Identification of *neuropeptide-like protein* gene families in *Caenorhabditis elegans* and other species

Arif N. Nathoo, Rachael A. Moeller*, Beth A. Westlund†, and Anne C. Hart‡

Massachusetts General Hospital Cancer Center and Harvard Medical School Department of Pathology, 149-7202 13th Street, Charlestown, MA 02129

Edited by H. Robert Horvitz, Massachusetts Institute of Technology, Cambridge, MA, and approved September 21, 2001 (received for review May 10, 2001)

Neuropeptides play critical roles in synaptic signaling in all nervous systems. Unlike classical neurotransmitters, peptidergic neurotransmitters are encoded as preproproteins that are posttranslationally processed to yield bioactive neuropeptides. To identify novel peptidergic neurotransmitters, the *Caenorhabditis elegans* genome was searched for predicted proteins with the structural hallmarks of neuropeptide preproproteins. Thirty-two *C. elegans neuropeptide-like protein* (*nlp*) genes were identified. The *nlp* genes define at least 11 families of putative neuropeptides with unique motifs; similar expressed sequence tags were identified in other invertebrate species for all 11 families. Six of these families are defined by putative bioactive motifs (FAFA, GGxYamide, MRxamide, LQFamide, LxDxamide, and GGARAF); the remaining five families are related to allatostatin, myomodulin, buccalin/drosulfakinin, orcokinin, and APGWamide neuropeptides (MGL/Famide, FRPamide, MSFamide, GFxGF, and YGGWamide families, respectively). Most *C. elegans nlp* gene expression is in neurons. The *C. elegans nlp* genes and similar genes encoding putative neuropeptides in other species are likely to play diverse roles in nervous system function.

Chemical signaling via neurotransmitters is critical for synaptic transmission of information between neurons. Neuropeptides are the most varied and numerous type of neurotransmitters. Invertebrate neuropeptides are thought primarily to modulate synaptic function of classical small-molecule neurotransmitters by means of seven transmembrane domain receptors. However, the recent identification of a FMRFamide-gated sodium channel from *Helix lucorum* suggests that they may also act as fast transmitters (1). In mammals, neuropeptides and their receptors are implicated in behaviors including feeding and sleep (2–5). Despite their clear roles in synaptic signaling and behavior, neuropeptide functions are still not understood.

Biochemical isolation of neuropeptides has been relatively successful in several invertebrate systems, including *Lymnaea stagnalis*, *Drosophila melanogaster*, and *Aplysia californica* (6–8), and has led to the identification of several invertebrate neuropeptide families. In the nematode *Caenorhabditis elegans,* 23 FMRFamide-related proteins (FaRP) neuropeptide genes, designated *flp-1* to *flp-23* (*FMRFamide-like proteins*), have been identified (9). Only *flp-1* has been characterized at the functional level. Animals lacking *flp-1* have abnormal behavior, including uncoordinated movement and hyperactivity (10). The only other *C. elegans* non-*flp* neuropeptide genes that have been identified are the 37 insulin-like genes (11, 12).

Dense-core synaptic vesicles are prevalent in presynaptic terminals of *C. elegans* neurons that are neither FaRP immunoreactive nor catecholaminergic (13), suggesting that non-FaRP neuropeptides are present. Additionally, about 130 genes encoding putative neuropeptide receptors were identified in the *C. elegans* genome (14). This large number of receptors is much higher than the number of *flp*-encoded FaRPs (9) and is reminiscent of the large number of putative "orphan" neuropeptide receptors in vertebrates. This finding suggests that several more families of neuropeptides are as yet undiscovered.

The relative paucity of non-*flp* neuropeptide genes previously identified in *C. elegans* (by either genetic or biochemical techniques) led to the suggestion that FaRPs could be responsible for the majority of neuropeptide signaling in this animal (10). Clearly, neuropeptide signaling is implicated in multiple *C. elegans* behaviors, including defecation and social behavior (15, 16). Described herein are 32 non-*flp* putative neuropeptide genes that are expressed primarily in *C. elegans* neurons. These genes encode 134 unique and 151 total putative neuropeptides. Related candidate neuropeptides are found in other species, suggesting that these genes are functionally conserved.

Materials and Methods

Structural searches used the PATTERNFIND program (http://www.isrec.isb-sib.ch/software/PATFND_form.html). BLAST searches (17, 18) yielded *neuropeptide-like protein* (*nlp*) homologs in other species. Signal peptide cleavage sites were defined by using SIGNALP 2.0 (http://www.cbs.dtu.dk/services/SignalP/) (19). Predicted NLPs and peptides were aligned by using CLUSTALW 1.8 (http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html) (20). *nlp* gene families were defined by using (primarily) motifs, phylogenetic tree construction (MEGALIGN, Dnastar, Madison, WI), and overall predicted peptide homology. A motif is defined as at least three of five identical amino acids at the N or C terminus of at least two predicted peptides within a *nlp* gene and/or family. Motifs also were used for additional searching with PATTERNFIND.

A mixed cDNA library (M. Vidal, Dana-Farber Cancer Institute, Boston) was screened for *nlp-1* through *nlp-21* by PCR amplification. Primers overlapped *nlp* gene predicted start and termination sites. Amplified products were directly cloned (pCR2.1-TOPO, Invitrogen). Twelve *nlp*::*gfp* reporter constructs were made by insertion of the *nlp* promoter into a promoterless green fluorescent protein (GFP)-expressing vector, pPD95.67. The sequences and subcloning sites are: *nlp-1*: *Nsi*I/*Xba*I into *Pst*I/*Xba*I; *nlp-3*: *Nhe*I/*Nsi*I into *Xba*I/*Pst*I; *nlp-16*: *Bam*HI/*Nsi*I into *Bam*HI/*Pst*I; *nlp-2*, *nlp-25* through *nlp-32*: *Bam*HI/*Xho*I into *Sal*I/*Bam*HI. For the remaining *nlp* genes, the putative regulatory region was amplified by PCR (reaction 1). The *gfp* coding regions from pPD95.67 (21) were amplified by using a *nlp*-gene specific primer that also contained GFP vector 5′ sequence and by using a 3′ GFP vector primer (GFP-C, AAGGGCCCGTACGGCCGACTAGTAGG) in an indepen-

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: FaRP, FMRFamide-related protein; *nlp*, *neuropeptide-like protein*; *flp*, *FMRFamide-like protein*; GFP, green fluorescent protein; EST, expressed sequence tag.

*Present address: Department of Environmental Studies, Brown University, Providence, RI 02912.

†Present address: Cambria Biosciences, 2 Preston Court, Bedford, MA 01703.

‡To whom reprint requests should be addressed. E-mail: hart@helix.mgh.harvard.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

dent PCR (reaction 2) (22). The two amplified products contained a 20- to 30-bp region of overlap, enabling amplification (reaction 3) of the full-length *nlp*::*gfp* fragment (template DNA from reactions 1 and 2) by using a primer from the *nlp* promoter and from the GFP vector (GFP-2C, GGAAACAGTTATGTTTGGTATATTGGG). The cellular expression pattern for *nlp-1* through *nlp-3* fragments was indistinguishable from the cellular expression pattern for *nlp-1* through *nlp-3* subcloned constructs. PCR primers are listed in Table 1, which is published as supporting information on the PNAS web site, www.pnas.org; in general the upstream regulatory regions were ≈2.5 kb or extended to the next predicted gene.

Transgenic lines were generated by coinjection of the *nlp*::*gfp* construct (50–70 ng/μl) and *lin-15* rescue construct (pJM24,100 ng/μl) into *lin-15(n765ts)* animals (23, 24). At least two independent transgenic lines were analyzed for each *nlp* gene; 5–10 animals were scored per line. 1,1′-Dioctadecyl-3,3,3′,3′-tetramethylindodicarbocyanine perchlorate (Molecular Probes) was used to stain amphid and phasmid neurons (25) to facilitate identification.

Results

The *C. elegans* genomic sequence (26) was scanned for homologs of previously characterized invertebrate neuropeptides to identify *C. elegans nlp* genes. *nlp* genes are defined as those encoding putative neuropeptides that do not end in RFG. Two different search paradigms were used: similarity-based searching and pattern-based searching.

Similarity-Based Searching Yields *nlp-1* and *nlp-2*. Roughly 600 neuropeptides from the GenBank protein database were used to search the *C. elegans* proteome by using the FASTA program (27) at GENESTREAM (http://www2.igh.cnrs.fr/bin/fasta-guess.cgi). This search, based on similarity of predicted *C. elegans* proteins to previously characterized neuropeptides, yielded only two non-FaRP genes. *nlp-1* and *nlp-2* encode putative bioactive peptides with modest similarity to *Aplysia californica* buccalin and myomodulin, respectively (28, 29).

Pattern-Based Searching Yields *nlp-3* Through *nlp-32*. Bioactive peptides encoded by a single invertebrate neuropeptide gene are often highly related. For example, the *C. elegans flp-1* gene encodes seven putative neuropeptides that are highly related to each other (30). Because of the frequent sequence similarity among bioactive peptides encoded within a given invertebrate neuropeptide gene and the presence of characteristic endoproteolytic cleavage sites (usually KR), pattern-based searching strategies were used to identify additional neuropeptide genes. The PATTERNFIND program at the Swiss Institute for Experimental Cancer Research was used to search for putative *C. elegans* neuropeptide proteins based on structural criteria. Search patterns were permutations of [KR]-[KR]-x(3,20)-[KR]-[KR]-x(3,20)-[KR]-[KR]. Permutations included increasing the number of amino acids between predicted cleavage sites, modulating the number of peptide repeats, and testing single amino acid endoproteolytic sites. Additional related genes/expressed sequence tags (ESTs) were subsequently identified at the National Center for Biotechnology Information by using BLAST (17, 18). Not only were *nlp-1* and *nlp-2* reidentified, but an additional 30 non-FaRP genes were found (*nlp-3* through *nlp-32,* see Fig. 1). Only *nlp-19* was not previously predicted as a gene by the *C. elegans* Genome Sequencing Consortium (26).

For final selection as a putative *C. elegans* neuropeptide gene the following criteria were required. (*i*) The predicted preproprotein fits the PATTERNFIND search criterion (above) and has at least 3/5 identical amino acids at the N or C terminus of the predicted peptides in addition to the flanking dibasic (or monobasic) cleavage site. Intragenic peptide homology of the *nlp* genes with multiple predicted neuropeptides is as high as 100% for many *nlp* genes. (*ii*) The putative preproprotein must be less than 400 aa in length. More than 98% of previously characterized neuropeptide preproproteins and half of *C. elegans* predicted proteins meet this criterion. (*iii*) The predicted preproprotein does not have a predicted biological activity by genetic analysis nor by homology; the objective was identification of new neuropeptide genes. (*iv*) The predicted protein must have a putative signal peptide. Twenty nine percent of the roughly 19,000 predicted *C. elegans* proteins have putative signal peptides. The *nlp* and *flp* genes were the only predicted *C. elegans* genes identified in our searching that met all of these criteria. The following exceptions should be noted: *nlp-22* contains only one putative peptide, and the putative signal peptide (19) for *nlp-4* is unclear.

Some *nlp* genes are grouped in clusters. *nlp-22* is located adjacent to *nlp-2* and *nlp-23*, which encode similar putative neuropeptides. Five *nlp* genes, *nlp-27* through *nlp-31,* which encode homologous putative peptides, are located on cosmid B0213. Two *nlp* genes, *nlp-13* and *nlp-9*, which encode putative peptides unrelated by sequence similarity, are located on cosmid E03D2. *nlp-25* and *nlp-26* are located on cosmid Y43F8C and encode putative peptides with slight similarity. The functional or evolutionary significance of this clustering is unclear.

Posttranslational modification of neuropeptides frequently is required for biological activity. All *flp* genes encode neuropeptides ending with a C-terminal RFG (31), which is presumably converted to RFamide *in vivo* by posttranslational chemical modification (32). The C-terminal glycine of various *nlp* predicted peptides is also likely removed, leaving a C-terminal amide in the putative active peptide, but we have not directly assessed this nor other posttranslational modifications.

***nlp* Gene Families Are Defined by Conserved Motifs.** Based on (*i*) the conserved motifs in predicted *nlp* neuropeptides and (*ii*) similar ESTs/neuropeptides in other species, families of related neuropeptide genes were constructed. Experimental evidence suggests that the level of homology over an entire neuropeptide is less important than the presence of a specific, conserved motif within the bioactive neuropeptide (33–36). Therefore, *nlp* genes were initially grouped into families based on 3- to 5-aa motifs found in predicted *nlp* neuropeptides. A motif is defined as at least three of five identical amino acids at the N or C terminus of at least two predicted peptides of a *nlp* gene.

Neuropeptides or ESTs from other species that encode similar putative peptides were identified by using BLAST. The similarity rarely extends beyond the predicted peptide and cleavage sites. Although ESTs are usually incompletely sequenced, these all have the structural hallmarks of neuropeptide preproprotein genes: endopeptidase sites and (usually) potential signal sequences. ESTs and previously characterized neuropeptides from other species were examined for the presence of *nlp* family motifs. The resulting conserved motifs and family assignments are briefly described below and enumerated in Fig. 1. Functional analysis will be required to determine the biological significance of family assignments and motifs.

nlp gene families are presented below in order of decreasing confidence that they encode neuropeptides. For this purpose, confidence is based solely on homology to previously characterized neuropeptides. *nlp* gene families most likely to encode neuropeptides based on homology are listed first. Next, *nlp* genes that share C-terminal motifs with previously characterized neuropeptides are listed. Then, *nlp* genes whose putative neuropeptides share motifs, putative cleavage sites, and (usually) signal peptides with ESTs in other species are listed. Finally, we list *nlp* genes that lack homologs in other species but fit the structural criterion for a putative neuropeptide gene.

Family, Gene cDNA	Location Cosmid	Predicted peptides	GFP Expression	Organism Related ESTs or neuropeptides	Family, Gene cDNA	Location Cosmid	Predicted peptide(s)	GFP Expression	Organism/ Related ESTs or neuropeptides
FAFA	**II; +3.1**	SPYRAFAFA	ASI, 4 head neurons, 2 tail neurons,	*Ancylostoma. caninum* (AW181631)	**MGL/Famide**	**X; +24.4**	MAAPKQMVFGFG	ASI, IL1, 2 head neurons, 1	*Blabrus cranifer* (AAC72893)
nlp-18	F33A8..2	ARYGFA	spermatheca, NSM, 2 ant. pharyngeal	*Globodera rostochiensis* (AW506028)	*nlp-6*	T23E7.4	YKPRSFAMGFG	tail neuron, intestine	*A. californica* (S64298),
pHA#157		SPYRTFAFA	neurons, rectal gland, intestine	*Heterodera glycines* (BF014583)			AAMRSFNMGFG		*A. carinum* (BF245278, AW626815)
		ASPYGFAFA					LIMGLG		
nlp-20	**IV; +3.5**	FAFAFA	4 head neurons, 4 tail neurons,	*Strongyloides stercoralis* (BE029530)	*nlp-5*	**II; +0.8**	SVSQLNQYAGFDTLGGMGLG	ASI, 2 head neurons,	
yk150e4	F45E4.8	SGPQAHEGAGMRFAFA	spermatheca, intestine,	*Pristionchus pacificus* (AI989039)	pHA#119	F35C11.1	ALSTFDSLGGMGLG	spermatheca, 1 male tail	
		APKEFARFARFARASFA	1 ant. pharyngeal neuron				ALQIFSSLDTLGGMGFG	neuron	
GFGFX	**X; +11.3**	AINPFLDSMG	ADF, ASE, ASH, AWB, ASJ, BAG, HSN, I1,	*Procarnbarus clarkii* (AB029169)	**LOFamide**	**I; -13.4**	x2 DYRPLQFG	1 tail neuron	*Meloidogyne javanica* (BE578742)
nlp-3	F48C11.3	AVNFFLDSIG	I2, I3, I4, M1, M3, NSMR, 3 head neurons,	*Onchocerca volvulus* (AA294559)	*nlp-12*	M01D7.5	DGYRPLQFG		
pHA#116		YFDSLAGQSLG	VNC, occ. I6, M2., pm1VL, intestine	*G. rostochiensis* (AW506277,					
nlp-14	**X; +0.6**	x2 ALDGILDGSGFGFD	ASI, ASK, ASE, PHA, VNC, 2 RVG neurons,	AW506483, AW505751)	**LxDxamide**	**II; +1.6**	HISPSYDVEIDAGNMRNLLDIG	IL1, 2 head neurons, VNC,	
yk438b12	D1009.4	x5 ALNSLDGAGFGFE	intestine	*Globodara pallida* (AW505686)	*nlp-11*	ZK1320.10	SAPMASDYGNQFQMYNRLIDAG	PVD, 3 tail neurons,	*A. carium* (AW181454)
pHA#149		x3 ALDGILDGAGFGFD		*A. carium* (AW735128)			SPAISPATYQPENAFGLSEALERAG	precomma embryos	
		ALNSLDGQGFGFE		*H. glycines* (BF014196)					
		x3 ALNSLDGNGFGFD		*Meloidogyne incognita* (BE225539,			STEHIIRV	7 head neurons, 1 lateral	
nlp-15	**I; +15.8**	AFDSLAGSGFDNGFN	ASH, CAN, HSN, BDU, 5 head neurons,	AW829404)	**No known**	**IV; +3.1**	SEGIPIIE	neuron, intestine	
yk314a1	CC4.2	x2 AFDSLAGSGFGAFN	VNC, 3 RVG neurons, 1 tail neuron,	*Brugia malayia* (AW409399, N43331)	*nlp-16*	T13A10.5	ATHSPEGHIVAKDDNRIGHIE		
pHA#152		AFDSLAGSGFSGFD	intestine	*Ostertagia ostertagi* (AF052589)	yk164b1		SSDSIIIGDIQ		
		AFDSLAQQGFTGFE		*S. stercoralis* (BE581045)			SVDEIIIGHIQ		
		AFDTVSTSGFDSDFKL		*Caricius maenas* (1093782)			NAEDIIIEIQ		
				Onconectes limosus (P37086)			SEIVDEIIDAGFIEIQ		
							STGEVSGIPEIIHLQ		
nlp-8	**I; +2.3**	AFDRFDNSGVFSFGA	ASK, ADL, HOB, 6 head neurons, 2 tail		*nlp-4*	**I; +4.8**	SLILFYLLVAFAAARPVSEEVDRV	none	
pHA#145	D2005.2	AFDRMDNSDFFGA	neurons, I2, g1D, pm5L, pm5R, 2 RVG,			F59C6.6	DYDPRTEAPRLAADDPVSDEDRV		
		SFDRMGGTEFGLM	processes in pharynx, intestine				DYTDRTDAPRYPVDPEAEGEDRV		
GGxYamide	**III; -4.4**	AIPFNGGMYG	ASK, ADL, CAN, 2 lateral neurons, 1 tail and	*A. carnium* (AW700795)	*nlp-19*	**X; +2.5**	IAGLRLPNFL	4 head neurons, VNC in	
nlp-10	F37A8.4	STMPFSGGMYG	1 male tail neuron, 2 ant. pharyngeal	*M. incognita* (AW570660)		K09C8	IGLRLPNML	males, NSM, 4 post.	
pHA#12		AAIPFSGGMYG	neurons	*H. glycines* (BF013619)			MGMRLPNIIFL	pharyngeal neurons,	
		GAMPFSGGMYG						spermatheca	
MSFamide	**X; -9.4**	x3 MDANAFRMSFG	ASI, AWC, PHB, BDU, 4 head neurons,	*Drosophila melanogaster* (P09040)	*nlp-26*	**V; +22.8**	QFGFGGQQSFGG	hypoderm	
nlp-1	C01C4.1	MDPNAFRMSFG	intestine	*A. caninum* (BF249618, AW736481)	yk264f11	Y43F8C.2	GGQFGGMQ		
		VNLDPNSFRMSFG		*M. incognita* (AW827912, AW871456)			GGFNGN		
				H. glycines (BF013500)			GGFGQQSQFGG		
				Haemonchus contortus (BF166642)			GGNQFNG		
							GGSQFNG		
							GGNQFGG		
							GGFGFG		
nlp-13	**V; -7.6**	NDFSRDIMSFG	3 head neurons, NSM, M2, I4,		**YGGWamide**	**V; +3.2**	QWGGGPYGGYGP	ASI, spermatheca, 1	*A. californica* (U85585)
yk169h4	E03D2.1	SGNTADLYDRRIMAFG	spermatheca, LUA, 1 tail cell, dorsal and		*nlp-24*	F35B12.7	GYGGGGYGGG	pharyngeal neuron	*D. melanogaster* (AI293735)
pHA#153		QPSYDRDIMSFG	ventral hypoderm, intestine		yk367e9		YGGYGG		*Amblyomma americanum* (BF008415)
		SAPSDFSRDIMSFG					FTGPYGGYGG		*M. incognita* (AW871114)
		SSSMYDRDIMSFG					GPYGYGG		*B. malayi* (AU079056)
		SPVDYDRPIMAFG					GPYGGGGGLYGALLG		*P. pacificus* (AW097724, AI986659,
		AEDYERQIMAFG							AW052628, AW097269)
nlp-7	**X; -0.2**	QADFDDPRMFTSSFG	ADL, AFD, ASE, ASI, PHA, VNC, 4 head		*nlp-25*	**V; +22.8**	QWGGGGYGNPYGGYG	none	
pHA#121	F16E9.2	SMDDLDDPRLMTMSFG	neurons, 2 RVG neurons			Y43F8C.1	GGGYGGGYGGGFGAQQAYNYQNAA		
FRPamide	**X; -14.8**	SIALGRSGFRPG	1 head neuron, secretory cells near vulva,	*Aplysia californica* (AA827697)	*nlp-27*	**V; -8.2**	QWGYGGMPYGGYGGMGGYGYGMGY	ASI, 3 head neurons,	
nlp-2	T24D8.5	SMAMGRLGLRPG	intestine	*Lymnaea stagnalis* (CAA65635)	yk120f3	B0213.2	MWGSPYGGYGGYGGYGGWG	spermatheca, hypoderm,	
		x3 SMAYGRQGFRPG		*A. caninum* (AW589130)				intestine	
nlp-22	**X; -14.8**	SIAIGRAGFRPG	none	*Toxocara canis* (AA874711)	*nlp-28*	**V; -8.2**	QWGYGGYG	none	
	T24D8.3			*B. malayi* (AA585635)		B0213.3	GYGGYGGYG		
							x2 GMYGGYG		
							GMYGGWG		
nlp-23	**X; -14.8**	LYISRQGFRPA	tail, dorsal and ventral hypoderm		*nlp-29*	**V; -8.2**	QWGYGGYG	hypoderm, intestine	
	T24D8.4	SMAIGRAGMRPG			yk410f3	B0213.4	GYGGYGGYG		
		AFAAGWNRG					x3 GMYGGYG		
							GMYGGWG		
GGAPAF	**III; +7.0**	GGARAMLIT	AFD, 5 head neurons, VNC, 1 ant.	*Caenorhabditis briggsae* (R04509)	*nlp-30*	**V; -8.2**	QWGYGGYG	hypoderm	
nlp-21	Y47D3B.2	GGARAFSADYGDDY	pharyngeal neuron, 1 tail neuron, embryo,	*G. rostochiensis* (AW506326)	yk151e4	B0213.5	GYGGYGGYG		
yk28553		GGARAFLTEM	intestine	*M. incognita* (AW571095)			GYGGYG		
pHA#133		GGARVFQGFEDE		*A. caninum* (AW181311)			GMWG		
		GGGRAFGDMM		*S. stercoralis* (BE030353)			PYGGYGWG		
		GGARAFVENS							
		(GGGRSFPVKPGRLDD)							
nlp-9	**V; -7.6**	GGARAFGFGYNAGNS	ASI, AWB, 4 head neurons, 1 tail neuron,		*nlp-31*	**V; -8.2**	QWGYGGYG	hypoderm, embryos	
yk252f11	E03D2.2	GGGRAFNHNANLFRFD	VNC, spermatheca, vulval muscles, intestine		yk172g8	B0213.6	x2 GYGGYGGYG		
		GGGRAFAGSWSPYLE					GYGGYG		
							GMYGGYG		
							PYGGYGWG		
MRXamide	**IV; +9.0**	GSLSNMMRIG	none	*H. contortus* (AW670629, BF186813)	*nlp-32*	**III; -27.1**	YGGWGG	none	
nlp-17	Y45F10A.5	QQEYVQFPNEGVVPCESCNLGTLMRIG				F30H5.2	GGWG		
yk445b3							GGG		
pHA#141							GYGG		
							GGGWGG		
							GGGWG		
							GGGG		
							GFYGGG		
							GWG		

Fig. 1. *nlp* genes encode putative neuropeptides. *nlp* genes are grouped into families (shaded); the family motif is listed in the first column along with the *nlp* gene name, ESTs, or cDNA designations. Only one yk EST (*C. elegans* EST project) and/or pHA cDNA clone (this paper) is listed even if multiple clones were identified. The second column indicates the chromosome, location (in map units), and cosmid designation. Predicted peptides are listed for each *nlp* gene in the third column. Peptides are predicted based on similarity within a *nlp* gene and putative mono- or dibasic endopeptidase cleavage. Putative peptides are listed sequentially for each *nlp* gene but identical putative peptides encoded by a *nlp* gene are listed only once; the number of times the putative peptide is repeated is indicated (2×, 3×). Predicted posttranslational modifications are not shown for individual predicted peptides. The last predicted *nlp-21* peptide (parentheses)

GFxGF Family. *nlp-3*, *nlp-8*, *nlp-14*, and *nlp-15* contain putative peptides with various similarities. *nlp-14* and *nlp-15* putative peptides contain a C-terminal GFxGF (or GFGF) motif. They are up to 69% identical to orcokinin, a myotropic neuropeptide from *Orconectes limosus* (crayfish) abdominal nerves (37), which also contains a GFGF motif. Removal of the C-terminal FGFN or the last phenylalanine of orcokinin causes a loss of biological activity, suggesting that this motif is functionally significant (50). This motif is also found in orcokinins from *Carcinus maenas* (shore crab) and *Procambarus clarkii* (red swamp crayfish) (38). Many candidate peptides containing the GFxGF motif are encoded by ESTs in multiple species.

nlp-3 and *nlp-8* putative peptides do not contain GFxGF motifs but are very similar to specific putative peptides encoded by *nlp-14* and *nlp-15*. The final putative peptide encoded by *nlp-3* (YFDSLAGQSLG) is 70% identical to part of a *nlp-15* putative peptide (AFDSLAGQGFTGFE). Also, *nlp-3*-related putative peptides encoded by a *Globodera rostochiensis* (potato cyst nematode) EST do not contain a GFxGF moiety.

nlp-8 and *nlp-15* both encode putative peptides starting with AFD that diverge at their C termini, and *nlp-8* putative peptides are up to 46% identical to *nlp-15* putative peptides. ESTs from *Meloidogyne incognita* and *Meloidogyne javanica* (root-knot nematodes) also encode putative peptides with this N-terminal motif and various other similarities. The relative functional significance of these various motifs found in ESTs *nlp-14*, *nlp-15*, *nlp-3*, and *nlp-8* remains unclear.

FRPamide Family. *nlp-2*, *nlp-22*, and *nlp-23* encode putative peptides ending in FRPG. Again, the C-terminal glycine is a likely target for amidation *in vivo*. *nlp-2* putative peptides are up to 60% identical to *A. californica* and up to 42% identical to *Lymnaea stagnalis* (great pond snail) myomodulin peptides (8, 39). Myomodulin peptides from these species share N-terminal homology with *C. elegans* FRPamide family members, but lack the FRPamide of the *C. elegans* putative peptides. Putative neuropeptides containing an FRPG are found within ESTs from other organisms including *Toxocara canis* (Tc-huf-316, AA874711), which is similar to *nlp-2* (40).

MSFamide Family. *nlp-1*, *nlp-7*, and *nlp-13* putative peptides contain MSFG and related C-terminal motifs (including MAFG). The C-terminal glycine is a likely target for amidation; *Drosophila melanogaster* drosulfakinin (fruit fly) also ends in SFG and is amidated *in vivo* (41). *nlp-1* putative peptides are up to 46% identical to *A. californica* (sea hare) buccalin, a neuropeptide neurotransmitter that modulates acetylcholine-induced muscle contraction (42). Multiple ESTs with SFG and/or AFG motifs are found in other species.

MGL/Famide Family. *nlp-5* and *nlp-6* contain putative peptides ending in MGLG and MGFG. The C-terminal glycine is a likely target for amidation. Allatostatin (43), a *Blaberus craniifer* (cockroach) neuropeptide, is up to 43% identical to *nlp-5*. Other allatostatins, helicostatin, *Aplysia* buccalin, and a putative peptide encoded by an *Ancylostoma caninum* EST (dog hookworm) also contain putative peptides with a C-terminal GLG motif; a putative peptide encoded by a *B. malayi* EST is up to 50% identical to the C-terminal end of a *nlp-5* putative peptide.

YGGWamide Family. *nlp-24*, *nlp-25*, and *nlp-27* through *nlp-32* encode putative peptides sharing YGGWG and YGGYG motifs. Interestingly, the APGWamide neuropeptide (U85585) from *A. californica* has a related C-terminal motif (44). ESTs encoding similar putative peptides are found in insects and nematodes. Four ESTs from the nematode *Pristionchus pacificus* encode similar putative peptides and appear to arise from different genes.

GGARAF Family. *nlp-9* and *nlp-21* encode putative peptides containing an N-terminal GGARAF motif. The conservation of this N-terminal motif is particularly striking in comparison to the lack of C-terminal conservation between *nlp-9* and *nlp-2* putative peptides—even within the same gene. Putative neuropeptides beginning with this motif are encoded by ESTs from other nematodes.

FAFA Family. *nlp-18* and *nlp-20* encode putative peptides ending in FAFA and related motifs. Candidate preproproteins that have the structural hallmarks of neuropeptides and containing multiple FAFA putative peptides are predicted in ESTs from other nematode species. Perusal of the genomic sequence suggests that *nlp-20* may be in an operon with a neutral endopeptidase, F45E4.7, which could be involved in *nlp-20* processing.

GGxYamide, LxDxamide, LQFamide, and MRxamide Families. Eight *nlp* genes (*nlp-4*, *nlp-10*, *nlp-11*, *nlp-12*, *nlp-16*, *nlp-17*, *nlp-19*, and *nlp-26*) are not found in multigene families within *C. elegans* and are not clearly related to well-characterized neuropeptides. Based on motifs found in ESTs from other species, four of these genes can be assigned to gene families. *nlp-10* is the *C. elegans* representative of the GGxYamide family, *nlp-11* represents the LxDxamide family, *nlp-12* belongs to the LQFamide family, and *nlp-17* represents the MRxGamide family. Interestingly, a peptide encoded by *nlp-12* was shown to modulate locomotion (45) and shown to have potent myoexcitatory activity in *A. suum* (N. Marks, personal communication). The remaining *nlp* genes that lack related ESTs in other species (*nlp-4*, *nlp-16*, *nlp-19*, and *nlp-26*) may also define additional neuropeptide gene families.

Identification of *nlp* Gene cDNAs. The *C. elegans* and the National Center for Biotechnology Information databases contain ESTs corresponding to 15 *nlp* genes: *nlp-9*, *nlp-12* through *nlp-17*, *nlp-20*, *nlp-21*, *nlp-24*, *nlp-26*, *nlp-27*, and *nlp-29* through *nlp-31*. A mixed stage library was screened for cDNAs corresponding to *nlp-1* through *nlp-21*. cDNA clones were identified for six additional *nlp* genes: *nlp-3*, *nlp-5*, *nlp-7*, *nlp-8*, *nlp-10*, and *nlp-18*. In total, 21 of 32 *nlp* genes have either a *C. elegans* cDNA or EST. Expression was detected in microarray experiments for seven of the 11 remaining *nlp* genes (J. Gaudet and S. E. Mango, personal communication). Of these 11 *nlp* genes, 10 have significant similarity to ESTs in other species, suggesting that they are also expressed genes.

Three *flp* genes are alternatively spliced (10). So far, only one *nlp* gene, *nlp-21*, is known to be alternatively spliced. The putative neuropeptides encoded by alternative transcripts of *nlp-21* differ by one putative peptide in a terminal exon: one *nlp-21* transcript encodes nine putative peptides whereas the second transcript encodes eight. However, primers for PCR screening of the cDNA library were based on software predic-

is differentially spliced. Cells expressing GFP from *nlp* gene reporter constructs are in the fourth column. Neurons/cells scored unequivocally by position and counterstaining: ASH, ASI, ASJ, ADL, ASK, AWB, PHA, PHB, HSN, BDU, spermatheca, vulval muscles, rectal gland cell, pharyngeal neurons, and hypoderm. Other classes of neurons were counted (head, tail, and lateral) but not always identified. Approximately 90% of cells expressing GFP in this figure are neurons. VNC, ventral nerve cord neurons; RVG, retrovesicular ganglion neurons. The last column lists the species and accession number for *nlp* gene family members in other species.

tions of intron/exon structure, possibly excluding identification of alternative transcripts for other *nlp* genes.

Many *nlp* Genes Are Expressed in Neurons and Secretory Cells. To address the cellular expression pattern of *nlp* genes, reporter constructs were generated by using putative 5′ regulatory sequences to drive expression of the GFP gene in transgenic animals (46). Thirty two *nlp* constructs were analyzed; many transgenic lines had complex neuronal expression patterns. (see Fig. 2, which is published as supporting information on the PNAS web site, www.pnas.org). Transgenic animals had no detectable GFP expression for six genes (*nlp-4*, *nlp-17*, *nlp-22*, *nlp-25*, *nlp-28*, and *nlp-32*) although three of these (*nlp-2*, *nlp-22*, and *nlp-28*) are detectable in microarray experiments (J. Gaudet and S. E. Mango, personal communication). Five *nlp* gene reporter constructs (*nlp-7*, *nlp-9*, *nlp-14*, *nlp-15*, and *nlp-21*) are expressed in neurons located in the ventral nerve cord, which is directly involved in locomotion. Eight reporter constructs (*nlp-3*, *nlp-8*, *nlp-13*, *nlp-18*, *nlp-19*, *nlp-20*, *nlp-21*, and *nlp-24*) are expressed in pharyngeal neurons that modulate pharyngeal pumping of food. Expression of GFP in the intestine was also seen for many *nlp* genes. A putative role for *C. elegans* intestinal neuropeptides in defection was previously proposed (16) and the vertebrate gut contains neuropeptides. Embryonic expression (precomma stage) was noted for two *nlp* reporter constructs, suggesting that *nlp-21* and *nlp-31* may function in development, as previously demonstrated for PACAP in vertebrates (47). Transgenes are rarely expressed in the *C. elegans* germ line; it is, therefore, unclear whether any *nlp* genes are expressed in the germ line. An additional neuroendocrine role is suggested for nine genes whose reporter constructs showed expression in somatic gonad tissues, including spermatheca, somatic gonad, uterine muscles, and secretory cells. The predominantly hypodermal expression of YGGWamide family members was striking in comparison to other *nlp* gene families. Members of this family may play a significant role in non-neuronal signaling or, despite similarity to the APGWamide neuropeptide from *Aplysia*, these genes may not encode neuropeptides.

Discussion

Using a pattern-based searching strategy, 32 previously uncharacterized putative neuropeptide genes were identified in *C. elegans*. An additional 51 ESTs encoding putative neuropeptides were identified in various invertebrate species. Our results suggest that the diversity of invertebrate neuropeptides is greater than previously assumed.

We predict that 92 *C. elegans* genes may encode neuropeptides, including 23 *flp* genes (9), 37 insulin-related genes (12), and 32 *nlp* genes. Many of the *nlp* genes are likely to encode neuropeptide preproproteins because they fit the following criteria. First, *nlp* genes are structurally related to previously characterized neuropeptide genes; most have a predicted signal peptide region, are of an appropriate length (<400 aa), and have acceptable sites of endoproteolytic cleavage, which flank the conserved motifs within peptides. Second, most *nlp* genes encode related putative neuropeptides, a common feature of invertebrate neuropeptide genes. Third, many *nlp* putative peptides show sequence similarity and/or share putative bioactive motifs with bona fide neuropeptides. Eighteen of 32 *nlp* genes are in families containing previously characterized neuropeptides. Consistent with their putative role as neuropeptides, *nlp* genes are expressed predominantly in neurons and endocrine tissues.

Pattern-based searching was significantly more productive than similarity searching, indicating the necessity of developing tools that fully use genomic information. However, the pattern-based searching described herein only detects putative neuropeptide genes encoding multiple, related bioactive putative peptides. Individual genes encoding multiple, unrelated bioactive peptides or genes encoding just one novel bioactive peptide were not identified except by similarity. For example, a *Brugia malayi* EST (filarial parasitic nematode, AI834180), which encodes multiple putative neuropeptides ending in FLHFG, was identified. But, the *C. elegans* insulin-related genes (11, 12), which do not contain repeated, similar peptides, were not reidentified.

Most *nlp* gene homologs are found in nematodes. *C. elegans* predicted proteins were searched for neuropeptide genes based on specific patterns and motifs, then related proteins/ESTs in other species were identified based on overall protein similarity. The EST database was not searched by using patterns for novel neuropeptides. Therefore, the relative lack of *nlp* gene homologs in non-nematode invertebrates results from either the scarcity of predicted proteins in the database from other species, evolutionary divergence, and/or a paucity of related, repeated-motif neuropeptides in other species. The latter is likely for *D. melanogaster*. Pattern-based searching of *D. melanogaster* predicted proteins was relatively unsuccessful; only previously identified/annotated FMRFamide genes were identified in the fruit fly.

No clear homolog of a *nlp* putative peptide was identified in vertebrates. Previously characterized vertebrate neuropeptide genes do not encode highly related peptides, which further complicates identification. Interestingly, the last *nlp-8* predicted peptide contains a C-terminal region (4/5 C-terminal amino acids) of sequence identity with human substance P. More sensitive searching techniques might identify new vertebrate neuropeptides with similarity to *nlp* putative neuropeptides. Bioactive RFamide family neuropeptides were previously identified in vertebrates by using immunological techniques (48, 49); a similar approach based on *nlp* gene family motifs might be fruitful.

Neuropeptides are integral to behavior and nervous systems in animals. Using data from the *C. elegans* genome sequencing project, 32 previously uncharacterized putative neuropeptide genes with homologs in other species were identified. Further characterization of the *nlp* genes is likely to provide a greater understanding of mechanisms involved in neuropeptide function in development and behavior.

We thank C. Li, P. Sengupta, C. Bargmann, O. Hobert, J. White, and the van den Heuvel laboratory for advice and N. Marks, J. Gaudet, and S. E. Mango for unpublished data. Pharyngeal neurons were identified by L. Avery (Univ. of Texas Southwestern Medical School); male-specific tail neurons were identified by M. Barr (Univ. of Wisconsin, Madison), and IL1 neurons were identified by L. Ryder (Worchester Polytechnic Institute). The cDNA library was provided by M. Vidal; the *C. elegans* Genetics Center provided *C. elegans* strains, and A. Fire (Carnegie Institution of Washington) provided pPD95.67. We gratefully acknowledge the support of Axys Pharmaceuticals (B.A.W.), the Searle Scholar Foundation, and the National Institutes of Health (A.C.H.).

1. Cottrell, G. A. (1997) *J. Exp. Biol.* **200,** 2377–2386.
2. Lin, L., Faraco, J., Li, R., Kadotani, H., Rogers, W., Lin, X., Qiu, X., de Jong, P. J., Nishino, S. & Mignot, E. (1999) *Cell* **98,** 365–376.
3. Chemelli, R. M., Willie, J. T., Sinton, C. M., Elmquist, J. K., Scammell, T., Lee, C., Richardson, J. A., Williams, S. C., Xiong, Y., Kisanuki, Y., *et al*. (1999) *Cell* **98,** 437–451.
4. Raffa, R. B., Heyman, J. & Porreca, F. (1986) *Neurosci. Lett.* **65,** 94–98.
5. Horvath, T. L., Diano, S. & van den Pol, A. N. (1999) *J. Neurosci.* **19,** 1072–1087.
6. Nambu, J. R., Murphy-Erdosh, C., Andrews, P. C., Feistner, G. J. & Scheller, R. H. (1988) *Neuron* **1,** 55–61.
7. Ebberink, R. H., Price, D. A., van Loenhout, H., Doble, K. E., Riehm, J. P., Geraerts, W. P. & Greenberg, M. J. (1987) *Peptides* **8,** 515–522.
8. Cropper, E. C., Tenenbaum, R., Kolks, M. A., Kupfermann, I. & Weiss, K. R. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 5483–5486.
9. Li, C., Kim, K. & Nelson, L. S. (1999) *Brain Res.* **848,** 26–34.
10. Nelson, L. S., Rosoff, M. L. & Li, C. (1998) *Science* **281,** 1686–1690.

11. Duret, L., Guex, N., Peitsch, M. C. & Bairoch, A. (1998) *Genome Res.* **8,** 348–353.
12. Pierce, S. B., Costa, M., Wisotzkey, R., Devadhar, S., Homburger, S. A., Buchman, A. R., Ferguson, K. C., Heller, J., Platt, D. M., Pasquinelli, A. A., *et al*. (2001) *Genes Dev.* **15,** 672–686.
13. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. (1986) *Philos. Trans. R. Soc. London* **314,** 1–340.
14. Bargmann, C. I. (1998) *Science* **282,** 2028–2033.
15. de Bono, M. & Bargmann, C. I. (1998) *Cell* **94,** 679–689.
16. Dal Santo, P., Logan, M. A., Chisholm, A. D. & Jorgensen, E. M. (1999) *Cell* **98,** 757–767.
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
18. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
19. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997) *Int. J. Neural Syst.* **8,** 581–599.
20. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998) *Trends Biochem. Sci.* **23,** 403–405.
21. Fire, A., Harrison, S. W. & Dixon, D. (1990) *Gene* **93,** 189–198.
22. Hobert, O., Moerman, D. G., Clark, K. A., Beckerle, M. C. & Ruvkun, G. (1999) *J. Cell Biol.* **144,** 45–57.
23. Huang, L., Tzou, P. & Sternberg, P. (1994) *Mol. Biol. Cell.* **5,** 395–412.
24. Mello, C. C., Kramer, J. M., Stinchcomb, D. & Ambros, V. (1991) *EMBO J.* **10,** 3959–3970.
25. Hart, A., Simms, S. & Kaplan, J. M. (1995) *Nature (London)* **378,** 82–85.
26. *C. elegans* Genome Sequencing Consortium (1998) *Science* **282,** 2012–2018.
27. Pearson, W. R. (2000) *Methods Mol. Biol.* **132,** 185–219.
28. Church, P. J. & Lloyd, P. E. (1991) *J. Neurosci.* **11,** 618–625.
29. Brezina, V., Bank, B., Cropper, E. C., Rosen, S., Vilim, F. S., Kupfermann, I. & Weiss, K. R. (1995) *J. Neurophysiol.* **74,** 54–72.
30. Rosoff, M. L., Doble, K. E., Price, D. A. & Li, C. (1993) *Peptides* **14,** 331–338.
31. Nelson, L. S., Kim, K., Memmott, J. E. & Li, C. (1998) *Mol. Brain Res.* **58,** 103–111.
32. Hoyle, C. H. V. (1996) *Neuropeptides: Essential Data* (Wiley, New York).
33. Matuszek, M. A., Comis, A. & Burcher, E. (1999) *Pharmacology* **58,** 227–235.
34. Takahashi, T., Matsushima, O., Morishita, F., Fujimoto, M., Ikeda, T., Minakata, H. & Nomoto, K. (1994) *Zool. Sci.* **11,** 33–38.
35. Geraghty, R. F., Irvine, G. B., Williams, C. H. & Cottrell, G. A. (1994) *Peptides* **15,** 73–81.
36. Schafer, H., Zheng, J., Morys-Wortmann, C., Folsch, U. R. & Schmidt, W. E. (1999) *Regul. Pept.* **79,** 83–92.
37. Stangier, J., Hilbich, C., Burdzik, S. & Keller, R. (1992) *Peptides* **13,** 859–864.
38. Yasuda-Kamatani, Y. & Yasuda, A. (2000) *Gen. Comp. Endocrinol.* **118,** 161–172.
39. Li, K. W., van Golen, F. A., van Minnen, J., van Veelen, P. A., van der Greef, J. & Geraerts, W. P. (1994) *Brain Res. Mol. Brain Res.* **25,** 355–358.
40. Tetteh, K. K., Loukas, A., Tripp, C. & Maizels, R. M. (1999) *Infect. Immun.* **67,** 4771–4779.
41. Nichols, R., Schneuwly, S. A. & Dixon, J. E. (1988) *J. Biol. Chem.* **263,** 12167–12170.
42. Cropper, E. C., Miller, M. W., Tenenbaum, R., Kolks, M. A., Kupfermann, I. & Weiss, K. R. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6177–6181.
43. Bendena, W. G., Donly, B. C. & Tobe, S. S. (1999) *Ann. N. Y. Acad. Sci.* **897,** 311–329.
44. Fan, X., Croll, R. P., Wu, B., Fang, L., Shen, Q., Painter, S. D. & Nagle, G. T. (1997) *J. Comp. Neurol.* **387,** 53–62.
45. Reinitz, C. A., Herfel, H. G., Messinger, L. A. & Stretton, A. O. (2000) *Mol. Biochem. Parasitol.* **111,** 185–197.
46. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. & Prasher, D. (1994) *Science* **263,** 802–805.
47. Suh, J., Lu, N., Nicot, A., Tatsuno, I. & DiCicco-Bloom, E. (2001) *Nat. Neurosci.* **4,** 123–124.
48. Dockray, G. J., Vaillant, C., Williams, R. G., Gayton, R. J. & Osborne, N. N. (1981) *Peptides* **2,** 25–30.
49. Perry, S. J., Yi-Kung Huang, E., Cronk, D., Bagust, J., Sharma, R., Walker, R. J., Wilson, S. & Burke, J. F. (1997) *FEBS Lett.* **409,** 426–430.
50. Bungart, D., Kegel, G., Burdzik, S. & Keller, R. (1995) *Peptides* **16,** 199–204.