Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Biologies 326 (2003) 339–348
COMPTES RENDUS BIOLOGIES

Genomics / Génomique

Short inverse complementary amino acid sequences generate protein complexity

Les courts inverses complémentaires de séquences d'aminoacides participent à la complexité des protéines

Daniel J. Goldstein [a,*], Christian Fondrat [b], Florence Muri [c,d], Gregory Nuel [c], Patricia Saragueta [e], Anne-Sophie Tocquet [c], Bernard Prum [c,*]

[a] *Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina*
[b] *Direction des systèmes d'information, université René-Descartes (Paris-5), 12, rue de l'École-de-Médecine, 75270 Paris cedex 06, France*
[c] *Laboratoire « Statistique et Génome », Upresa CNRS 8071 et département de mathématique, Génopole, tour Évry-2, 2e étage, 523, place des Terrasses-de-l'Agora, 91000 Évry, France*
[d] *Département STID, Institut universitaire de technologie, université René-Descartes, Académie de Paris, 143, avenue de Versailles, 75016 Paris, France*
[e] *Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina*

Received 20 January 2003; accepted 11 March 2003

Presented by François Cuzin

Abstract

Inversions of short genomic sequences play a central role in the generation of protein complexity. More than half of the 1300 motifs registered in ProSite have protein inverse complementary sequences (princoms) among proteins registered in SwissProt. The observed number of princoms occurrences exceeds by far the expected number ($p < 10^{-10}$). Princoms often endow their host proteins with a whole new range of biochemical and physiological capabilities, including the possibility of intramolecular and intermolecular disulfide bond formation. These results support the idea that, like the duplications, the inversions of small genomic fragments have been a fundamental mechanism for shaping genomes. ***To cite this article: D.J. Goldstein et al., C. R. Biologies 326 (2003).***

Résumé

Le retournement de courtes séquences génomiques joue un rôle central dans la génération de la complexité des protéines. Plus de la moitié des inverses complémentaires des 1300 motifs de ProSite se retrouvent dans les protéines de SwissProt. Le nombre d'occurrences dépasse fortement le nombre attendu ($p < 10^{-10}$). Ces résultats renforcent l'idée que, comme les duplications,

* Corresponding author.
E-mail addresses: djgol@biolo.bg.fcen.uba.ar (D.J. Goldstein), prum@genopole.cnrs.fr (B. Prum).

1631-0691/03/$ – see front matter
doi:10.1016/S1631-0691(03)00077-5

les inversions de courts segments génomiques ont été un mécanisme fondamental dans l'élaboration des protéines. ***Pour citer cet article : D.J. Goldstein et al., C. R. Biologies 326 (2003).***

Keywords: inverse complementary; genomic inversions; genomic complexity

Mots-clés : complémentarité inverse ; inversions génomiques ; complexité génomique

1. Introduction

Double helical DNA is made up by two complementary polynucleotide chains, Watson (W_s) and Crick (C_s) with opposing polarities. Messenger RNAs (mRNA, represented by arrows) are transcribed from both strands with a $5' \rightarrow 3'$ polarity.

$3' \leftarrow 5'$

$5'W_s \ \ldots \ W_s3'$

$3'C_s \ \ldots \ C_s5'$

$5' \rightarrow 3'$

Duplications with or without inversions of the duplicated DNA segments are common genetic phenomena. Inversions generate transposed segments $[C/W]^T$. Given a complementary double sequence $[W/C]_j$ in a supporting stand, $[W/C]_s$,

$W_s \ \ldots \ W_j \ \ldots \ W_s$

$C_s \ \ldots \ C_j \ \ldots \ C_s$

its inversion and reinsertion will generate a structure of the type

$W_s \ \ldots \ W_j \ \ldots \ C_j^T \ \ldots \ W_s$

$C_s \ \ldots \ C_j \ \ldots \ W_j^T \ \ldots \ C_s$

with $[C_j^T \ W_j^T]$ adjacent to or separated from $[W_j \ C_j]$. The obvious question is how $[C^T \ W^T]$ will be transcribed, i.e., from the Watson strand W_s or from the Crick strand C_s.

As an example, the sequence catgctgct on the Watson strand will face its complementary sequence on the Crick strand:

5′ catgctgct 3′

3′ gtacgacga 5′

In the case of an inversion and reinsertion, it will appear elsewhere in the genome

5′ agcagcatg 3′

3′ tcgtcgtac 5′

Contemporary accounts of gene structure assume that when a protein-encoding double helical DNA segment is inverted, the transposed C^T in W is transcribed with the polarity of C if it carries its own promoter or is driven by a W_s promoter. If the insertion of C_j^T is out of frame, the encoded polypeptide will be a different molecular species, with a length determined by the emergence of a non-sense triplet.

The transposed strand, W_j^T in C_s, codes for the protein inverse complementary sequence (princoms) of the protein coded by C_j. In a previous paper, we presented evidence that inverse complementary (i.c.) DNA-sequences are transcribed and translated in the direction of C_s, the supporting strand [1]. The proteins codified by C_j and W_j^T are princoms pairs. When a princoms pair coexists in the same transcriptional and translational unit, the gene of the protein will contain both inverse complementary (i.c.) sequences. Two i.c. DNA sequences may be found in different structural genes, either in the same chromosome or in another chromosome (of the same or of different species) – and therefore code for a princoms pair – these DNA sequences being phylogenetically related or not.

2. Methods

ProSite is a catalogue of patterns identified by sequence or profile (weight matrix). This paper is based on the analysis of the May 2000 release (release 38), which contains more than 1300 patterns. Since profiles allow for the detection of signatures in sequences with a high degree of divergence, we only worked with patterns. ProSite patterns are sequences of brackets, each

Table 1
Amino acids in the original sequence and the residue found in the princoms (I.C.)

aa	IC	aa	IC	aa	IC	aa	IC
F	EK	Y	IV	K	FL	T	CGRS
L	EKQ	H	MV	E	FL	C	AT
M	H	Q	L	S	RGAT	W	P
I	DNY	N	IV	P	RGW	R	APST
V	DNYH	D	IV	A	CGRS	G	APST

containing the list of the possible amino acids that can be found in a given position (e.g., [NF]), separated by a number d of gaps, each characterized by lower and upper boundary values (e.g., $a \leqslant d \leqslant b$).

The princoms of a ProSite pattern were obtained by replacing the amino acids within each bracket by the amino acids encoded by the i.c. of all their possible codons (see Table 1), and the order of the brackets and the gaps, inverted.

We asked how many of the ProSite patterns have princoms in the proteins registered in SwissProt. The probability of a bracket equals the sum of the frequencies of its letters (e.g., $p_{[\mathrm{KEIV}]} = p_{\mathrm{K}} + p_{\mathrm{E}} + p_{\mathrm{I}} + p_{\mathrm{V}}$). By this way, we take into account the fact that a given protein generally has more than one princoms.

The probability p of the occurrence of a sequence of brackets separated by gaps is the product of the probabilities of the brackets in the sequence, multiplied by the factor $b - a + 1$ for each gap of length d with $a \leqslant d \leqslant b$. Taking into account the number N of amino acids in SwissProt (80 000 entries with $N = 29\,085\,265$ amino acids), the expected number E of the observed number of occurrences, O, for this sequence is equal to the product Np. The standard deviation s.d. of O was calculated through a Poisson approximation. The statistical signification of the observed number of occurrences, O, was based on the statistic $(O - E)/\mathrm{s.d.}$ [2].

3. Results

To avoid very large files, we selected the patterns for which the expected number E of princoms occurrences does not exceed 100, the cut-off number 100 being arbitrary. This selection generated a set of 594 sequences. Out of these 594 motifs, the mathematical expectation of the number of occurrences greater than 1.96 s.d. is less than 30, and we observed 273 (Group A, $p < 10^{-10}$). The mathematical expectation of the number of occurrences greater than 5 s.d. is less than 2×10^{-4}, and we observed 93 (Group B, $p < 10^{-10}$). From these results, we conclude that the princoms that we reported earlier are a subset of a much larger set. This strongly suggests that princoms are a common feature in the proteome.

The amino acid composition of the Group-B patterns is significantly different from that of all ProSite entries (Table 2). Cysteine is almost 2.5 times more frequent in Group B than in the whole ProSite; arginine, glycine, lysine, alanine, and tryptophan almost 1.25 times. The over-representation of these amino acids could be explained by the fact that all of them can be generated by a single site mutation of the cysteine codon, while lysine is a habitual replacement of arginine. A similar behaviour could be expected from serine, whose codons can also be derived from those of cysteine by a one-letter change. The codons for cysteine, tryptophan, arginine and glycine, as well as one of the stop codons, all have a central guanine, and the frequency of central guanine in the codon usage of Group B (40.60%) is significantly higher than in the whole ProSite ($p < 10^{-18}$). In the case of serine, we estimated its frequency as 1/3 (Table 3).

Protein data banks have certain inherent characteristics that can lead to sampling biases – e.g., the arbitrary selection of proteins studied and reported, the inconsistencies of the annotation systems and the existence of numerous entries corresponding to the same protein in different species (say ‘redundancy’). Furthermore, the criteria for identification of the patterns registered in ProSite are not exhaustive.

We addressed the problem of redundancy by comparing the statistical signification of princoms data obtained from the whole SwissProt and that derived from four specialized data banks – Human proteins in SwissProt (5913 proteins), Yeast Protein Database

Table 2
Statistics and frequencies of individual amino acids in ProSite and ProSite*. The occurrence of each amino acid in the complete ProSite and in ProSite*A (the subset of Group A, the 93 ProSite **m** and **s** having a highly significant number of princoms). When a position is degenerate, generating a k-letter bracket, we assigned a value of $1/k$ to each member of the bracket. The last column corresponds to the ratio of the frequency in ProSite* with respect to the frequency in ProSite

	ProSite		Group B		Ratio
A	773.6	6.05%	46.2	5.09%	0.84
R	631.7	4.94%	55.5	6.11%	1.24
N	441.3	3.45%	24.6	2.71%	0.79
D	651.2	5.09%	40.6	4.47%	0.88
C	811.2	6.34%	142.0	15.64%	2.47
Q	275.6	2.15%	14.8	1.63%	0.76
E	532.1	4.16%	26.4	2.91%	0.70
G	1519.1	11.88%	133.1	14.66%	1.23
H	453.7	3.55%	28.5	3.14%	0.88
I	715.7	5.60%	36.7	4.04%	0.72
L	927.8	7.25%	52.5	5.78%	0.80
K	515.7	4.03%	45.4	5.00%	1.24
M	538.1	4.21%	25.5	2.81%	0.67
F	597.1	4.67%	36.2	3.99%	0.85
P	524.2	4.10%	29.5	3.25%	0.79
S	771.3	6.03%	51.8	5.70%	0.95
T	575.1	4.50%	33.1	3.65%	0.81
W	238.9	1.87%	20.8	2.29%	1.23
Y	479.9	3.75%	25.3	2.79%	0.74
V	815.3	6.38%	39.6	4.36%	0.68
Total	12789	100.00%	908	100.00%	

Table 3
Comparison of the nucleotide composition (a, t, g, c) of the whole SwissProt (column 1), ProSite (column 2), and Prosite*A

	SwissProt	ProSite	Group B
a	28.2%	24.3%	21.7%
t	24.6%	27.0%	27.4%
g	24.1%	27.4%	30.5%
c	23.1%	21.3%	20.4%
g + c	47.2%	48.7%	50.9%

(4531 proteins), an Enzyme sub-bank (1519 proteins), and a PDB sub-bank (2139). The statistical signification was comparable except in the case of the PDbank (Table 4) [3].

3.1. The biological signification of princoms

Do princoms play a functional role in proteins? We arbitrarily selected the first pattern in Group A, PS 01113. This pattern corresponds to the domain signature of C1q, a subunit of the C1 enzyme complex that activates the serum complement system. We found 38 princoms of the C1q motif in SwissProt, a number that significantly exceeds the expected number 16.40 ($p < 3 \times 10^{-8}$) (Table 5). These princoms are found in eukaryotes (animals and plants), prokaryotes and viruses.

The 38 host proteins containing princoms of C1q form a heterogeneous group, both structurally and functionally, which includes several types of intracellular and extracellular proteins. The intracellular proteins are nucleotide-binding proteins, DNA- and RNA-binding proteins, ribosomal proteins, and Ca^{2+}-binding proteins. The extracellular proteins are the constant region of an immunoglobulins heavy chain (IgG-1c), a von Willebrand factor (vWF), and the heme-transporter hemopexin (Hp). While the princoms of C1q in the intracellular proteins cannot be readily associated with any known function, those present in the IgG, vWF and Hp play biochemical roles.

- IgGc. The hinge region of the IgG-1c (P01868 and P01869), which includes the cysteine involved in

Table 4
Number of ProSite motifs having a number of princoms that exceeds the expected value by more than 1.96 and 5 s.d. All these results have a signification $p < 10^{-10}$, except the 37 for PDB/1.96 s.d., where $p = 0.10$

	SwissProt	Human	Yeast	Enzyme	PDB
(1.96 s.d.)	273	69	72	58	37
(5 s.d.)	93	19	16	8	1

Table 5
List of the proteins containing princoms of ProSite motif PS01113 (C1q). ProSite motif: $Fx_5[ND]x_4[FYWL]x_6Fx_5GxYxFx[FY]$. Princoms: $[EIKV]x[EK]x[IV]x[APST]x_5[EK]x_6[EIKPQV]x_4[IV]x_5[EK]$

P02997	*Escherichia coli*	ELONGATION FACTOR TS (EF-TS)
Q43894	*Haemophilus influenzae.*	ELONGATION FACTOR TS (EF-TS)
Q38913	*Saccharomyces cerevisiae*	FAD SYNTHETASE
P50907	*Wolbachia pipientis*	CELL DIVISION PROTEIN FTSZ
P45485	*Wolbachia sp*	CELL DIVISION PROTEIN FTSZ
Q10719	*Saccharomyces pombe*	CELL FUSION PROTEIN FUS1
P01868	*Mouse*	IG GAMMA-1 CHAIN C REGION
P01869	*Mouse*	IG GAMMA-1 CHAIN C REGION
P20058	*Rabbit*	HEMOPEXIN PRECURSOR
O29490	*Archaeoglobus fulgidus*	PROBABLE TRANSLATION IF-2
P38249	*Saccharomyces cerevisiae*	EUKARYOTIC TRANSLATION IF-3
P29681	*Drosophila melanogaster*	20-HYDROXYECDYSONE
Q06738	*Arabidopsis thaliana*	DESSICATION-RESPONSIVE PROTEIN
O23676	*Arabidopsis thaliana*	MAGO NASHI PROTEIN HOMOLOG
O51737	*Borrelia burgdorferi*	DNA MISMATCH REPAIR PROTEIN
P33238	*Domestic duck*	INTERFERON-INDUCED GTP-BINDING
Q90597	*Chicken*	INTERFERON-INDUCED GTP-BINDING
P33937	*Escherichia coli*	PERIPLASMIC NITRATE REDUCTASE PREC
P36608	*Caenorhabditis elegans*	NEURONAL CALCIUM SENSOR 1
Q09711	*Saccharomyces pombe*	HYPOTHETICAL CALCIUM-BINDING
Q08637	*Enterococcus hirae*	V-TYPE SODIUM ATP SYNTHASE
P27341	*Sulfolobus acidocaldarius*	TRANSCRIPTION ANTITERMINATION
Q42667	*Citrus limon*	PHENYLALANINE AMMONIA-LYASE
P05738	*Saccharomyces cerevisiae*	60S RIBOSOMAL PROTEIN L9-A
P51401	*Saccharomyces cerevisiae*	60S RIBOSOMAL PROTEIN L9-B
P48119	*Cyanophora paradoxa*	DNA-DIRECTED RNA POLYMERASE BETA
P12954	*Saccharomyces cerevisiae*	ATP-DEPENDENT DNA HELICASE
P45740	*Bacillus subtilis*	THIAMINE BIOSYNTHESIS
P20985	*Vaccinia virus*	PROTEIN A6
P33633	*Escherichia coli*	PROTEIN IN SRMB-UNG INTERGENIC
Q28295	*Dog*	VON WILLEBRAND FACTOR PRECURSOR
Q57624	*Methanococcus jannaschii*	GLUTAMYL-TRNA AMIDOTRANSFERASE
Q57692	*Methanococcus jannaschii*	HYPOTHETICAL PROTEIN MJ0240
O58012	*Pyrococcus horikoshii*	HYPOTHETICAL PROTEIN PH0274
Q57968	*Methanococcus jannaschii*	HYPOTHETICAL PROTEIN MJ0548
P57992	*Drosophila melanogaster*	YEMANUCLEIN-ALPHA
Q04693	*Drosophila melanogaster*	HYPOTHETICAL
P46327	*Bacillus subtilis*	HYPOTHETICAL

the formation of the heavy chain-light chain disulfide bond, is provided by a princoms of C1q. This particular princoms has remarkable similarities with several plant, insect, and vertebrate metallothioneins (Table 6).

- vWF. This protein (Q28295) belongs to a protein family endowed with a C-terminal cysteine knot (CTCK) [4]. Approximately a third of the vWF–CTCK is contributed by the princoms of C1q. This particular part of vWF presents a similarity

Table 6
List of proteins having a protein sequence highly homologous to the princoms contained in IgC-1c (P01868 and P01869)

P01868	*Mus musculus*	IG GAMMA-1 CHAIN C REGION
P01869	*Mus musculus*	IG GAMMA-1 CHAIN C REGION (MEMBRANE BOUND)
P20759	*Rattus norvegicus*	IG GAMMA-1 CHAIN C REGION
P20760	*Rattus norvegicus*	IG GAMMA-2A CHAIN C REGION
P01863	*Mus musculus*	IG GAMMA-2A CHAIN C REGION, AALLELE
P01865	*Mus musculus*	IG GAMMA-2A CHAIN C REGION, MEMBRANE-BOUND
P01857	*Homo sapiens*	IG GAMMA-1 CHAIN C REGION
P01870	*Oryctolagus cuniculus*	IG GAMMA CHAIN C REGION
P01859	*Homo sapiens*	IG GAMMA-2 CHAIN C REGION
P01860	*Homo sapiens*	IG GAMMA-3 CHAIN C REGION
P20761	*Rattus norvegicus*	IG GAMMA-2 CHAIN C REGION
P01862	*Cavia porcellus*	IG GAMMA-2 CHAIN C REGION
Q02223	*Homo sapiens*	B-CELL MATURATION PROTEIN
P01861	*Homo sapiens*	IG GAMMA-4 CHAIN C REGION
P15566	*Tachypleus gigas*	COAGULOGEN
P02681	*Tachypleus tridenta*	COAGULOGEN
P15265	*Mus musculus*	SPERM MITOCHONDRIAL CAPSULE SELENOPROTEIN
Q96353	*Brassica napus*	METALLOTHIONEIN-LIKE PROTEIN TYPE 2
P30570	*Triticum aestivum*	ZINC-METALLOTHIONEIN CLASS II
P30569	*Triticum aestivum*	ZINC-METALLOTHIONEIN CLASS II
Q40158	*Lycopersicon escule*	METALLOTHIONEIN-LIKE PROTEIN TYPE 2 B
P03997	*Carcinoscorpius*	COAGULOGEN
P02804	*Cricetulus griseus*	METALLOTHIONEIN-I
P01866	*Mus musculus*	IG GAMMA-2B CHAIN C REGION
P01867	*Mus musculus*	IG GAMMA-2B CHAIN C REGION
Q38805	*Arabidopsis thaliana*	METALLOTHIONEIN-LIKE PROTEIN 2B
P56168	*Brassica juncea*	METALLOTHIONEIN-LIKE PROTEIN TYPE 2
P56172	*Brassica juncea*	METALLOTHIONEIN-LIKE PROTEIN TYPE 2
Q39269	*Brassica rapa ssp. P*	METALLOTHIONEIN-LIKE PROTEIN
P56170	*Brassica juncea*	METALLOTHIONEIN-LIKE PROTEIN
P02803	*Rattus norvegicus*	METALLOTHIONEIN-LIKE PROTEIN
Q42258	*Arabidopsis thaliana*	EC PROTEIN HOMOLOG 3
P80290	*Oryctolagus cuniculus*	METALLOTHIONEIN-LIKE PROTEIN
P18055	*Oryctolagus cuniculus*	METALLOTHIONEIN-IIA
Q42377	*Arabidopsis thaliana*	EC PROTEIN HOMOLOG 2
P93746	*Arabidopsis thaliana*	EC PROTEIN HOMOLOG
P43390	*Actinidia chinensis*	METALLOTHIONEIN-LIKE PROTEIN TYPE 2
Q42494	*Brassica rapa*	METALLOTHIONEIN-LIKE PROTEIN TYPE 2
P25860	*Arabidopsis thaliana*	METALLOTHIONEIN-LIKE PROTEIN
P33654	*Streptomyces cacaoi*	HYPOTHETICAL 14.2 KDA PROTEIN
P43396	*Coffea arabica*	METALLOTHIONEIN-LIKE PROTEIN 1
P14425	*Stenella coeruleoalba*	METALLOTHIONEIN-II
P04459	*Gallus gallus*	KERATIN, SCALE
P18563	*Cavia porcellus*	INTEGRIN BETA-6
Q52106	*Acinetobacter calcoac.*	MERCURIC TRANSPORT PROTEIN
P41927	*Yarrowia lipolytica*	METALLOTHIONEIN-I
P11844	*Homo sapiens*	GAMMA CRYSTALLIN A
P20762	*Rattus norvegicus*	IG GAMMA-2C CHAIN C REGION
P15229	*Buthus sindicus*	SMALL TOXIN

Table 7
Rabbit hemopexin. Underlined: princoms. Bold: princoms and princoms* of C1q

PMVKASGIPIALGVWGLCWSLATVNSVPLTSAHGNVTEGESGTKPEADVIE
QCSDGWSFDATTLDDNGTMLFFKDEFVWKSHRGIREL**ISERWKNFIGPVD**
AAFRHGHTSVYLIKGDKVWVYTSEKNEKVYPKSLQDEFPGIPFPLDAAVE
CHRGECQDEGILFFQGNRKWFWDLTTGTKKERSWPAVGNCTSALRWLGRY
YCFQGNQFLRFNPVSGEVPPGYPLDVRDYFLSCPG*RGHR*SSHRNSTQHGH
ESTRCDPDLVLSAMVSDNHGATYVFSGSHYWRLDTNRDGWHSWP**IAHQWP**
QGPSTVDAAFSWEDKLYLIQDTKVYVFLTKGGYTLVNGYPKRLEKELGSP
PVISLEAVDAAFVCPGSSRLHIMAGRRLWWLDLKSGAQATWTELPWPHEK
VDGALCMEKPLGPNSCSTSGPNLYLIHGPNLYCYRHVDKLNAAKNLPQPQ

with several plant, insect, and vertebrate metallothioneins (data not shown).

- Hp. This protein (P20058) consists of a single polypeptide chain divided in two similar domains, the probable result of a duplication of an ancestral gene. The two Hp domains, in positions 32 to 235 and 239 to 460, and separated by a four residue hinge, share about 25% sequence similarity and the same 3D structure [5–8]. The princoms of C1q lie in the amino terminal domain of rabbit Hp (position 125–155), where it provides the metal-binding histidine in position 152. The C-terminal domain also has a trace of the princoms of C1q, (princoms* of C1q), in positions 334–365. Surprisingly, the princoms and the princoms* of C1q present in Hp have partial similarity with the sequence of the characteristic ProSite pattern of Hp, [LIVMFY]–[DENQS]–[STA]–[AV]–[LIVMFY], the polypeptide segment between the sequence corresponding to the ProSite pattern Hp and the princoms of C1q (Table 7).

These data suggest that each Hp domain itself is the result of the duplication of a smaller ancestor gene. If this were so, there should be other vestiges of the princoms of C1q. We found these traces in the N-terminal domain (positions 88–103) and in the C-terminal domain (positions 295–310). As expected, similar princoms* of C1q can be detected in rat, pig, and human hemopexins, and in several matrix metalloproteinases which contain an Hp domain. However, these sequences are also found in the extracellular domain of the γ-aminobutyric acid (GABA) receptors GAB1 (human, bovin), GAB2 (human, mouse), GAB3 (human, mouse, chick), and a hypothetical protein [P40882] of *Pseudomonas aeruginosa*.

3.2. Apoptotic proteins and the hemoglobins

The proteins belonging to the Bcl family of apoptosis regulators have four domains, BH1, BH2, BH3, and BH4, each of them characterized by a consensus pattern. Bcl-2 and Bcl-x block apoptosis, and Bax, Bak, and the BH3-only proteins are proapoptotic.

There are 709 princoms of the ProSite BH-2 pattern (PS 01258) in SwissProt. This number of princoms of BH-2 exceeds by far the expected number $E = 357 \pm 18.9$ ($p < 10^{-50}$). One of the BH-2 princoms is found in positions 44–55 of the mouse BCLX (Q64373) and in equivalent positions of rat, pig, and human BCLX proteins. On the other hand, the 998 princoms of the BH-3 pattern (ProSite 01259) found in SwissProt barely surpasses the expected number $E = 907.8 \pm 30.1$, and the statistical signification is very poor (1%). A closer analysis of these results indicates that one of the princoms of BH-3 is the segment 37–51 of the Rana catesbiana hemoglobin β-chain [9]. This sequence comprises helices C and D; four of its amino acids (phenylalanine 43, phenylalanine 44, leucine 48, and leucine 57) are highly conserved in all hemoglobin α- and β-chains, and the myoglobins. This shows that the number of princoms of BH-3 is much higher than that detected by our program, which did not detect those princoms of BH-3 present in hemoglobins and myoglobins when in position 48 there is a methionine instead of [VI], and in position 44 when a glycine appears instead [PSAT].

The determination of the three dimensional (3D) structures of several apoptotic proteins led to the realization that they share together with the membrane spanning domain of colicins and the diphtheria toxins the myoglobin fold. Although the 3D structure of the Rana catesbiana hemoglobin has not yet been directly determined, it may be safely assumed that its β-

chain shares the same fold of the rest of hemoglobin β-chains. The segment 43–57 (princoms of BH-3) adopts a helical secondary structure in the β-chains of human, bovine, equine, and avian hemoglobins. Although the segment is predominantly α-helical, in some species it also has a short 3-helical (3–10) stretch, generally separated from the α-helix by an hydrogen-bounded turn. In Bcl-x, the fifteen amino acids that form the BH3 domain signature (86–100) also form an α-helix.

4. Discussion

Duplications and inversions are characteristic genomic features, and play a central role in the evolution of chromosomal architecture. Large size, low-copy repeats with high-sequence identity (several kb to Mb duplicons) lead to deletions, duplications, inversions, and inverted duplications. Contemporary mosaic proteins are often the result of the iteration of small-size genetic domains (duplicated segments of up to 1 kb). Many of these iterated domains preserve their characteristic sequence patterns motifs and signatures as well as their 3D structure. A substantial amount of motifs and signatures are cysteine C-rich, and the constancy of the positions of cysteine allows the classification of proteins in families and superfamilies, and to identify new members belonging to them [10,11].

While studying the patters of cysteine signatures in several families of autacoid peptides, we became aware of the fact that in the precursor polypeptides cysteine-rich regions alternate with threonine and/or alanine-rich regions. This clustered distribution of cysteine, threonines, and alanines is also found in vertebrate and invertebrate membrane glycoproteins, mucins, metalloenzymes of the extracellular matrix, proteoglycans, DNA-binding proteins, nuclear membrane proteins, and viral capsides. Since threonine and alanine are encoded by the inverse complementary codons of cysteine, we asked whether the threonine/alanine-rich regions were in fact the result of inversions of duplicated cysteine-rich domains [1]. To answer this question, we applied Hidden Markov Models (HMM) in the statistical tool R'HOM [12,13] to study the DNA encoding the threonine/alanine-rich regions flanking the three cysteine-rich trefoil patterns present in two small proteins MUA1-XENLA [sw P10667] and MUC1-XENLA [sw Q05049]. Our results showed that the cysteines and these threonine/alanine-rich regions actually are princoms pairs. These trefoil peptides can be described, therefore, as mosaics made up by the linear combination of direct and inverse gene segments. The analysis of the amino acid sequences of other peptides containing cysteine signatures revealed that the also have princoms pairs, e.g., the prepropeptides of six endothelins, and the Zn^{2+} finger proteins of the classes 1, 2, 4, 4^{*} (half of the type 4 signature), and 5 knots. In other cases, the princoms pairs are found in different polypeptides: the i.c. sequence of the cysteine signature of the somatomedins is present in 39 different proteins, but not in the somatomedin prepropetides.

In this paper, we provide evidence that the princoms pairs reported in our previous paper are just a small subset of a larger universe set of primcoms inserted in contemporary host-proteins registered in SwissProt. Our results are not due to biases introduced by the inherent characteristics of the protein data banks (problems of annotation and redundancy), because essentially the same results are obtained with different protein data banks. From this date, we conclude that many proteins are mosaics composed by direct and i.c. sequences (princoms pairs). Our present results allow us to generalize these findings and to postulate that many proteins contain sequences that are the princoms of known ProSite entries. Since we only analysed entries in protein data bank, we do not know if there are traces of princoms in intergenic regions. However, there is evidence that inverted segments are also translocated to non-coding, intergenic regions in the fish *Tetraodon nigrotiridis* (J. Weissenbach, personal communication). Furthermore, our data show that the role of genetic inversions in the determination of protein structure extends beyond the case of RAG2 and RAG1-mediated V/C recombinations to create antibody diversity. We have found examples in which the ancestral gene that has given rise to a multidomain protein by *n*-plication is in fact the princoms of a sequence found in a different, totally unrelated kinds of proteins.

Furthermore, we show that princoms change significantly the biochemical and physiological characteristics of the grafted proteins by providing new opportunities for intra-molecular and inter-molecular bonding, thus conferring distinct biochemical and physi-

ological functions to their host-proteins. In the case of PS01113, one of the princoms of C1q provides the hinge of IgG heavy chain, another gives vWF its CTCK, and still other makes the heme-binding residues of Hp. In the case of Hp, the molecule itself is the result of the tetraplication of a primordial princoms of C1q. In fact, the princoms of C1q is the characteristic ProSite motif of Hp. The grafting of a princoms opens the possibility of substantial structural modifications of the host-protein, e.g., polypeptide length, stability, catalytic specificity, folding, and associativity. Princoms inserted in phase and devoid of non-sense codons do not disrupt the reading of the host-protein, but if they contain a non-sense codon they will cause premature end of translation. When inserted out of phase, princoms will shift the open reading frame of the protein, and the new reading frame will replaces previous stop signals and introduce a new stop. It is plausible that the grafting of a princoms could cause radical 3D changes or confer new catalytic profiles to the host-protein. They may contribute hinge regions and divide a single domain in two domains, and the hinge may contain target sequences for proteolytic processing. Finally, princoms may create new interactive surfaces leading to non-covalent homodimerization or heterodimerization. We do not know the actual size of the set of princoms. Since we limited our search to the i.c. sequences of the ProSite patterns, we do not know yet the real length of the duplicated and inverted sequences. Work is in progress to devise mathematical and computational tools to find the real princoms length.

Our findings give raise to several structural questions. We have shown that a princoms pair, one present in the hemoglobin β-chain and the other in domain BH-3 of apoptotic proteins, have essentially the same secondary (α-helical) structure. Is this a general phenomenon? Do all the highly similar princoms pairs have the same secondary structure in solution? If they do, will they conserve it when grafted in their host proteins, or do they adopt a new secondary structure as a function of the context provided by the host protein? In fact, the existence of highly similar princoms pairs in different types of proteins is an experiment of nature for testing the generality of the findings of Milnor and Kim [14] concerning the importance of context-dependent effects in protein folding.

Chromosomes are mosaics of ancestral and horizontally transmitted sequences [15]. Genomes evolve by acquiring new sequences through duplications, inversions, horizontal genetic transfers, transposition events, and rearrangements (duplications and inversions), and ulterior divergence [15–19]. The widespread occurrence of palindromes and their biological relevance in genetic regulation and RNA structure and function indicates that inversions played a crucial role in the diversification of the portfolio of biological opportunities at the polynucleotide level throughout evolution [20,21]. Princoms offer a new way for detecting generalized lateral transfer among kingdoms, taxa, and species. Lateral gene transfer is a significant mechanism in the evolution (diversification and speciation) of bacterial genomes, introducing traits of antibiotic resistance, virulence attributes, and metabolic properties, and accounts for the ability of bacteria to exploit new environments [22]. So far, the detection and identification of cases of lateral gene transfer in bacteria relays on the finding of unusually high degrees of similarity between the donor and the recipient strains, atypical base compositions, and patterns of codon usage bias, as well as the detection of vestiges of genetic elements involved in their transference and integration. However, this approach is restricted to cases in which the putative recipient and the donor species (or taxons) are known, and is prone to underestimate the actual number of transferred genes [15]. On the other hand, princoms allow the detection of small lateral transfers. Since the statistical methods used so far to draw protein phylogenetic trees do not take into consideration the horizontal transfer of dincoms and their corresponding princoms, new approaches are needed for inferring the historical patterns of protein evolution, including their estimated time of grafting. This would provide a more refined appraisal of the tempo of evolution and reflect the multiple sources of genetic material from which a given protein family derives its contemporary structure.

The existence of princoms pairs implies the existence of DNA inverse complementary sequences (dincoms pairs), and their corresponding RNA inverse complementary sequences (rincoms pairs). The existence of families of short, non-coding RNAs having the characteristics of rincoms has been recently reported [23–27]. Rincoms can generate self-folding sequences that can change alternative splicing targets,

give raise to anti-sense RNA and post-transcriptional gene silencing (PTGS) structures, form regulatory hairpins in primary transcripts and messenger RNAs and thus affect the expression of genes and the rate of protein synthesis.

References

[1] D.J. Goldstein, F. Muri, P. Saragueta, B. Prum, Inverse complementary homologues of short cysteine signatures, C. R. Acad. Sci. Paris, Ser. III 323 (2000) 167–172.

[2] G.W. Snedecor, W.G. Cochran, Statistical Methods, 7th ed., Iowa State University Press, Ames, Iowa, 1980.

[3] S. Pasek, Étude du nombre d'occurrences d'inverses complémentaires de motifs protéiques dans les banques de protéines, Mémoire IUP 2, Laboratoire « Statistique et Génome », Genopole, Évry, France, 2001.

[4] J.E. Sadler, Biochemistry and genetics of von Willebrand factor, Annu. Rev. Biochem. 67 (1998) 395–424.

[5] H.R. Faber, C.R. Groom, H.M. Baker, W.T. Morgan, A. Smith, E.N. Baker, 1.8-A crystal structure of the C-terminal domain of rabbit serum haemopexin, Structure 3 (1995) 551–559.

[6] W.T. Morgan, P. Muster, F. Tatum, S.M. Kao, J. Alam, A. Smith, Identification of the histidine residues of hemopexin that coordinate with heme-iron and of a receptor-binding region, J. Biol. Chem. 268 (1993) 6256–6262.

[7] E. Morgunova, A. Tuuttila, U. Bergmann, M. Isupov, Y. Lindqvist, G. Schneider, K. Tryggvason, Structure of human pro-matrix metalloproteinase-2: activation mechanism revealed, Science 284 (1999) 1667–1670.

[8] D. Wellner, K.C. Cheng, U. Mueller-Eberhard, N-terminal amino acid sequences of the hemopexins from chicken, rat and rabbit, Biochem. Biophys. Res. Commun. 155 (1988) 622–625.

[9] L.T. Tam, G.P. Gray, A.F. Riggs, The hemoglobins of the bullfrog Rana catesbeiana. The structure of the beta chain of component C and the role of the alpha chain in the formation of intermolecular disulfide bonds, J. Biol. Chem. 261 (1986) 8290–8294.

[10] C. Brandon, J. Tooze, Introduction to Protein Structure, 2nd edn., Garland Publishing Inc., New York, 1999.

[11] T.E. Creighton, Proteins: Structures and Molecular Properties, 2nd edn., W.H. Freeman and Company, New York, NY, 1993.

[12] L. Bize, F. Muri, F. Samson, F. Rodolphe, S.D. Ehrlich, B. Prum, P. Bessières, in: S. Istrail, P. Pevzner, M. Waterman (Eds.), Recomb99, Proc. 3rd Annu. Int. Conf. Comp. Mol. Biol., ACM Press, New York, 1999.

[13] F. Muri-Majoube, B. Prum, Une approche statistique de l'analyse des génomes, La Gazette des Mathématiciens 89 (2001) 63 & 98.

[14] D.L. Milnor, P.S. Kim, Context-dependent secondary structure formation of a designed protein sequence, Nature 300 (1996) 730–734.

[15] H. Ochman, J.G. Lawrence, E.A. Groisman, Lateral gene transfer and the nature of bacterial innovation, Nature 405 (2000) 299–304.

[16] R. Jain, M.C. Rivera, J.A. Lake, Horizontal gene transfer among genomes: the complexity hypothesis, Proc. Natl Acad. Sci. 96 (1999) 3801–3806.

[17] T. Komano, Shufflons: multiple inversion systems and integrons, Annu. Rev. Genet. 33 (1999) 171–191.

[18] S. Ohno, Evolution by Gene Duplication, Springer-Verlag, New York, 1970.

[19] S. Ohno, The notion of primordial building blocks in constructing genes and transcriptional and processing errors due to the random occurrence of oligonucleotide signal sequences, Adv. Exp. Med. Biol. 190 (1985) 627–636.

[20] D. Romero, C. Palacios, Gene amplification and genomic plasticity in prokaryotes, Annu. Rev. Genet. 31 (1997) 91–111.

[21] D. Romero, J. Martinez-Salazar, E. Ortiz, C. Rodriguez, E. Valencia-Morales, Repeated sequences in bacterial chromosomes and plasmids: a glimpse from sequenced genomes, Res. Microbiol. 150 (1999) 735–743.

[22] J.G. Lawrence, H. Ochman, Molecular archeology of the *Escherichia coli* genome, Proc. Natl Acad. Sci. USA 95 (1998) 9413–9417.

[23] D.R. Eddy, Non-coding RNA genes and the modern RNA world, Nature Review/Genetics 2 (2001) 919–929.

[24] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, T. Tuschl, Identification of novel genes coding for small expressed RNAs, Science 294 (2001) 853–858.

[25] N.C. Lau, L.P. Lim, E.G. Weinstein, D.P. Bartel, An abundant class of tiny RNAs with propable regulatory roles in Caernohabditis elegans, Science 294 (2001) 858–862.

[26] R.C. Lee, V. Ambros, An extensive class of small RNAd in Caernohabditits elegans, Science 294 (2001) 862–864.

[27] G. Ruvkun, Glimpses of a tiny RNA world, Science 294 (2001) 797–799.