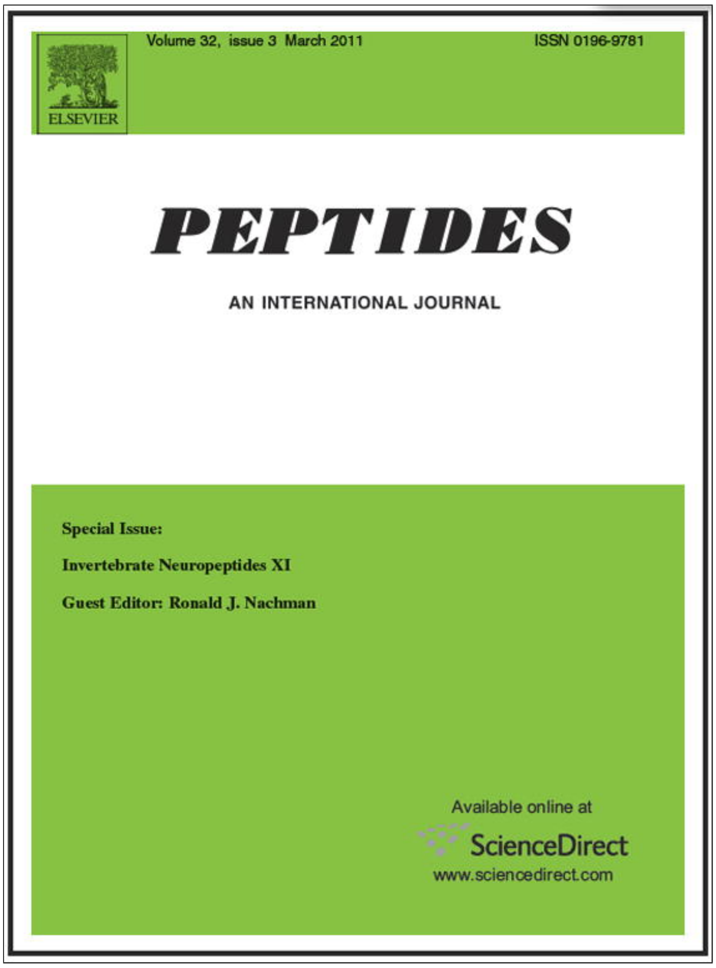

Volume 32, issue 3 March 2011

ISSN 0196-9781

ELSEVIER

PEPTIDES

AN INTERNATIONAL JOURNAL

Special Issue:

Invertebrate Neuropeptides XI

Guest Editor: Ronald J. Nachman

Available online at

ScienceDirect

www.sciencedirect.com

Peptides 32 (2011) 509–517

Contents lists available at ScienceDirect

Peptides

journal homepage: www.elsevier.com/locate/peptides

Influence of codon usage bias on FGLamide-allatostatin mRNA secondary structure

Francisco Martínez-Pérez[a,b], William G. Bendena[c], Belinda S.W. Chang[a,d,e], Stephen S. Tobe[a,*]

[a] Department of Cell and Systems Biology, University of Toronto, 110 St. George St., Toronto, ON M5S 3G5, Canada
[b] Department of Genetics and Molecular Biology, CINVESTAV, D.F., Mexico
[c] Department of Biology, Queen's University, Kingston, ON, Canada
[d] Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada
[e] Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON, Canada

ARTICLE INFO

Article history:
Received 28 July 2010
Received in revised form 6 October 2010
Accepted 6 October 2010
Available online 13 October 2010

Keywords:
mRNA
Secondary structure
Allatostatin
Cockroach
Diploptera
Codon bias

ABSTRACT

The FGLamide allatostatins (ASTs) are invertebrate neuropeptides which inhibit juvenile hormone biosynthesis in Dictyoptera and related orders. They also show myomodulatory activity. FGLamide AST nucleotide frequencies and codon bias were investigated with respect to possible effects on mRNA secondary structure. 367 putative FGLamide ASTs and their potential endoproteolytic cleavage sites were identified from 40 species of crustaceans, chelicerates and insects. Among these, 55% comprised only 11 amino acids. An FGLamide AST consensus was identified to be $(X)_{1\rightarrow16}$Y(S/A/N/G)FGLGKR, with a strong bias for the codons UUU encoding for Phe and AAA for Lys, which can form strong Watson–Crick pairing in all peptides analyzed. The physical distance between these codons favor a loop structure from Ser/Ala-Phe to Lys-Arg. Other loop and hairpin loops were also inferred from the codon frequencies in the N-terminal motif, and the first amino acids from the C-terminal motif, or the dibasic potential endoproteolytic cleavage site. Our results indicate that nucleotide frequencies and codon usage bias in FGLamide ASTs tend to favor mRNA folds in the codon sequence in the C-terminal active peptide core and at the dibasic potential endoproteolytic cleavage site.

1. Introduction

The database sequences for genomes, genes, messenger RNA (mRNA), and proteins provide rich sources of new information with which many regulatory elements and evolutionary relationships can be derived by *in silico* analysis. With this approach, it is now possible to propose a global and integrative view relative to physiological events. *In silico* analysis has uncovered 431 amino acid sequences of an arthropod neuropeptide family, the FGLamide allatostatins (ASTs). These include sequences uncovered from genome projects in 20 Insecta species, *Daphnia pulex* and the black-legged tick *Ixodes scapularis* [8,13,24,39] by inverse PCR and genomic libraries in 9 insect species [2,12] and with cDNA from *Bombyx mori* [33] and from four crustacean species ([42], GenBank: AB245091; EU000307; AB106899). Furthermore, FGLamide AST-like peptides have been proposed in other invertebrates [2,17,32,39].

FGLamide ASTs are inhibitors of juvenile hormone (JH) biosynthesis in cockroaches, termites and crickets but in other insect orders, FGLamide ASTs have no effect on JH production [23,34,40,41]. In the central and peripheral nervous systems of insects, FGLamide ASTs may also serve as neurotransmitters or neuromodulators [35]. Furthermore, following their release from endocrine cells in the gut, FGLamide ASTs have hormonal activity [34,4] and a potent inhibitory effect on spontaneous contractions of the gut of the blowfly [10]. In crustaceans, the FGLamide ASTs can stimulate production of selected intermediates in the JH biosynthetic pathway [22].

FGLamide AST genes show differences in the number and length of introns and exons [24]. FGLamide-AST mRNA encodes a precursor polypeptide with a signal peptide that directs the prepropeptide into the rough endoplasmic reticulum. Processing of the FGLamide-AST precursor is through recognition of potential endoproteolytic cleavage sites, K/R, R/R or R/K and carboxyl-terminal amidation that releases 4 or more active peptides [2,35–37]. Each FGLamide AST is different in sequence at the N-terminal region whereas the C-terminal motif (Y/F)/XFG(L/I/V)G is conserved in all FGLamide ASTs and is essential for biological activity [18,35].

Codon usage frequencies among FGLamide AST and associated potential endoproteolytic cleavage sites have not been extensively investigated. In related proteins, synonymous changes that alter

* Corresponding author. Tel.: +1 416 978 3517; fax: +1 416 978 3522.
E-mail address: stephen.tobe@utoronto.ca (S.S. Tobe).

0196-9781/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.peptides.2010.10.007

the nucleotides but are silent at the amino acid level can occur at different frequencies and this event has been termed ‘codon usage bias’ (for recent reviews, see [9,14,15]). In the cell, the presence or absence of synonymous codons for any amino acid improve the intron–exon distribution [19,26,29] and participate in mRNA folding to give rise to secondary structures [6,30]. mRNA secondary structure participates in splicing regulation [5,16], in the regulation of translation [1,20] and may regulate the expression of a variety of motifs in translation [26].

Codon usage bias in FGLamide ASTs and its relationship to mRNA folding has not been established. For example, the C-terminus motif has a Gly which can be encoded by four synonymous codons GGN, but the degree of codon usage bias, particularly in comparison with other FGLamide ASTs, remains to be investigated.

To determine the codon usage bias for FGLamide ASTs and its possible relationship with mRNA secondary structure, the amino acid composition and codon frequencies for all putative FGLamide ASTs currently found in the databases were determined. *In silico* analysis showed that some FGLamide AST mRNAs were folded by Watson–Crick pairing between the codon UUU (specifying Phe) from the core sequence of the C-terminal region and the codon AAA (specifying Lys) of the endoproteolytic cleavage site to create a loop structure. However, in the same region, with other codon frequencies, the FGLamide AST mRNAs folding gave rise to hairpin loops.

2. Materials and methods

2.1. FGLamide AST codons and amino acid data base

FGLamide AST cDNAs precursors were obtained from 40 species (2 Chelicerata, 5 Crustacea, and 33 Insecta) from the following databases: National Center for Biotechnology Information, Fly Base [38], Human Genome Sequencing Center, and the Joint Genome Institute (Table 1). FGLamide ASTs nucleotide and amino acid sequences were determined for each mRNA according to previously reported peptide sequences [17,24]. The database from FGLamide AST was constructed with the EXCEL program.

2.2. Determination of N and C terminal regions to FGLamide AST

To establish the codons and amino acid sequence for each N-terminal region, we considered that the C-terminal region [(Y/F)XFG(L/I/V)G] and the 2 amino acids creating the potential endoproteolytic cleavage sites for all FGLamide ASTs are invariant (8 amino acids in total). Therefore, the number of amino acids comprising the N-terminal region was calculated as the difference between the total amino acid number of each FGLamide AST minus 8. The concept is derived with the equation: *N terminal = AA to FGLamide AST* − *8*.

2.3. Codon usage and secondary structure prediction from FGLamide AST mRNA

Codon usage from each AST was determined with the Countcodon program [version 4] [27]. The means from codon frequency and distribution for each amino acid from each FGLamide AST were edited using EXCEL. mRNA secondary structures were computed using Mfold 2.3 program [25,43]. The parameters used in the calculation of minimum free energy structures and base-pairing probabilities for each FGLamide AST mRNA were according to the authors of the program.

3. Results

3.1. Amino acid number to FGLamide AST in the Arthropoda

FGLamide precursor sequences were analyzed from 40 species. This analysis revealed 367 putative FGLamide ASTs released by potential endoproteolytic cleavage sites (comprising 2 amino acids). Of these, 238 FGLamide-AST peptide sequences were found in Insecta, 122 in Crustacea, and 7 in Chelicerata. Collectively, the average peptide size is 11 amino acids. Including the 8 amino acid constant region, the smallest peptide of 9 amino acids was found in 19 Arthropods whereas among the longest sequences were a 21 amino acid putative peptide in Crustacea; a 17 amino acid putative peptide in Chelicerata and 18 peptides with 22 amino acids in Insecta. From the FGLamide AST analyzed, 55% showed 3 amino acids in the N-terminal region whereas 12%,

Table 1
FGLamide AST mRNA from Arthropoda obtained from databases.

	Species	Accession no.		Species	Accession no.
1	[a]*P. clarkii*	AB106899	21	[c]*P. americana*	X91029
2	[a]*P. interruptus*	AB245091	22	[c]*S. longipalpa*	AF0680639
3	[a]*M. rosenbergii*	DQ088626	23	[c]*R. flavipes*	FJ668632
4	[a]*C. finmarchicus*	EU000307	24	[c]*A. aegypti*	U66841
5	[a]*D. pulex*	[d]GNO_0600073	25	[c]*A. gambiae*	NW_045754 (3324916–3325667)
6	[b]*I. scapularis*	DS971562 (339806–340315)	26	[c]*C. quinquefasciatus*	DS232003 (299898–300658)
7	[b]*D. variabilis*	EU620228	27	[c]*C. vomitoria*	East et al. [11]
8	[c]*A. mellifera*	XM_001120780	28	[c]*L. cuprina*	East et al. [11]
9	[c]*N. vitripennis*	Martínez-Pérez et al. [24]	29	[c]*D. erecta*	XM_001981909 (1–456)
10	[c]*A. pisum*	Martínez-Pérez et al. [24]	30	[c]*D. melanogaster*	AF263923 (265–720)
11	[c]*H. armigera*	AF015296	31	[c]*D. mojavensis*	XM_001998495 (1–471)
12	[c]*B. mori*	AF309090	32	[c]*D. pseudoobscura*	XM_001357677
13	[c]*S. frugiperda*	AJ508906	33	[c]*D. sechellia*	XM_002032516 (1–456)
14	[c]*P. humanus*	DS235203 (13378–14310)	34	[c]*D. virilis*	XM_002052990 (1–471)
15	[c]*S. gregaria*	Z58819	35	[c]*D. willistoni*	XM_002073224 (1–474)
16	[c]*G. bimaculatus*	AJ302036	36	[c]*D. yakuba*	XM_002099154 (1–453)
17	[c]*B. craniifer*	AF068062	37	[c]*D. ananassae*	XM_001964412 (1–465)
18	[c]*B. germanica*	AF068061	38	[c]*D. grimshawi*	XM_001990019 (1–474)
19	[c]*B. orientalis*	AF068064	39	[c]*D. persimilis*	XM_002020181
20	[c]*D. punctata*	U00444	40	[c]*D. simulans*	XM_002104731

[a] Crustacea.
[b] Chelicerata.
[c] Insecta.
[d] Daphnia pulex V1.0.

Table 2
Amino acid number in the hypervariable region (Xn) of FGLamide AST sequence (X)n YXFGLGKR.

	Amino acids in the hypervariable region (Xn)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	16	Total
5 Crustacea	2	9	103	3	0	0	0	0	2	1	0	1	1	0	122
2 Arachnida	2	0	0	2	0	1	1	0	1	0	0	0	0	0	7
33 Insecta	15	12	99	21	15	10	6	8	2	2	21	1	12	14	238
	19	21	202	26	15	11	7	8	5	3	21	2	13	14	367

had 2 or 4 amino acids. The remainders were variable in number. None of the N-terminal sequences contained either Trp or Cys (Table 2).

3.2. FGLamide AST N and C-terminal regions and amino acid organization in Arthropoda

The first amino acid in the C-terminal core region is either Tyr (87.2%) or Phe (12%). The only exception to Phe/Tyr in this position is an Asp in the lobster *Panulirus interruptus* FGLamide AST26. The second position of the C-terminal core was the most variable with 11 different amino acids appearing at this position. Amino acid frequencies at this position included Ser (31%), Ala (25.9%), Asn (22.1%) and Gly (10%). Less frequently, Asp, Leu, Glu, Gln, His, Val and Pro were also found here. All FGLamide ASTs had Phe in the 3rd position and Gly in the 4th. In the 5th position, Leu was most frequent (94.8%) whereas Ile and Val were in low frequency. The Gly, necessary as an amidation signal, was invariable at the 6th position in all FGLamide ASTs. Examination of the potential endoproteolytic cleavage sites revealed that Lys in the 7th position was invariable in Crustacea and Chelicerata whereas in Insecta, the 7th position was variable with Lys (87%) and Arg (13%). Arg was greatly favored in the 8th position (96.4%) with Lys as the only alternative (Table 3).

3.3. Influence of synonymous codons in FGLamide AST mRNA folding

The amino acid analysis showed that the FGLamide AST consensus sequence, comprising the most frequent amino acids in all species analyzed, is $(X)_{1\rightarrow 16}$Y(S/A/N/G)FGLGKR. Within this consensus sequence, codons would have the potential to form Watson–Crick pairing at the RNA level. The synonymous codons GCG for Ala and TCT for Ser in the 2nd core position could potentially pair with the codons CGC and AGA specifying Arg at the 8th position. Independently, the codon TTT specifying Phe at the 3rd position has the potential to pair with codon AAA specifying Lys at the 7th position. Considering that mRNA secondary structure is determined by the sequence of synonymous codons and the distance between them, four putative loops could potentially be formed in FGLamide AST mRNA, based on the presence of the highest frequency amino acids: Ser, Ala, Asn and Gly at the 2nd position in the consensus (Fig. 1).

3.4. Synonymous codons preferred in FGLamide ASTs to encode the C-terminal motif

To establish the codons related to the putative mRNA secondary structure, the codon usage bias of the C-terminal core domain from

Table 3
Amino acid frequency for C terminal domain of FGLamide AST.

Position	AA	5 Crustacea	2 Chelicerata	33 Insecta	Arthropoda
1	Y	118	5	197	320
	F	3	2	41	46
	D	1	0	0	1
		122	7	238	367
2	S	33	1	80	114
	A	68	0	27	95
	N	7	2	72	81
	G	10	4	23	37
	D	2	0	25	27
	H	0	0	4	4
	L	0	0	2	2
	E	0	0	2	2
	Q	1	0	1	2
	P	0	0	2	2
	V	1	0	0	1
		122	7	238	367
3	F	122	7	238	367
4	G	122	7	238	367
5	L	115	7	226	348
	I	6	0	11	17
	V	1	0	1	2
		122	7	238	367
6	G	122	7	236	365
7	K	122	7	207	336
	R	0	0	31	31
		122	7	238	367
8	K	11	0	1	12
	R	110	7	237	354
		1	0	0	1
		122	7	238	367

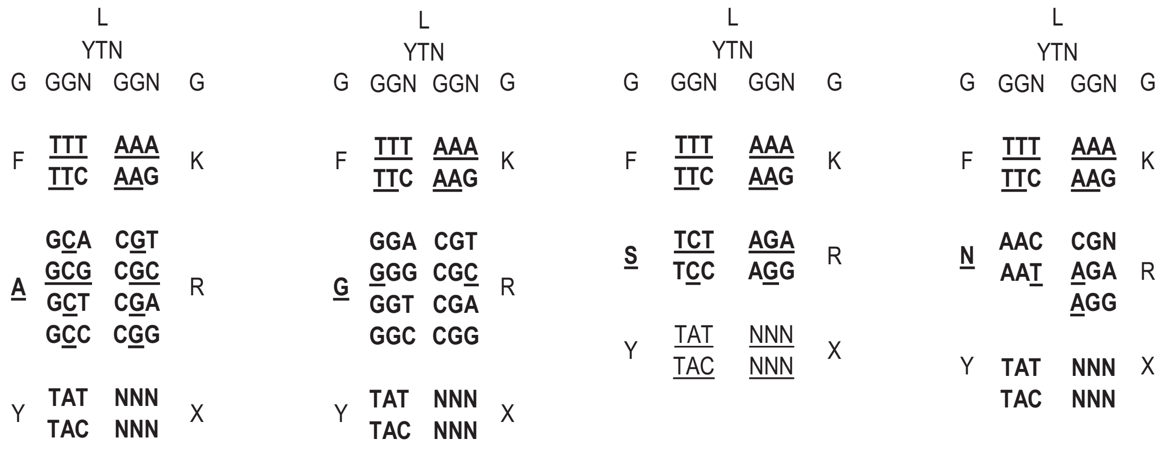

Fig. 1. Hydrogen bonds between nucleotides for the C terminal domain of FGLamides AST. Four putative loops can be obtained according to codon usage for C terminal domains. The underlined nucleotides show the putative hydrogen bonds.

FGLamide AST was obtained. As shown in Table 4, the codon GGC specifying Gly was the most used in all sequences. Chelicerata and Insecta had a similar codon usage frequency for Tyr and Leu as well as the Arg in the 8th position within the endoproteolytic cleavage site. In contrast, comparison between crustaceans and insects revealed the same codon but with different frequencies for Ser, Ala, Asn and Gly (2nd position), Phe (3rd position), and Lys (7th position).

The codon frequency of codons that potentially give rise to folded secondary mRNA structures showed that the codon AAA specifying Phe was the most frequent in relation to the codon TTT specifying Lys, whereas the codon GCG specifying Ala was less frequent in comparison to the codon CGC specifying Arg in all FGLamide AST mRNAs. Similarly, the codon TCT specifying Ser was less frequent in comparison with the codon AGA specifying Arg in crustaceans and insects (Table 4).

3.5. mRNA secondary structures in FGLamide ASTs in Arthropoda

The secondary structures of different FGLamide AST mRNA could potentially be obtained by pairing of codons in the C-terminal domain and endoproteolytic cleavage site. Watson–Crick pairing could create mRNA folding between codons specifying Ser-Phe and Gly-Lys. The codon TCT specifying Ser is found in the 2nd position of AST4 from *Procambarus clarkii* and the 8th position from *Blatta orientalis* providing the potential to fold with the conserved Gly-Lys. Similar loop structures could be created in FGLamide AST 1 from *P. interruptus* and *Drosophila persimelis* with the codons GCT and GCG specifying Ala respectively (Table 5).

Three additional RNA secondary structures were determined with the synonymous codons TCT specifying Ser and GCA, GCT, GCC and GCG specifying Ala in the 2nd position. In the first structure, hairpin loops in the FGLamide AST mRNA were predicted because the Watson–Crick pairing was between the codons specifying the N-terminal amino acids with the codons for Gly-Lys-Arg from the C-terminal domain. This mRNA fold was observed in FGLamide AST 11 from *P. americana*, AST 43 from *Macrobrachium rosenbergii*, AST 20 from *P. clarkii* and AST 4 from *Gryllus bimaculatus* (Table 5).

The second mRNA secondary structure created a loop, because the Watson–Crick pairing again utilized the first codons in the N-terminus and paired with codons specifying Tyr-Ala-Phe or Ala-Phe-Gly which were in the first amino acids of the C-terminal motif. This mRNA fold could potentially be made in FGLamide AST 16 from *M. rosenbergii*, AST 13 from *P. interruptus*, AST 21–22 from *P. clarkii* and AST 1 from *D. simulans* (Table 5). The third FGLamide AST RNA secondary structure had several combinations, with the codons GGT and GGC specifying the 2nd position Ala. Some mRNAs had the potential to create two hairpin loops; one loop was in the codons of the N-terminal motif and the other loop in the C-terminal domain. In other mRNAs, two loops could be generated using codons in the C-terminal motif. These conformations had the potential to form in FGLamide AST mRNA: AST 7 from *Acyrthosiphon pisum*, AST 15 from *P. interruptus* and FGLamide AST 1 from *D. melanogaster*, *D. sechellia* and *D. erecta* (Table 5).

All potential loops and hairpin loops found had different free energies since the mRNA fold was a function of the amino acid number in the N-terminal region (Fig. 2).

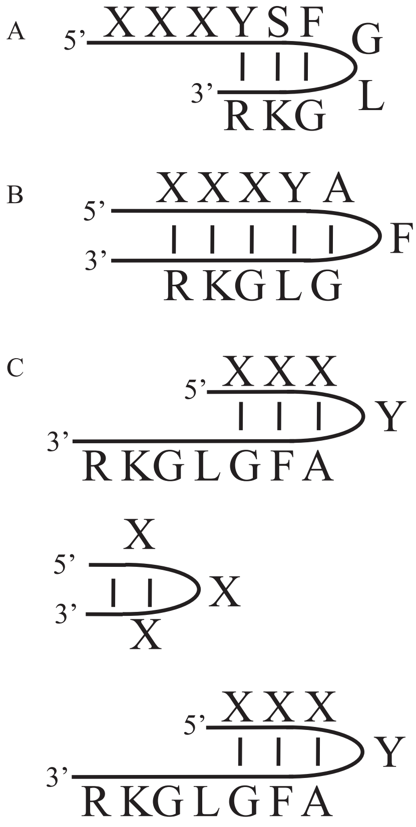

Fig. 2. FGLamide AST Secondary structures. The drawing represents the putative mRNA secondary structures. When codon bias to Tyr-Ser-Phe and Gly-Lys-Arg favor Watson–Crick pairing, a loop is obtained (A). The codons of the N-terminal motif favor different loops in the mRNA (B and C).

Table 4
Codon usage observed for C-terminal core of FGLamide AST.

Position	AA	FGLamide AST	Codon	Crustacea	Chelicerata	Insecta	Position	AA	FGLamide AST	Codon	Crustacea	Chelicerata	Insecta
1	Y	320	TAT	49.22	12.50	28.91	5	L	348	CTT	30.51	0.00	14.61
			TAC	45.98	70.83	53.39				CTC	8.75	29.17	11.51
	F	46	TTT	0.57	0.00	5.24				CTA	8.05	0.00	4.69
			TTC	3.78	16.67	12.46				CTG	19.73	58.33	43.94
	D	1	GAT	0.00	0.00	0.00			348	TTA	4.57	0.00	4.53
			GAC	0.44	0.00	0.00				TTG	8.40	12.50	17.58
2	S	80	**TCT**	2.78	0	3.79		I	17	ATT	8.57	0.00	1.09
			TCC	58.33	0	66.49				ATC	8.57	0.00	0.67
			TCA	21.30	0	0.33				ATA	0.00	0.00	1.15
			TCG	0.00	0	6.38		V	2	GTT	2.86	0.00	0.00
			AGC	10.19	100	1.27				GTC	0.00	0.00	0.00
			AGT	7.41	0	6.98				GTA	0.00	0.00	0.00
	A	95	GCT	18.80	0	12.90				GTG	0.00	0.00	0.22
			GCC	58.90	0	62.77	6	G	365	GGT	16.24	0.00	16.65
			GCA	20.78	0	0.07				GGC	49.11	70.83	55.53
			GCG	1.52	0	6.32				GGA	26.43	29.17	20.63
	N	76	AAC	66.67	100	89.03				GGG	8.22	0.00	7.19
			AAT	33.33	0	13.10	7	K	336	**AAA**	42.29	29.17	34.66
	G	37	GGT	20.00	12.5	0.53				AAG	57.71	70.83	44.50
			GGC	60.00	25	12.77		R	31	CGT	0.00	0.00	2.06
			GGA	20.00	12.5	0.27				**CGC**	0.00	0.00	0.00
			GGG	0.00	50	0.13				CGA	0.00	0.00	10.29
	D/H L/E Q/P V	79	NNN	N.D.	N.D.	N.D.				CGG	0.00	0.00	3.92
										AGA	0.00	0.00	0.59
										AGG	0.00	0.00	3.98
							8	K	12	**AAA**	5.46	0.00	0.00
3	F	367	**TTT**	57.46	41.67	55.55				AAG	1.02	0.00	0.22
			TTC	42.54	58.33	44.45		R	354	CGT	3.02	12.50	22.08
4	G	367.00	GGT	31.76	12.50	32.44				CGC	12.17	58.33	24.68
			GGC	31.02	70.83	30.03				**CGA**	11.14	0.00	15.41
			GGA	26.14	16.67	25.16				CGG	10.41	29.17	13.63
			GGG	11.08	0.00	12.37				**AGA**	43.25	0.00	11.87
										AGG	10.19	0.00	12.11

The codons related with the secondary structure are indicated by underline.

Table 5
Putative FGLamide AST RNA secondary structures.

Species and number FGLamide AST		Codon and amino acid sequence	Putative RNA secondary structure	Initial dG

P. clarkii 4	Ser-TCT	A D M Y S F G L G K R
		1 GCA GAC ATG TA**C** **TCT** **TTC** GGG CTG **GGA** **AAG** **AG**A 33
			10 20
			GCAGACAUGUA| GGG
			           CUCUUUC \\
			           GAGAAAG C
			A----------^ GGU
			            30
				−8.1

B. orientalis 8	Ser TCT	D R M Y S F G L G K R
		1 GAC AGA ATG TAT ***TCT*** T**TT** **G**GG CTA GG**C** **AA**G **AGA** 33
			10 20
			GACAGAAUGUAU| GGC
			           UCUUUUG \\
			           AGAGAAC U
			------------^ GGA
			              30
				−4.6

P. interruptus 1	Ala GCT	H N N Y A F G L G K R
		1 CAC AAC AAC TAT **GCT** TT**C** GGC CTC G**G**G **AAG** C**G**A
			10 20
			CACAACAACUA| GGC
			           UGCUUUC \\
			           GCGAAGG C
			A----------^ GCU
			            30
				−6.9

D. persimelis 1	Ala GCG	V E R Y **A** F G L G R R
		1 GTA GAA CGA TAT **GCG** TT**C** **GG**A TTG G**GC** **C**GT **CGC**
			10 20
			GUAGAACGAUAU| UU AU
			           GCG CGG U
			           CGC GCC G
			------------^ U- GG
			              30
				−6.2

P. americana 11	Ser TCT	S P Q G H R F S F G L G
		1 TCC C**CT** **C**AA **GGT** **CA**C **AGA** **TT**C TCT TTC **GGT** **CTT** **GGC**
		K R
		A**AG** **A**GA
			10 20
			UCCC| AAG C CUC
			    CUC GUCA AGAUU \\
			    GAG CGGU UCUGG U
			A---^ AA- - CUU
			    40 30
				−8.5

M. rosenbergii 43	Ala GCA	A G P Y A F G L G K R
		1 **GCG** GG**T** **CCC** TAC ***GC**A* TTT G**GC** CTT **GGA** AAA **CGT**
			10
			   GG-| CUAC AU
			GCG UCC GC \\
			UGC AGG CG U
			   AAA^ UUC- GU
			  30 20
				−5.8

P. clarkii 20	Ala GCC	T A G P Y A F G L G K R 1 A**CT** **GC**A GG**A** **CCT** TAT **GCC** TTT **GGT** CT**A** **GGT** AA**G** **CGG**	10 A| AGG UAU U CUGC ACCU GCC \\ GGCG UGGA UGG U -^ AA- UC- U 30	−10.20

G. bimculatus 4	Ala GCC	G P D H R F A F G L G K
		1 GG**G** **CCC** GA**C** **C**AC C**GG** TTC G**CC** TTC **GG**C CT**G** **GGC** AAG
		R
		CGG
			10
			GG---- A-| ACC U
			      GCCCG CC GG U
			      CGGGU GG CC C
			GGCGAA CC^ CUU G
			        30 20
				−11.8

M. rosenbergii 16Ala GCA

```
  P   R   H   Y   A   F   G   L   G   K   R
1 CCG AGG CAC TAT GCA TTT GGA CTG GGT AAG CGT
```

```
-------------     -| AC
             CCGAG GC  \
             GGUUU CG  U
UGCGAAUGGGUCA     A^ UA
  30        20
```

–5.6

P. interruptus 13 Ala GCT

```
  P   N   N   Y   A   F   G   L   G   K   R
1 CCA AAT AAC TAT GCT TTT GGA CTT GGC AAA CGA
```

```
                     10
-------------|    UAAC
             CCAAA    U
             GGUUU    A
AGCAAACGGUUCA^    UCGU
  30        20
```

–2.9

P. clarkii 21, 22 Ala GCC

```
  A   G   P   Y   A   F   G   L   G   K   R
1 GCT GGA CCT TAT GCC TTT GGT CTA GGT AAG CGG
```

```
                15
G-------|      UUAU
        CUGGACC    G
        GAUCUGG    C
GGCGAAUG^      UUUC
  30        20
```

–10.20

D.simulans 1   Ala GCC

```
  V   E   R   Y   A   F   G   L   G   R   R
1 GTG GAG CGG TAC GCC TTC GGC CTG GGA CGA CGG
```

```
                     10
------------| G   C  U
            GU GAG GG A
            CG CUU CC C
GGCAGCAGGGUC^ G   -  G
  30        20
```

–5.5

A.pisum 7   Ala GCT

```
  T   A   D   D   M   G   H   G   Q   R   F
1 ACG GCG GAC GAT ATG GGT CAC GGC CAG CGG TTC

 A   F   G   L   G   K   R
GCT TTC GGG CTG GGC AAA CGA
```

```
.-A| GC
   CG  \
   GC  G
\ -^ AG
  10

         20            30
  AUAUGG   A     ----    G
        GUC CGGCC    AGCG \
        CGG GUCGG    UCGC U
  AGCAAA   -     GCUU    U
     50          40
```

–14.3

P. interruptus 15 Ala GCC

```
  S   R   Q   Y   A   F   G   L   G   K   R
1 TCT CGA CAG TAC GCC TTC GGC CTC GGT AAA CGG
```

```
         10
.-UCUCGACAGUAC|  U
              GCC \
              CGG U
\ ------------^  C
              20
  CU  GU
    CG  \
    GC  A
  G-  AA
     30
```

–2.7

Table 5 (*Continued*)

Species and number FGLamide AST	Codon and amino acid sequence	Putative RNA secondary structure	Initial dG
D. melanogaster 1*D. sechellia* 1*D. erecta* 1Ala GCC	V E R Y A F G L G R R 1 GTG GAG CGG TAC GCC TTC GGT CTG GGA CGA CGG	.-GU -| G GGA GCG \\ CUU CGC U \\ -- C^ A 20 G- UG GUC G CAG G GG CA 30	−5.2

4. Discussion

This is the first study in which the codon usage in FGLamide ASTs was investigated, to establish potential influence on the mRNA secondary structure. Codon bias in prokaryotic and eukaryotic cells can contribute to mRNA folding and secondary structure, which can be critical to regulating expression [6,30]. Codon interactions can bring distant, complementary sections of one RNA molecule into close spatial proximity [21,26].

Three lines of evidence underlie our proposed FGLamide AST secondary structure model: (1) this is the first report that considered the influence of the doublet of basic amino acid combination Lys or Arg which are potential endoproteolytic cleavage site in the FGLamide AST precursor. (2) Using this approach, we observed that synonymous codons UUU and UUC that specify Phe showed Watson–Crick pairing with the codons AAA and AAG that specify Lys; likewise, the codons GCN specifying Ala and UCU specifying Ser paired with the codons CGN and AGA specifying Arg. These amino acids are present in the majority of the FGLamide AST precursors reported previously [35]. (3) We have used the position and distance from the last amino acids in the FGLamide precursor in our analysis. In this way, Ser/Ala-Phe and Lys-Arg are separated by the last amino acids from the active peptide and from the Gly necessary for peptide amidation. The codon usage data support the hypothesis that the FGLamide AST mRNA could have loop structures, based on synonymous codons located within the C-terminal motif and the endoproteolytic cleavage site.

Genomic and cDNA data from Crustacea and Insecta have shown that all FGLamide ASTs arise from a polypeptide precursor with similar organization but differing number of peptides [17,24,35]. There are 367 FGLamide AST peptides in precursors from 40 species reported in databases. In contrast to other neuropeptides such as melanocortins, cholecystokinin or oxytocin [3,28,31], FGLamide ASTs do not contain Trp or Cys. FGLamide AST in the precursor showed an average peptide length of 11 amino acids with a variable N-terminal sequence and a conserved C-terminal motif. The C-terminal 5 amino acids are essential to inhibit JH biosynthesis in cockroaches, termites and crickets [18]. A dibasic cleavage site follows a Gly residue in every AST sequence.

In most FGLamide AST precursors analyzed, at least one peptide had the amino acid sequence $(X)_{1\rightarrow 16}$Y(S > A > N > G > X)FGLGKR; however, the sequences available are heavily biased toward insect and crustacean species. It seems likely that greater variability may occur as other invertebrate FGLamide ASTs are discovered. For example, in all FGLamide ASTs in the precursor of the copepod *C. finmarchicus,* Leu at position 6 is replaced by Ile or Val (GenBank: EU000307).

In most FGLamide ASTs analyzed, the potential dibasic endoproteolytic cleavage site was Lys-Arg with less frequent use of Arg-Arg. These differences suggest that different endopeptidases participate in prepro-FGLamide AST processing in a regulatory fashion within species or may reflect species-specific enzyme utilization. For example, a metalloendopeptidase purified from rat testis had higher cleavage activity against the Lys-Arg doublet of preproneurotensin than the Arg-Arg doublet of dynorphin A or atrial natriuretic factor [7]. Variation in potential dibasic endoproteolytic cleavage sites was observed in FGLamide AST precursors of *L. cuprina*, *C. vomitoria* and 12 *Drosophila* species. In these species, Arg-Arg as well as Lys-Arg participate in the formation of the active peptides [4,11].

The evolutionary conservation of specific codons for amino acids in the C-terminal core and endoproteolytic cleavage site suggests that selective pressures may have been greater in these regions than at the N-terminus. The importance of the conservation of synonymous codons may have been to create and conserve mRNA secondary structures that involve Watson–Crick pairing. Several models for folding have been predicted in our analysis that suggest that one to two loops or hairpin structures could occur in the mRNA region specifying the C-terminal core or the endoproteolytic cleavage site. The validity of these models remains to be tested experimentally. However, our theoretical approach supports the notion that Watson–Crick pairing between the codons for Phe and the codons for Lys in any FGLamide AST and the codons for Ser and Ala, the most frequently occurring amino acids in the C-terminal motif, pairing with the codons for Arg. In both pairings, the codon pairing was favored by the presence of the codons for Gly-Lys-Gly.

We found the loop structure proposed in the one FGLamide AST mRNA with Ser codons in the precursor from *P. clarkii*, *P. interruptus* and *B. orientalis* whereas with Ala codons, the loop or hairpin loops were found in *P. interruptus*, *M. rosenbergii*, *P. americana*, *D. persimelis* and *G. bimaculatus* although Watson–Crick pairing was with the codons of the N-terminus.

Other examples of the influence of the codon bias in the FGLamide AST mRNA fold were established in the FGLamide ASTs, VERYAFGLGRR from *D. melanogaster*, *D. sechellia* and *D. erecta*. The codon GGT for the first Gly favors two unstable loops. *D. simulans* has the same peptide but the only difference with respect to the other *Drosophila* sp. is that the synonymous codon GGC favors a stable loop.

In conclusion, our results show that the FGLamide AST mRNA secondary structure is determined by the codon usage from the mRNA region that specifies the C-terminal active peptide region and endoproteolytic cleavage sites. mRNA folding will occur by Watson–Crick pairing between codons encoding the N-terminal domain and the first codons in the C-terminal domain or with the codons for the potential endoproteolytic cleavage site.

Acknowledgments

Francisco José Martínez Pérez was recipient of postdoctoral fellowship from CONACYT (77390) and he is member of S.N.I. This research was supported in part by Discovery grants from the Natural Sciences and Engineering Research Council of Canada (SST, WGB and BSWC).

References

[1] Babendure JR, Babendure JL, Ding JH, Tsien RY. Control of mammalian translation by mRNA structure near caps. RNA 2006;12:851–61.

[2] Bellés X, Graham LA, Bendena WG, Ding QI, Edwards JP, Weaver RJ, et al. The molecular evolution of the allatostatin precursor in cockroaches. Peptides 1999;20:11–22.

[3] Bertolini A, Tacchi R, Vergoni AV. Brain effects of melanocortins. Pharmacol Res 2009;59:13–47.

[4] Bowser PRF, Tobe SS. Comparative genomic analysis of allatostatins encoding (Ast) genes in *Drosophila* species and prediction of regulatory elements by phylogenetic footprinting. Peptides 2007;28:83–93.

[5] Buratti E, Baralle FE. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol 2004;24:10505–14.

[6] Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, et al. RNA secondary structure and compensatory evolution. Genes Genet Syst 1999;74:271–86.

[7] Chesneau V, Pierotti AR, Barré N, Créminon C, Tougard C, Cohen P. Isolation and characterization of a dibasic selective metalloendopeptidase from rat testes that cleaves at the amino terminus of arginine residues. J Biol Chem 1994;269:2056–61.

[8] Christie AE, Cashman CR, Brennan HR, Ma M, Sousa GL, Li L, et al. Identification of putative crustacean neuropeptides using in silico analyses of publicly accessible expressed sequence tags. Gen Comp Endocrinol 2008;156(246):64.

[9] Duret L. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 2002;12:640–9.

[10] Duve H, Johnsen AH, Maestro JL, Scott AG, East PD, Thorpe A. Identification of the dipteran Leu-callatostatin peptide family: the pattern of precursor processing revealed by isolation studies in *Calliphora vomitoria*. Regul Pept 1996;67:11–9.

[11] East P, Tregenza K, Duve H, Thorpe A. Identification of the dipteran Leu-callatostatin peptide family: characterisation of the prohormone gene from *Calliphora vomitoria* and *Lucilia cuprina*. Regul Pept 1996;67:1–9.

[12] Elliott KL, Hehman GL, Stay B. Isolation of the gene for the precursor of Phe-Gly-Leu-amide allatostatins in the termite *Reticulitermes flavipes*. Peptides 2009;30:855–60.

[13] Gard AL, Lenz PH, Shaw JR, Christie AE. Identification of putative peptide paracrines/hormones in the water flea *Daphnia pulex* (Crustacea; Branchiopoda; Cladocera) using transcriptomics and immunohistochemistry. Gen Comp Endocrinol 2009;160:271–87.

[14] Hershberg R, Petrov DA. Selection on codon bias. Ann Rev Genet 2008;42:287–99.

[15] Hershberg R, Petrov DA. General rules for optimal codon choice. PLoS Genet 2009;5:e1000556.

[16] Hiller M, Zhang Z, Backofen R, Stamm S. Pre-mRNA secondary structures influence exon recognition. PLoS Genet 2007;3:e204.

[17] Hult EF, Weadick CJ, Chang BS, Tobe SS. Reconstruction of ancestral FGLamide-type insect allatostatins: a novel approach to the study of allatostatin function and evolution. J Insect Physiol 2008;54:959–68.

[18] Kai ZP, Ling Y, Liu WJ, Zhao F, Yang XL. The study of solution conformation of allatostatins by 2-D NMR and molecular modeling. Biochem Biophys Acta 2006;1764:70–5.

[19] Konecny J, Schöniger M, Hofacker I, Weitze MD, Hofacker GL. Concurrent neutral evolution of mRNA secondary structures and encoded proteins. J Mol Evol 2000;50:238–42.

[20] Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 2005;361:13–37.

[21] Krishnan NM, Seligmann H, Rao BJ. Relationship between mRNA secondary structure and sequence variability in Chloroplast genes: possible life history implications. BMC Genomics 2008;28(9):48.

[22] Kwok R, Zhang JR, Tobe SS. Regulation of methyl farnesoate production by mandibular organs in the crayfish, *Procambarus clarkii*: a possible role for allatostatins. J Insect Physiol 2005;51:367–78.

[23] Lorenz MW, Kellner R, Hoffmann KH. Identification of two allatostatins from the cricket, *Gryllus bimaculatus* de geer (Ensifera, Gryllidae): additional members of a family of neuropeptides inhibiting juvenile hormone biosynthesis. Regul Pept 1995;57:227–36.

[24] Martínez-Pérez F, Bendena WG, Chang BS, Tobe SS. FGLamide allatostatin genes in Arthropoda: introns early or late? Peptides 2009;30:1241–8.

[25] Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 1999;288:911–40.

[26] Meyer IM, Miklós I. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucleic Acid Res 2005;33:6338–48.

[27] Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucleic Acid Res 2000;28:292.

[28] Rehfeld JF, Friis-Hansen L, Goetze JP, Hansen TV. The biology of cholecystokinin and gastrin peptides. Curr Top Med Chem 2007;7:1154–65.

[29] Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD. Can codon usage bias explain intron phase distributions and exon symmetry? J Mol Evol 2005;60:99–104.

[30] Shabalina SA, Ogurtsov AY, Spiridonov NA. A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acid Res 2006;34:2428–37.

[31] Slusarz MJ, Slusarz R, Ciarkowski J. Molecular dynamics simulation of human neurohypophyseal hormone receptors complexed with oxytocin-modeling of an activated state. J Peptide Sci 2006;12:171–89.

[32] Smart D, Johnston CF, Curry WJ, Williamson R, Maule AG, Skuce PJ, et al. Peptides related to the Diploptera punctata allatostatins in nonarthropod invertebrates: an immunocytochemical survey. J Comp Neurol 1994;347:426–32.

[33] Secher T, Lenz C, Cazzamali G, Sørensen G, Williamson M, Hansen GN, et al. Molecular cloning of a functional allatostatin gut/brain receptor and an allatostatin preprohormone from the silkworm Bombyx mori. J Biol Chem 2001;276:47052–60.

[34] Stay B. A review of the role of neurosecretion in the control of juvenile hormone synthesis: a tribute to Berta Scharrer. Insect Biochem Mol Biol 2000;30:653–62.

[35] Stay B, Tobe SS. The role of allatostatins in juvenile hormone synthesis in insects and crustaceans. Ann Rev Entomol 2007;52:277–99.

[36] Tobe SS. Bendena WG The regulation of juvenile hormone production in arthropods. Functional and evolutionary perspectives. Ann N Y Acad Sci 1999;897:300–10.

[37] Tobe SS, Bendena WG. In: Kastin AJ, editor. Handbook of Biologically Active Peptides 2006. Burlington: Elsevier Academic Press; 2006. p. 201–6.

[38] Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, et al. FlyBase Consortium. FlyBase: enhancing *Drosophila* Gene Ontology annotations. Nucleic Acid Res 2009;37:D555–9.

[39] Weaver RJ, Audsley N. Neuropeptide regulators of juvenile hormone synthesis: structures, functions, distribution, and unanswered questions. Ann N York Acad Sci 2009;1163:316–29.

[40] Woodhead AP, Stay B, Seidel SL, Khan MA, Tobe SS. Primary structure of four allatostatins: neuropeptide inhibitors of juvenile hormone synthesis. Proc Natl Acad Sci USA 1989;86:5997–6001.

[41] Yagi KJ, Kwok R, Chan KK, Setter RR, Myles TG, Tobe SS, et al. Phe-Gly-Leu-amide allatostatin in the termite *Reticulitermes flavipes*: content in brain and corpus allatum and effect on juvenile hormone synthesis. J Insect Physiol 2005;51:357–65.

[42] Yin GL, Chen Q, Yang WJ. Naturally occurring antisense RNA of allatostatin gene in the prawn, *Macrobrachium rosenbergii*. Comp Biochem Physiol B: Biochem Mol Biol 2007;146:20–5.

[43] Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acid Res 2003;31:3406–15.