5). A faint expression of REG Iβ and PAP IB was observed in liver and placenta respectively. In small intestine, we only observed a clear expression of REG Iα, REG Iβ and HIP/PAP whereas PAP IB expression was nearly absent (Fig 5E). Differential expression patterns between HIP/PAP and PAP IB were also reported in small intestine, liver, hepato-carcinoma, stomach, kidney and testis [18]. First, these results showed that REG Iα gene probably has pleiotropic roles in the organism; by contrast, the expression of REG Iβ and PAP genes seemed more restrictive. Second, results suggested that, in spite of recent divergence, REG Iα_Human and REG Iβ_Human genes, HIP/PAP and PAP IB, might already have different roles in the organism.

3.6. PAP IB and REG Iα/lithostathine are a structural family within the CTLD domain superfamily

To correlate likely functional variations with structural differences between PAP IB and REG Iα/lithostathine, we compared the three-dimensional structures of the newly characterized PAP IB (coded by PAP IB_Human gene, FI) and the most representative Reg protein, REG Iα/lithostathine (coded by REG Iα_Human gene, FII). The known three-

dimensional structure of REG Iα/lithostathine ([35] and [39]) allowed us to model that of PAP IB (see the Materials and Methods section). The resulting model is shown in Figure 6A. The r.m.s.d. value between the PAP IB modeled structure and the REG Iα/lithostathine crystallographic one was 0.5 Å. The only distinctive topological feature was the 5-residue-longer PAP IB loop between the β4 and β5 strands (as indicated by a solid arrow pointing to this blue region). In fact, this loop, which is specific to the lithostathine structural family, belongs to a larger region that is highly divergent from that of the snake venom lectin-type toxins (data not shown).

3.7. REG Iα/lithostathine and PAP IB showed a distinctive charge distribution

Interestingly, beyond their similar fold, REG Iα/lithostathine and PAP IB differ in the density distribution of their charged residues (Fig. 6B-D). This is especially visible at the level of the differential loop, which is negatively charged for PAP IB whereas it is positively charged for REG Iα/lithostathine (Fig. 6C-D). In addition, other regions also differ in charge throughout the whole sequence. Interestingly, HIP/PAP, which also belongs to the FIa sub-family like PAP IB (see Fig. 3), displays the same structural characteristics as PAP IB (Fig. 6B). This peculiarity may have important consequences in terms of function. All in all, these results underline that, in spite of the same general fold, REG Iα/lithostathine and PAP IB may differ in their function via local structural differences.

4. Discussion

Here we described a new member of the Reg family, named PAP IB, located on

chromosome 2. PAP IB is in the same cluster as other Reg genes, suggesting that this multigene family probably arose by gene duplication (see below). The Reg family is specific to mammals, and phylogenetic analysis of 18 mammalian sequences revealed the existence of different species-specific paralogous genes clustering in three families. Not only are these families of unequal size but also their diversification differs with the lineage. In the PAP family (FI), phylogenetic relationships in rodents indicate that duplication occurred at least 4 times (subgroups b to e) but only once in the human lineage. At the present time, however, it is not possible to know which genes are orthologous in the two lineages.

The Reg family possibly evolved through duplication events that have conferred a direct advantage to the organism (positive selection) or by recurrent duplications that have led to a polymorphism with chromosomes carrying different copy numbers [40]. This notion is largely corroborated by our detailed structural analysis of REG Iα/lithostathine and PAP IB, which belong to two main distinct branches of the phylogenetic tree (FI and FII, respectively). Indeed, human REG Iα/lithostathine and PAP IB belong to a single structural clade based on the remarkable conservation of the types of residues that govern folding (*i.e.* the hydrophobic/aromatic residues and the Cys, Pro, and Gly residues). This similar fold is clearly distinct from those formed by other members of the CTLD superfamily, as shown by our structural alignments (unpublished results). The similar fold might therefore reflect a general function common to Reg proteins, such as interactions with similar protein partners or involvement in similar signalization pathways. For example, REG Iα/lithostathine and HIP/PAP (the closest relative to PAP IB according to our phylogenetic tree) are both inflammatory proteins overexpressed in the brain of patients with Alzheimer's disease [13].

In the course of evolution, duplicate genes have many fates: they can conserve the same functions, be lost, evolve to pseudogenes, or generate highly different functions [41]. Therefore, linking genome sequence analysis to protein structure, and then to function, is the

major endeavor of the post-genomic era (for reviews, see [42], [43] and [44]). Here, the structure of the newly described PAP IB protein was predicted through comparative modeling. Taking into account both the 50% sequence identity between the two sequences and the high-resolution crystallographic structure of REG Iα/lithostathine, we were in the requested conditions to accurately build a homology-derived model of PAP IB. That the PAP IB modeled backbone differed from the one of REG Iα/lithostathine by only 0.5 Å highlights the inherent topological similarity between the two proteins. According to Wilson *et al.* [45], proteins/domains with a common fold and sequence identities down to about 40% have the same specific function, whereas between 40% and 25%, they share only the "broad" function. This stands true in the case of the RRM (RNA Recognition Motif) domains: as a whole family of about 200 members, they share only 24% sequence identity (according to Pfam database) for a common single-stranded RNA recognition activity, but, as a specific sub-class of AU-rich RNA binders, they possess 37% sequence identity [46]. In fact, the 40% threshold is only a "rule of thumb", subject to variations depending on the domain family. Thus, the 50% sequence identity level between PAP IB and REG Iα/lithostathine is enough to direct the general Reg-type fold, but it may not be sufficient to ensure the same specific function to the two proteins. As suggested by Andrade and colleagues [47], if the overall fold is responsible for invariable aspects of function, more subtle properties would be responsible for the divergent aspects of function. This observation raises a major question: Are there other specific functions that could be related to structural properties other than the overall fold ?

One potential source of functional diversity could be the charge of the protein. As first defined by Lichtarge et al. [48], the powerful concept of evolutionary trace has revealed that surface map comparison can rationalize the functional variation within a divergent group of proteins ([49], [50] and [51]). Indeed, besides their different tissue expression, PAP IB and REG Iα/lithostathine display notable different structural properties at a finer level. For

instance, the specific loop of Reg protein family, between β4 and β5, is both 5 residues longer and oppositely charged in PAP IB relative to REG Iα/lithostathine. The different charge distributions of PAP IB and REG Iα/lithostathine that we observed might generate specific responses to different ligands that need to be further investigated.

Our results confirm that the complexity of organisms is not linked to the number of genes, but rather to the ability of an organism to duplicate, diverge and recombine, thus creating new functions. This work illustrates that the marriage of phylogeny systematics with structural studies promises new insights into the nature of biological structure/function relationships. Both the genome organization and the three-dimensional structure of proteins should be part of the study of evolution.

Acknowledgements

We thank J. Lanet and J.L. Franc for help in transfection and Western blot experiments. EL is supported by GRAL (Groupe de Recherche sur la maladie d'ALzheimer) and France Alzheimer. ME was a recipient of a contract from Région Midi-Pyrénées #01008888. This work was supported in part by grants from the "G.I.S.-Infections à prions" from the French government (#F37). This paper is the contribution N°2005-002 from the "Institut des Sciences de l'Évolution" of Montpellier, France.

Figure legends

Figure 1. *Immunodetection of PAP IB after CHO transfection*. Lane M: Molecular weight markers. Lane C: Control experiment after transfection with pcDNA 3.1/V5-His vector. Lane PAP IB: Extracellular milieu after transfection with PAP IB construct. The sample was analyzed on immunoblots with anti-V5 antibody which recognizes the C-terminal V5 tag of PAP IB. Molecular weight standards are given by ChemiBlot Molecular Weight Markers-Trial from Chemicon international, Inc (Temecula, California).

Figure 2. *Sequence alignment of the Reg family proteins*. All members of the Reg family are secretory proteins. Signal peptide cleavage sites were predicted according to Nielsen et al. [52], except for REG Iα/lithostathine_Human biochemically determined [53]. Residues are indicated by a single-letter amino acid code. Sequence gaps introduced to optimize the alignment are indicated by dashes. The box corresponds to the CTLD domain [33]. The invariant amino acids in C-type lectin [34] are indicated by a black diamond-shaped ("♦"). "+" signs refer to the residues involved in the two calcium binding sites of C-type lectins. Gray highlighted amino acids represent residues strictly conserved throughout the whole Reg family. The star sign (*****) indicates the mutation in Reg proteins preventing calcium binding. S_1, S_2, and S_3 represent the three disulfide bridges. See also the paper by Okamoto and Takasawa [54].

Figure 3. *Phylogram of the REG family members obtained with the Bayesian approach*. The eel and viper sequences were used as outgroups. The JTT substitution model was used with rate heterogeneity between sites estimated by a gamma distribution and a proportion of invariable sites. Numbers at nodes refer, from left to right, to Bayesian posterior probabilities

and bootstrap proportions in Maximum Likelihood. FI, FII and FIII represent the three families with regard to the human sequences. Clades *a-e* represent the five subgroups observed in FI family.

Figure 4. *Chromosomal organization of human Reg genes*. Arrows indicate the sense of translation. ATG start (upper) and STOP (lower) codon numbers are indicated for each gene. Number 1 was attributed to the ATG of HIP/PAP. Φ is a Reg related pseudogene.

Figure 5. *Comparison of tissue expression pattern of REG and PAP mRNAs in the sample MTC panel*. The expressions of Reg and PAP genes were analyzed by RT-PCR with specific primers. After amplification cycles, samples were electrophoresed on 2.5% agarose/EtBr gels: lane 1 (heart), lane 2 (brain), lane 3 (placenta), lane 4 (lung), lane 5 (liver), lane 6 (skeletal muscle), lane 7 (kidney), lane 8 (pancreas), and lane 9 (control experiment). Panel A: REG Iα; panel B: REG Iβ; panel C: HIP/PAP; panel D: PAP IB. Panel E : expression in small intestine of REG Iα (lane 10), REG Iβ (lane 11), HIP/PAP (lane 12), PAP IB (lane 13). Molecular weight standards (M) are given by X174 RF DNA/HaeIII from Invitrogen Corp. Control experiments showed no detectable bands.

Figure 6. *Structural analysis* (A) Model of the three-dimensional structure of the human PAP IB built by homology with REG Iα/lithostathine. The 10 structurally conserved elements '$\beta_1\beta_2\alpha_1\beta_3\alpha_2\beta_4\beta_5\beta_6\beta_7\beta_8$' are numbered sequentially and are color-coded (α-helices in brown and β-strands in indigo). The blue arrow points to the distinctive feature, the longer loop between β4 and β5 than in REG Iα/lithostathine. The five-residue insertion QGSEP, characteristic of PAP IB, is in blue. (B) Multiple-sequence alignment of PAP IB, HIP/PAP and REG Iα/lithostathine showing the respective distributions of acidic residues in brown and

basic residues in turquoise (C, D) Charged residue distribution at the surface of human PAP IB and REG Iα/lithostathine. Acidic and basic residues are displayed in Connolly surface mode and color-coded as in B. The brown and green arrows point to the differential loop, respectively 5 residues longer and negatively charged in PAP IB and short and positively charged in REG Iα/lithostathine.

References

[1]J.P. Bernard, Z. Adrich, G. Montalto, A. De Caro, M. De Reggi, H. Sarles and J.C. Dagorn, Inhibition of nucleation and crystal growth of calcium carbonate by human lithostathine, Gastroenterology 103 (1992) 1277-84.

[2]S. Geider, A. Baronnet, C. Cerini, S. Nitsche, J.P. Astier, R. Michel, R. Boistelle, Y. Berland, J.C. Dagorn and J.M. Verdier, Pancreatic lithostathine as a calcite habit modifier, J Biol Chem 271 (1996) 26302-6.

[3]E. Anastasi, E. Ponte, R. Gradini, A. Bulotta, P. Sale, C. Tiberti, H. Okamoto, F. Dotta and U.D. Mario, Expression of Reg and cytokeratin 20 during ductal cell differentiation and proliferation in a mouse model of autoimmune diabetes, Eur J Endocrinol 141 (1999) 644-52.

[4]H. Okamoto, The Reg gene family and Reg proteins: with special attention to the regeneration of pancreatic beta-cells, J Hepatobiliary Pancreat Surg 6 (1999) 254-62.

[5]M.E. Zenilman, J. Chen and T.H. Magnuson, Effect of reg protein on rat pancreatic ductal cells, Pancreas 17 (1998) 256-61.

[6]J.L. Levine, K.J. Patel, Q. Zheng, A.R. Shuldiner and M.E. Zenilman, A recombinant rat regenerating protein is mitogenic to pancreatic derived cells, J Surg Res 89 (2000) 60-5.

[7]H. Nishimune, S. Vasseur, S. Wiese, M.C. Birling, B. Holtmann, M. Sendtner, J.L. Iovanna and C.E. Henderson, Reg-2 is a motoneuron neurotrophic factor and a signalling intermediate in the CNTF survival pathway, Nat Cell Biol 2 (2000) 906-14.

[8]C. Lasserre, L. Christa, M.T. Simon, P. Vernier and C. Bréchot, A novel gene (HIP) activated in human primary liver cancer, Cancer Res 52 (1992) 5089-95.

[9]T. Itoh, N. Sawabu, Y. Motoo, A. Funakoshi and H. Teraoka, The human pancreatitis-associated protein (PAP)-encoding gene generates multiple transcripts through alternative use of 5' exons, Gene 155 (1995) 283-7.

[10]H. Rechreche, G. Montalto, G.V. Mallo, S. Vasseur, L. Marasa, P. Soubeyran, J.C. Dagorn and J.L. Iovanna, pap, reg Ialpha and reg Ibeta mRNAs are concomitantly up-regulated during human colorectal carcinogenesis, Int J Cancer 81 (1999) 688-94.

[11]T. Watanabe, Y. Yonemura, H. Yonekura, Y. Suzuki, H. Miyashita, K. Sugiyama, S. Moriizumi, M. Unno, O. Tanaka, H. Kondo and et al., Pancreatic beta-cell replication and amelioration of surgical diabetes by Reg protein, Proc Natl Acad Sci U S A 91 (1994) 3589-92.

[12]D.J. Gross, L. Weiss, I. Reibstein, J. van den Brand, H. Okamoto, A. Clark and S. Slavin, Amelioration of diabetes in nonobese diabetic mice with advanced disease by linomide-induced immunoregulation combined with Reg protein treatment, Endocrinology 139 (1998) 2369-74.

[13]L. Duplan, B. Michel, J. Boucraut, S. Barthellemy, S. Desplat-Jego, V. Marin, D. Gambarelli, D. Bernard, P. Berthezene, B. Alescio-Lautier and J.M. Verdier, Lithostathine and pancreatitis-associated protein are involved in the very early stages of Alzheimer's disease, Neurobiol Aging 22 (2001) 79-88.

[14]E. Laurine, C. Gregoire, M. Fandrich, S. Engemann, S. Marchal, L. Thion, M. Mohr, B. Monsarrat, B. Michel, C.M. Dobson, E. Wanker, M. Erard and J.M. Verdier, Lithostathine Quadruple-helical Filaments Form Proteinase K-resistant Deposits in Creutzfeldt-Jakob Disease, J Biol Chem 278 (2003) 51770-51778.

[15]E.V. Koonin, L. Aravind and A.S. Kondrashov, The impact of comparative genomics on our understanding of evolution, Cell 101 (2000) 573-6.

[16]A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J Mol Biol 247 (1995) 536-40.

[17]C. Bartoli, J.C. Dagorn, M. Fontes and J.L. Bergé-Lefranc, A limited genomic region contains the human REG and REG-related genes, Eur J Hum Genet 3 (1995) 344-50.

[18]K. Nata, Y. Liu, L. Xu, T. Ikeda, T. Akiyama, N. Noguchi, S. Kawaguchi, A. Yamauchi, I. Takahashi, N.J. Shervani, T. Onogawa, S. Takasawa and H. Okamoto, Molecular cloning, expression and chromosomal localization of a novel human REG family gene, *REG III*, Gene 340 (2004) 161-170.

[19]Y. Narushima, M. Unno, K. Nakagawara, M. Mori, H. Miyashita, Y. Suzuki, N. Noguchi, S. Takasawa, T. Kumagai, H. Yonekura and H. Okamoto, Structure, chromosomal localization and expression of mouse genes encoding type III Reg, RegIII alpha, RegIII beta, RegIII gamma, Gene 185 (1997) 159-68.

[20]E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860-921.

[21]J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, et al., The sequence of the human genome, Science 291 (2001) 1304-51.

[22]H. Philippe, MUST, a computer package of Management Utilities for Sequences and Trees, Nucleic Acids Res 21 (1993) 5264-72.

[23]R. Rafaeloff, G.L. Pittenger, S.W. Barlow, X.F. Qin, B. Yan, L. Rosenberg, W.P. Duguid and A.I. Vinik, Cloning and sequencing of the pancreatic islet neogenesis associated protein (INGAP) gene and its expression in islet neogenesis in hamsters, J Clin Invest 99 (1997) 2100-9.

[24]D.T. Jones, W.R. Taylor and J.M. Thornton, The rapid generation of mutation data matrices from protein sequences, Comput Appl Biosci 8 (1992) 275-82.

[25]S. Guindon and O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst Biol 52 (2003) 696-704.

[26]J.P. Huelsenbeck and F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics 17 (2001) 754-5.

[27]J. Greer, Comparative modeling of homologous proteins, Methods Enzymol 202 (1991) 239-52.

[28]A.L. Morris, M.W. MacArthur, E.G. Hutchinson and J.M. Thornton, Stereochemical quality of protein structure coordinates, Proteins 12 (1992) 345-64.

[29]R.A. Laskowski, D.S. Moss and J.M. Thornton, Main-chain bond lengths and bond angles in protein structures, J Mol Biol 231 (1993) 1049-67.

[30]L. Wesson and D. Eisenberg, Atomic solvation parameters applied to molecular dynamics of proteins in solution, Protein Sci 1 (1992) 227-35.

[31]J.A. Blake, M.T. Davisson, J.T. Eppig, L.J. Maltais, S. Povey, J.A. White and J.E. Womack, A report on the international nomenclature workshop held may 1997 at the Jackson Laboratory, Bar Harbor, Maine, USA, Genomics 45 (1997) 464-8.

[32]X. Huang, A contig assembly program based on sensitive detection of fragment overlaps, Genomics 14 (1992) 18-25.

[33]K. Drickamer, Two distinct classes of carbohydrate-recognition domains in animal lectins, J Biol Chem 263 (1988) 9557-60.

[34] K. Drickamer and M.E. Taylor, Biology of animal lectins, Annu Rev Cell Biol 9 (1993) 237-64.

[35]J.A. Bertrand, D. Pignol, J.P. Bernard, J.M. Verdier, J.C. Dagorn and J.C. Fontecilla-Camps, Crystal structure of human lithostathine, the pancreatic inhibitor of stone formation, Embo J 15 (1996) 2678-84.

[36]B.I. Lee, D. Mustafi, W. Cho and Y. Nakagawa, Characterization of calcium binding properties of lithostathine, J Biol Inorg Chem 8 (2003) 341-7.

[37]S. Kumar, K. Tamura and M. Nei, MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers, Comput Appl Biosci 10 (1994) 189-91.

[38]W.J. Murphy, E. Eizirik, S.J. O'Brien, O. Madsen, M. Scally, C.J. Douady, E. Teeling, O.A. Ryder, M.J. Stanhope, W.W. de Jong and M.S. Springer, Resolution of the early placental mammal radiation using Bayesian phylogenetics, Science 294 (2001) 2348-51.

[39]V. Gerbaud, D. Pignol, E. Loret, J.A. Bertrand, Y. Berland, J.C. Fontecilla-Camps, J.P. Canselier, N. Gabas and J.M. Verdier, Mechanism of calcite crystal growth inhibition by the N-terminal undecapeptide of lithostathine, J Biol Chem 275 (2000) 1057-64.

[40]A.G. Clark, Invasion and maintenance of a gene duplication, Proc Natl Acad Sci U S A 91 (1994) 2950-4.

[41]J. Zhang, Evolution by gene duplication: an update, Trends Ecol Evol 18 (2003) 292-298.

[42]C.A. Orengo, A.E. Todd and J.M. Thornton, From protein structure to function, Current Opinion in Structural Biology 9 (1999) 374-82.

[43]D. Baker and A. Sali, Protein structure prediction and structural genomics, Science 294 (2001) 93-6.

[44]B. Rost, B. Honig and A. Valencia, Bioinformatics in structural genomics, Bioinformatics 18 (2002) 897-8.

[45]C.A. Wilson, J. Kreychman and M. Gerstein, Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores, Journal of Molecular Biology 297 (2000) 233-49.

[46]L. Thion and M. Erard, Structure/function relationships within the RNA recognition motif family applied to the hermes gene product, Protein & Peptide Letters 9 (2002) 127-32.

[47]M.A. Andrade, S.I. O'Donoghue and B. Rost, Adaptation of protein surfaces to subcellular location, J Mol Biol 276 (1998) 517-25.

[48]O. Lichtarge, H.R. Bourne and F.E. Cohen, An evolutionary trace method defines binding surfaces common to protein families, J Mol Biol 257 (1996) 342-58.

[49]K. Pawlowski and A. Godzik, Surface map comparison: studying function diversity of homologous proteins, J Mol Biol 309 (2001) 793-806.

[50]D.R. Livesay, P. Jambeck, A. Rojnuckarin and S. Subramaniam, Conservation of electrostatic properties within enzyme families and superfamilies, Biochemistry 42 (2003) 3464-73.

[51]F.E. May, S.T. Church, S. Major and B.R. Westley, The closely related estrogen-regulated trefoil proteins TFF1 and TFF3 have markedly different hydrodynamic properties, overall charge, and distribution of surface charge, Biochemistry 42 (2003) 8250-9.

[52]H. Nielsen, J. Engelbrecht, S. Brunak and G. von Heijne, Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, Protein Eng 10 (1997) 1-6.

[53]A.M. De Caro, Z. Adrich, B. Fournet, C. Capon, J.J. Bonicel, J.D. De Caro and M. Rovery, N-terminal sequence extension in the glycosylated forms of human pancreatic stone protein. The 5-oxoproline N-terminal chain is O- glycosylated on the 5th amino acid residue, Biochim Biophys Acta 994 (1989) 281-4.

[54]H. Okamoto and S. Takasawa, Recent advances in the Okamoto model: the CD38-cyclic ADP-ribose signal system and the regenerating gene protein (Reg)-Reg receptor system in beta-cells, Diabetes 51 Suppl 3 (2002) S462-73.

Figure 1

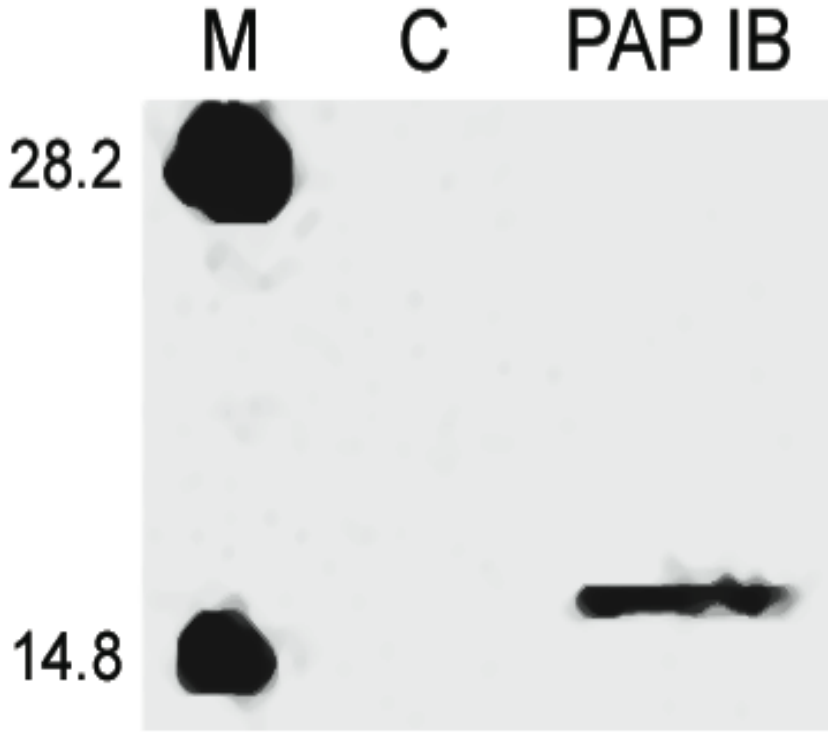


FIGURE 2

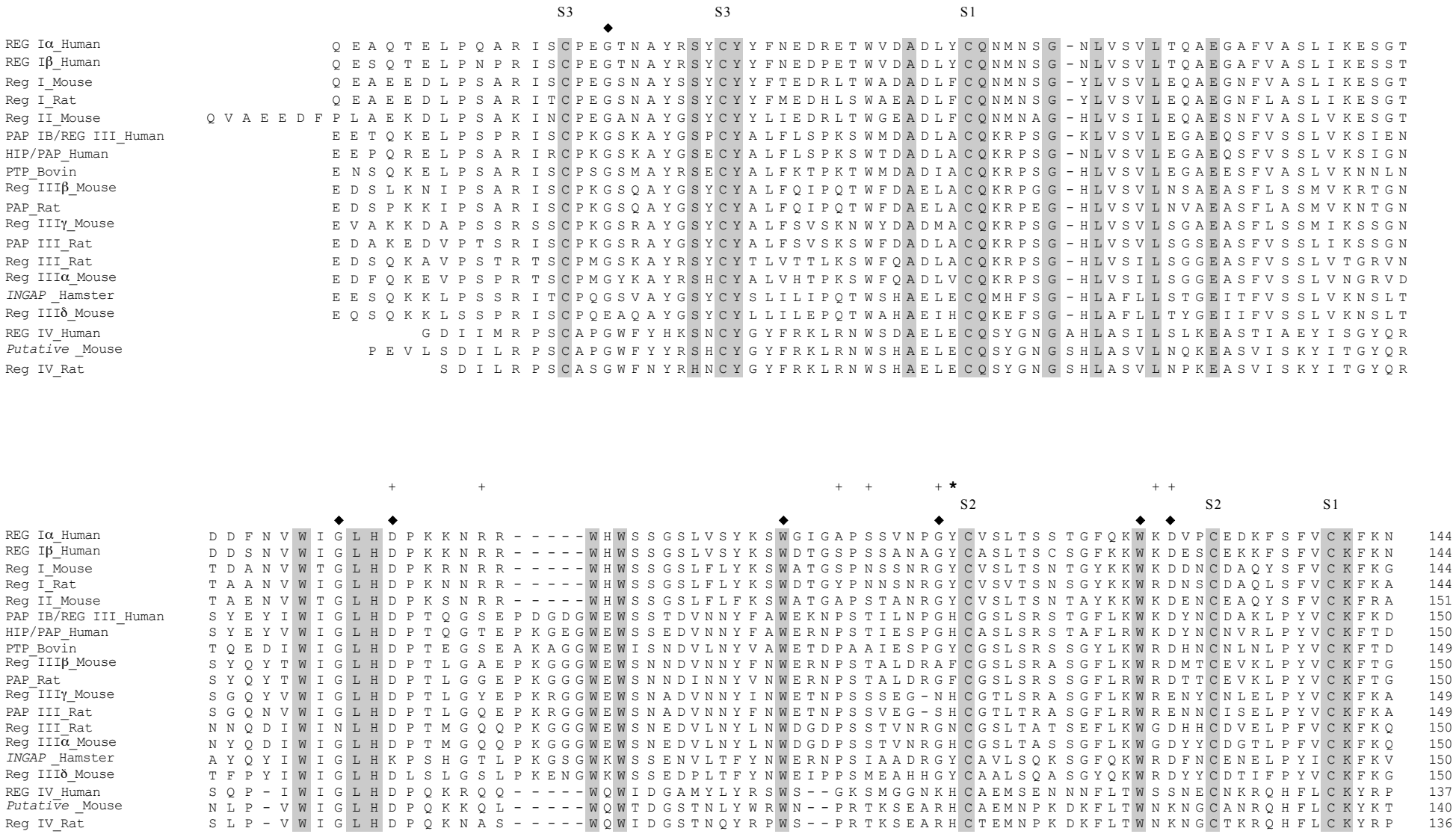

S3
S3
S1
S2
S2
S1
REG Iα_Human
REG Iβ_Human
Reg I_Mouse
Reg I_Rat
Reg II_Mouse
PAP IB/REG III_Human
HIP/PAP_Human
PTP_Bovin
Reg IIIβ_Mouse
PAP_Rat
Reg IIIγ_Mouse
PAP III_Rat
Reg III_Rat
Reg IIIα_Mouse
INGAP _Hamster
Reg IIIδ_Mouse
REG IV_Human
Putative _Mouse
Reg IV_Rat

Figure 3

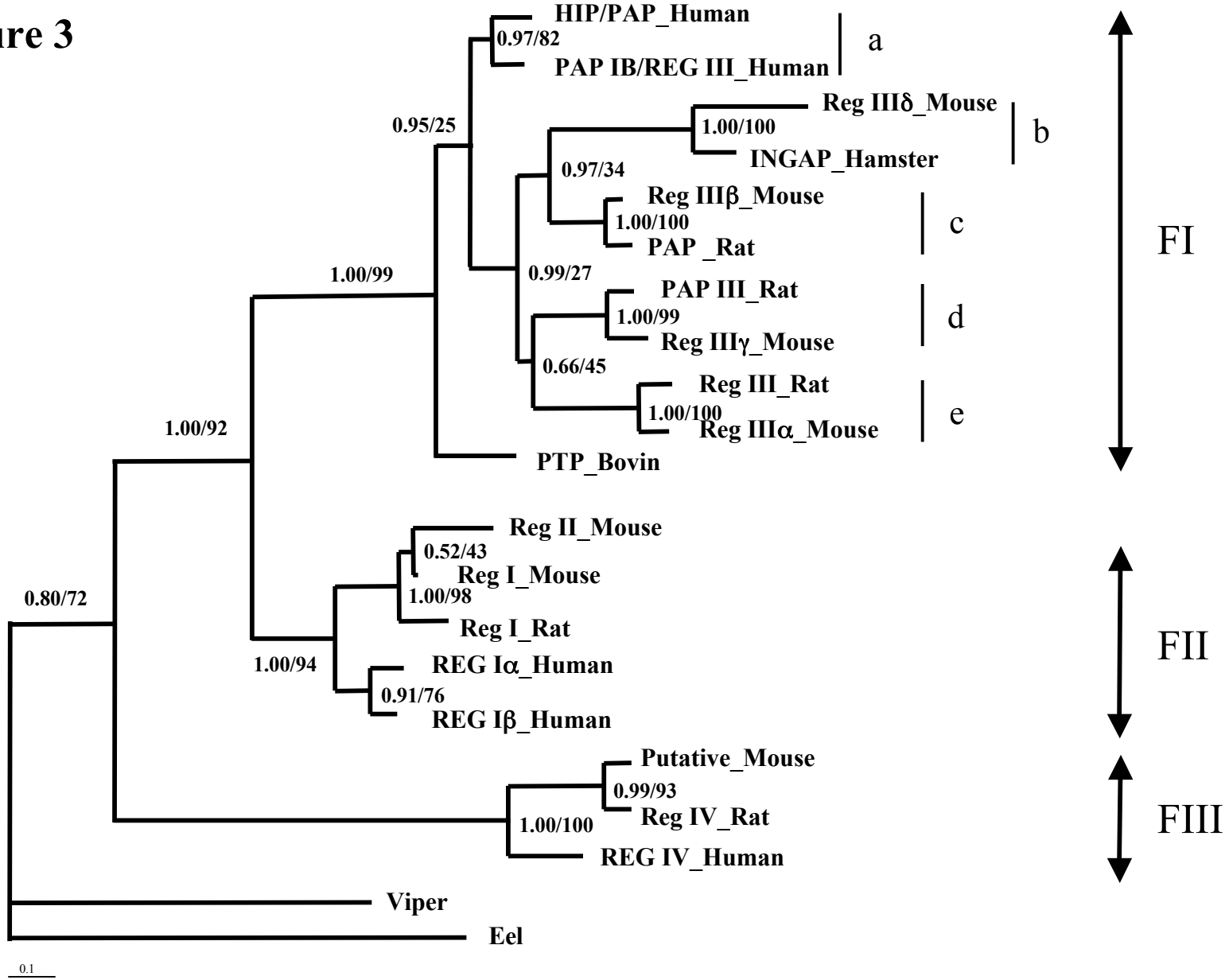

HIP/PAP_Human
0.97/82
PAP IB/REG III_Human
a
Reg IIIδ_Mouse
1.00/100
INGAP_Hamster
b
0.95/25
0.97/34
Reg IIIβ_Mouse
1.00/100
PAP _Rat
c
1.00/99
0.99/27
PAP III_Rat
1.00/99
Reg IIIγ_Mouse
d
0.66/45
Reg III_Rat
1.00/100
Reg IIIα_Mouse
e
1.00/92
PTP_Bovin
FI
Reg II_Mouse
0.52/43
Reg I_Mouse
0.80/72
1.00/98
Reg I_Rat
1.00/94
REG Iα_Human
0.91/76
REG Iβ_Human
FII
Putative_Mouse
0.99/93
Reg IV_Rat
1.00/100
REG IV_Human
FIII
Viper
Eel
0.1

Figure 4

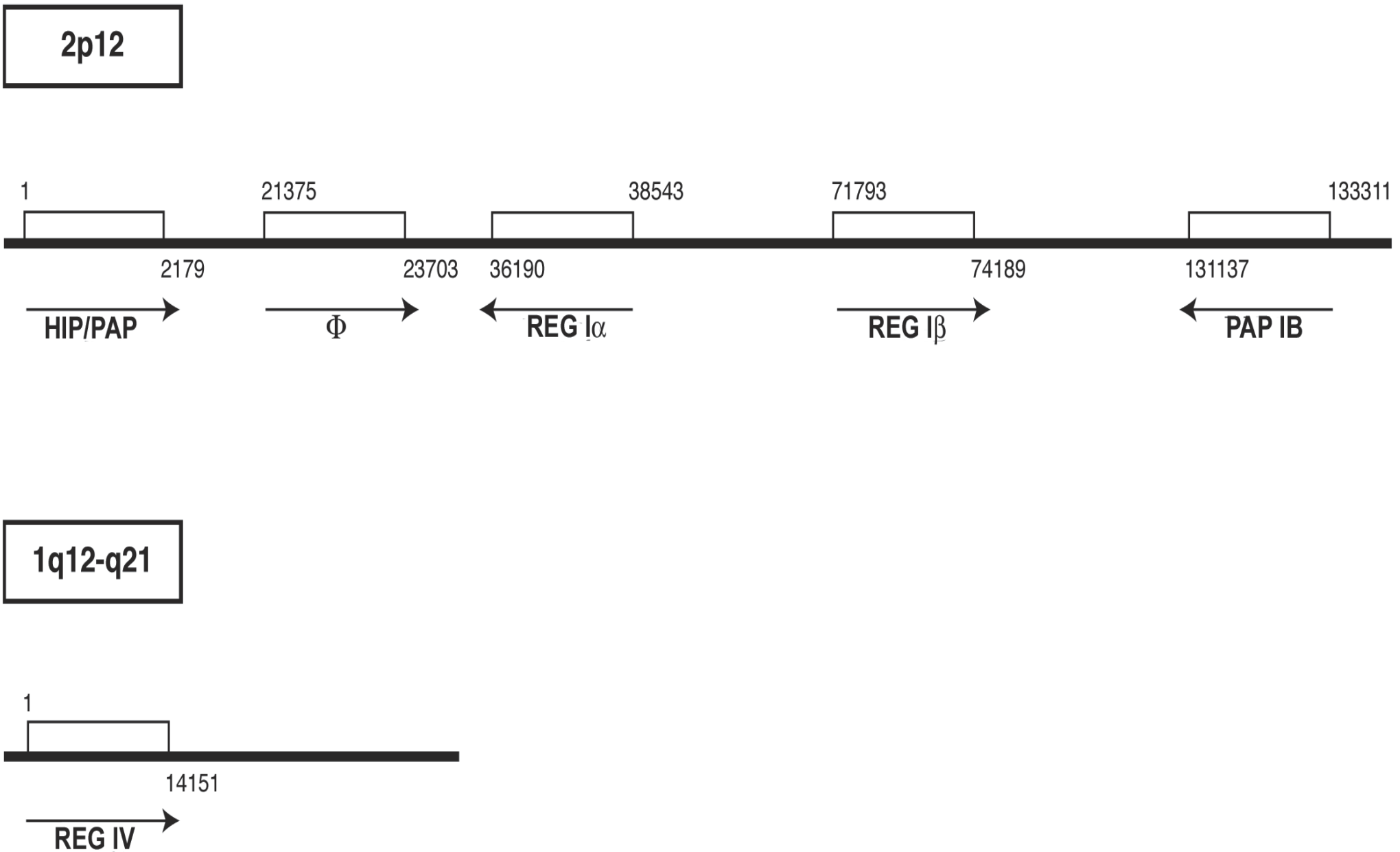

2p12
1
21375
38543
71793
133311
2179
23703
36190
74189
131137
HIP/PAP
Φ
REG Iα
REG Iβ
PAP IB
1q12-q21
1
14151
REG IV

Fig. 5

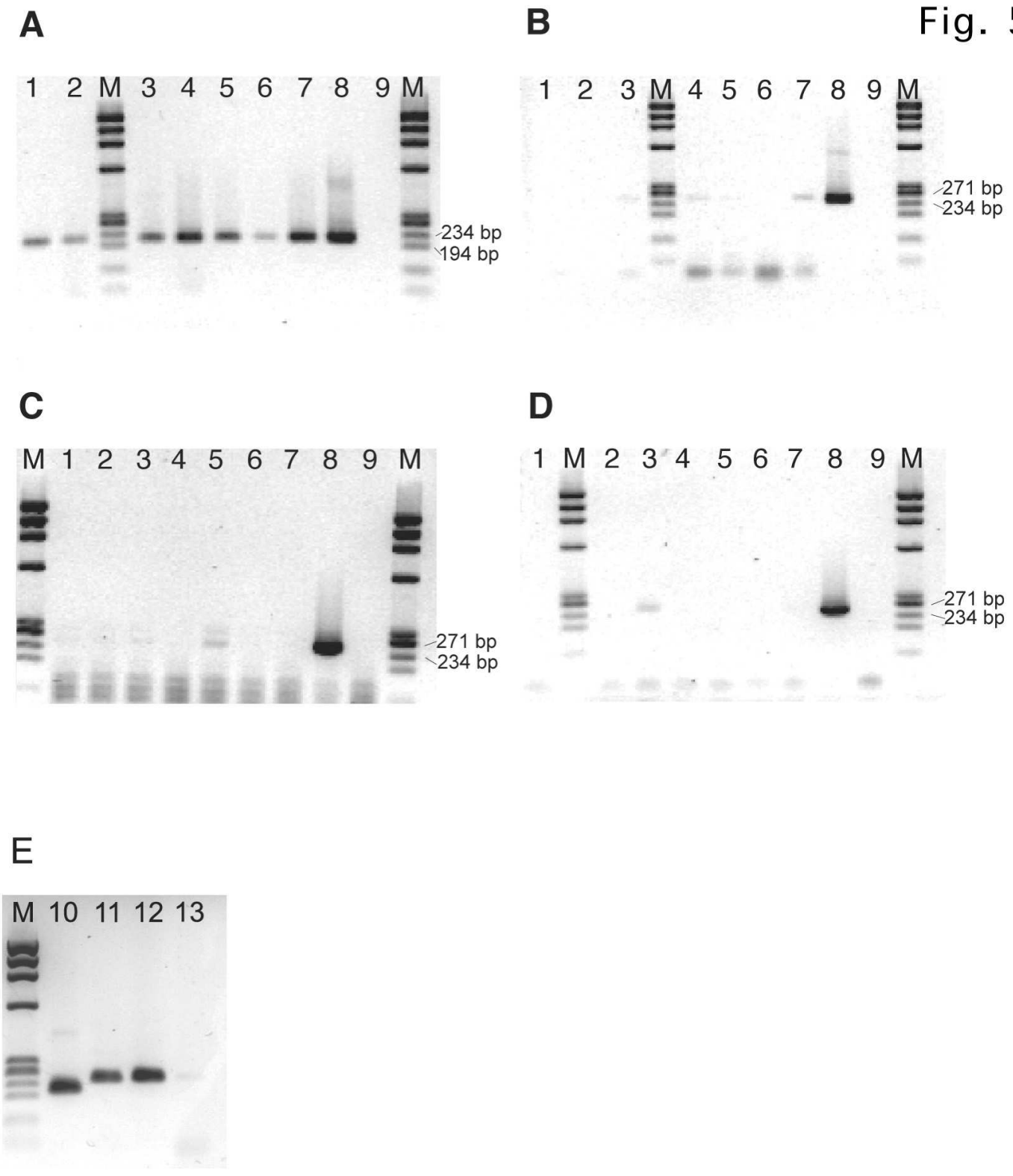
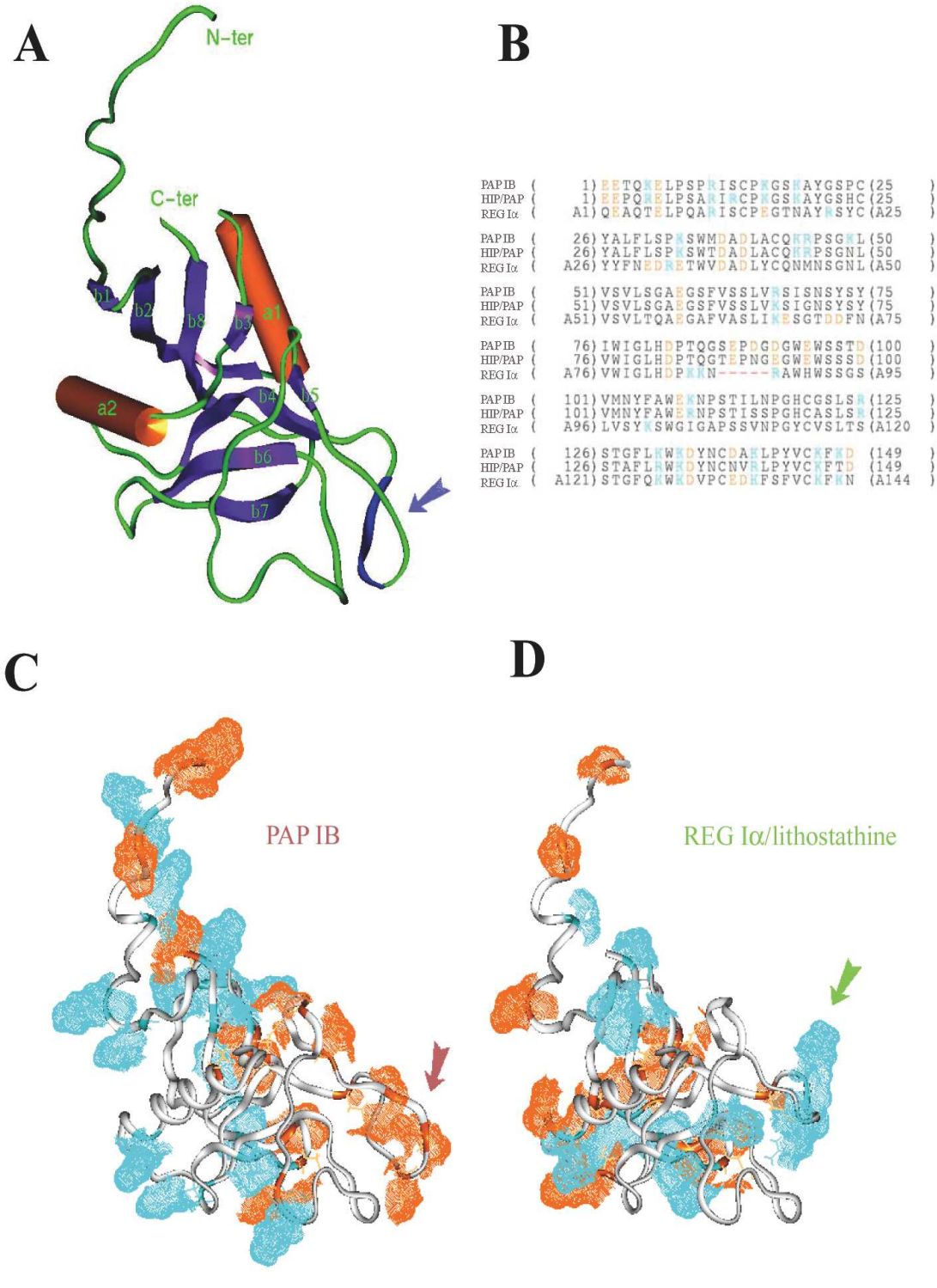

A
1 2 M 3 4 5 6 7 8 9 M
234 bp
194 bp
B
1 2 3 M 4 5 6 7 8 9 M
271 bp
234 bp
C
M 1 2 3 4 5 6 7 8 9 M
271 bp
234 bp
D
1 M 2 3 4 5 6 7 8 9 M
271 bp
234 bp
E
M 10 11 12 13
271 bp
234 bp
194 bp

Figure 6


A
N-ter
C-ter
b1
b2
b8
b3
a1
a2
b4
b5
b6
b7
B
PAP IB (1) EETQKELPSPKISCPKGSKAYGSPC (25)
HIP/PAP (1) EEPQAELPSAKIKCPKGSKAYGSHC (25)
REG Iα (A1) QEAQTELPQAKISCPEGTNAYKSYC (A25)
PAP IB (26) YALFLSPKSWMDADLACQKRPSGKL (50)
HIP/PAP (26) YALFLSPKSWTDADLACQKRPSGNL (50)
REG Iα (A26) YYFNEDKETWVDADLYCQNMNSGNL (A50)
PAP IB (51) VSVLSGAEGSFVSSLVKSISNSYSY (75)
HIP/PAP (51) VSVLSGAEGSFVSSLVKSIGNSYSY (75)
REG Iα (A51) VSVLTQAEGAFVASLIKESGTDDFN (A75)
PAP IB (76) IWIGLHDPTQGSEPDGDGWEWSSTD (100)
HIP/PAP (76) VWIGLHDPTQGTEPNGEGWEWSSSD (100)
REG Iα (A76) VWIGLHDPKKN-----KAWHWSSGS (A95)
PAP IB (101) VMNYFAWEKNPSTILNPGHCGSLSK (125)
HIP/PAP (101) VMNYFAWEKNPSTISSPGHCASLSK (125)
REG Iα (A96) LVSYKSWGIGAPSSVNPGYCVSLTS (A120)
PAP IB (126) STGFLKWKDYNCDAKLPYVCKFKD (149)
HIP/PAP (126) STAFLRWKDYNCNVRLPYVCKFTD (149)
REG Iα (A121) STGFQKWKDVPCEDKFSFVCKFKN (A144)
C
PAP IB
D
REG Iα/lithostathine