Searching for novel peptide hormones in the human genome

Olivier Mirabeau

UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline : Biologie Informatique

Ecole Doctorale : Sciences chimiques et biologiques pour la santé

Formation doctorale : Biologie-Santé

Recherche de nouvelles hormones peptidiques codées par le génome humain

par

Olivier Mirabeau

présentée et soutenue publiquement le 30 janvier 2008

JURY

M.	Hubert	Vaudry	Rapporteur
M.	Jean-Philippe	Vert	Rapporteur
Mme	Nadia	Rosenthal	Examinatrice
M.	Jean	Martinez	Président
M.	Olivier	Gascuel	Directeur
M.	Cornelius	Gross	Examinateur

Résumé

Cette thèse porte sur la découverte de gènes humains non caractérisés codant pour des précurseurs à hormones peptidiques. Les hormones peptidiques (PH) ont un rôle important dans la plupart des processus physiologiques du corps humain. Ce sont de petites protéines sécrétées générées après clivage de précurseurs plus larges codés par le génome.

Dans la première partie de la thèse, l'on introduit des algorithmes, basés sur les chaînes de Markov cachées (HMM), qui vont nous permettent de modéliser les séquences protéiques des précurseurs à hormones peptidiques. On montre que l'on peut dégager des caractéristiques particulières au niveau de la séquence chez ce groupe de protéines et l'on s'attarde en particulier sur la modélisation de deux signaux toujours présents chez ces protéines, les peptides signaux et les sites de clivage par les prohormones convertases. On présente ensuite des algorithmes qui prennent en compte le degré de conservation des résidus le long d'alignements de protéines orthologues. On montre que ces nouveaux algorithmes améliorent de manière significative les résultats obtenus à l'aide des algorithmes classiques. Enfin, après lancement de l'algorithme sur des données de protéomes, l'on dégage une liste de candidats dont certains ont pu être étudiés au laboratoire.

La deuxième et la troisième partie de la thèse présentent les conclusions que l'on peut tirer des données de Western blot relatives aux profils de sécrétion et de découpage (processing) de chacun des deux candidats les plus prometteurs, « spexine » et « augurine ». On présente des données d'expression sur la souris (hybridation *in situ,* immunohistochimie,…) que l'on a récemment obtenues sur ces nouvelles hormones peptidiques potentielles ainsi que des données fonctionnelles sur la « spexine ». En conclusion, l'on avance des hypothèses quant aux fonctions de ces deux protéines. Si les fonctions de ces nouveaux peptides nous sont encore inconnues, leur expression chez la souris, tant au niveau de l'ARN messager que de la protéine, révèle des pistes qui devraient soulever un intérêt certain chez les spécialistes du domaine des peptides.

Enfin, dans la quatrième et dernière partie de la thèse, l'on présente pour quatre autres candidats (dont on n'a pu mener une étude approfondie) des données préliminaires d'expression de gène et de sécrétion *in vitro* après transfection de l'ADN codant pour ces protéines dans des cellules issues de lignées cellulaires pancréatiques.

Mots clés : hormones peptidiques, neuropeptides, chaines de Markov cacahees, analyse de sequences proteiques.

Abstract

The goal of my thesis was to discover novel peptide hormones (PH) in the human genome.

Peptide hormones (PH) are an important class of molecules that are involved in a broad range of normal and pathological physiological processes. PH are small secreted proteins that are processed from larger proteins, called precursors. In the first part of the thesis I introduce hidden Markov models (HMM)-based algorithms that allow for the modelling of those PH precursor sequences. This leads me to derive models of some of their well-known characteristic features, including signal peptides and prohormone convertase cleavage sites. Next, I make use of orthology information and modify the HMM algorithms so that the HMM can incorporate a conservation score that is calculated along protein sequence alignments. I show that this idea brings significant improvements in the quality of predictions and I conclude the chapter by drawing a list of potential candidate peptide hormones that was obtained by running the HMM algorithms on genome-wide protein sequence data.

In the second and third part of the thesis I present data on the most interesting candidates that we named “spexin” and “augurin”. I show *in vitro* subcellular localization, secretion and processing of those proteins after transfection of their coding DNA into endocrine cells. I then show mRNA and protein *in vivo* expression data in the mouse for those two proteins, and present functional data on spexin. I conclude both chapters by making some speculations on the functional significance of those two proteins.

In the fourth part of the thesis I present data on the expression, secretion and processing of four extra candidate peptide hormones.

Acknowledgments-Remerciements-Ringraziamenti

Merci avant tout à mon directeur de thèse au LIRMM Olivier Gascuel, pour m'avoir lancé dans la direction des HMM, et pour avoir cru en moi jusqu'au bout.

Un grand merci à mes prestigieux rapporteurs Hubert Vaudry et Jean-Philippe Vert pour avoir accepté mon invitation improvisée. Je remercie également Jean Martinez pour m'avoir fait l'honneur de présider mon jury de thèse.

Warmest thanks to Nadia for bringing me there in the first place, for her amazing personality, for putting her trust in me, for her energy, infectious enthusiasm, her generous research financial support and for cutting me slack at the beginning of my PhD!

My special thanks go to Cornelius Gross, who was essential throughout my PhD, for his kind support and understanding, for giving me confidence at the beginning, for his great ideas and communicative passion for science.

I am indebted to Ewan Birney for accepting to be in my EMBL TAC, for supervising my bioinformatics work, for hosting me at the EBI at the beginning of my PhD and being so supportive all along.

I am grateful to the generous EMBL PhD program which gave me this opportunity, especially Anne Ephrussi for her understanding.

Thank you to all my colleagues at the EMBL who helped me along the way with their advices and good mood.

Special thanks to our precious histologist Emerald Perlas for his refreshing enthusiasm, wonderful *in situs* and histochemistry and for inspiring discussions.

Thank you to the Gross lab for their (love and) affection, Tiago Ferreira, Luisa Lo Iacono, Enrica Audero, Valeria Carola, Olga Ermakova, Theodoris Tsetsenis, Giovanni Frazzetto, Apar Jain and Amaicha Depino.

Thank you also to the members of Nadia's lab for their help and good nature, Nadine Winn, Paschalis Kratsios, Esfir Slonimski, Arianna Nenci, Lieve Temermann,, Catarina Catela, Tommaso Nastasi, Marion Huth, Katik Salimova Foteini Mourkioti, Michele Pelosi, Christian Fasci, Eli Reuveni, Marianne Hede (E-tidsskrifter, really useful!), Ekaterina Salimova, Pascal Te Welscher, Katia Semenova.

From the EMBL and collaborators of the EMBL I owe much to David Tosh (for the fantastic insulin-packed RIN cell line), Enrique Lara-Pezzi (my first PCR… sigh…), Tiago Ferreira (for your warm company and laughs in the late hours at the lab and especially for letting me ride your motorino), Paschalis Kratsios (for this precious aliquot of 20mM dNTPs), Shin Kang (without you I would still be trying to clone that 56 blunt sequence into the TA vector), Walter Witke (if only the FLAG wasn't acidic!), generally the Witke lab for being so generous in letting me have whatever I want from their lab, Pietro Pilo-Boyl (calcetto, cell fractionation, political and philosophical discussions), Enrica Audero (il capo for organising everything and her precious help), Arianna Nenci (for becoming the genespring expert so I can work on my project). Special thanks to Nadine Winn, Luisa Lo Iacono, Tiago Ferreira and Elke Kurz for many good memories.

I’d like to thank also to Mark Carter, Jose Gonzalez, Augusto De Paolis, the caretakers and the admininistration for running things so smoothly at the EMBL.

Grazie a Roberta Possenti per la sua simpatia e i preziosi consigli e segreti su come si deve lavorare con i peptidi! Grazie anche a Cinzia Severini per i suoi dati eccezionali.

Merci aux thésards (et ex-thésards) du LIRMM pour leur accueil chaleureux, tout particulièrement Jean-Baka Domelevo-Entfellner et Raluca Uricaru pour leur précieuse aide logistique ainsi que leur amitié. Dans cette catégorie, je remercie également Nicolas Philippe, ainsi que Samuel Blanquart, Cécile Bonnard et Celine Scornavacca. Un grand merci au monsieur HMM du LIRMM, Laurent Bréhélin, pour sa disponibilité et ses suggestions pertinentes. Merci enfin à Catherine Larose de l’Ecole doctorale Biologie-Santé pour m’avoir si gentiment aidé dans mes démarches administratives.

Merci à mes précédents mentors Christian Saguez, Robin Munro et Georg Casari, pour m’avoir accordé leur confiance.

Merci à mes potes d’Arrayscout qui m’ont initié aux joies de la programmation java, Nicolas Rodriguez, Charles Girardot, Andreas Buness et Sylvie Lefranc. Merci aussi au team bioinfo d’Epigène, Guillaume Kerboul, Eric Perche et David Monteau qui ont accompagné mes premiers pas dans la vie active. Merci à Julien Gagneur, pour son amitié et son coaching.

Cela va sans dire, je remercie ma famille, qui reste essentielle.

Un tendre merci enfin à Friederike Jönsson, pour m’avoir supporté (au sens français du terme) et apporté son soutien technique et psychologique sur la durée. Cette thèse lui doit beaucoup.

Table of contents

List of figures

Abbreviations

°C	Celsius degrees
aa	amino acid(s)
Ab	antibody
ANP	atrial natriuretic peptide
BBB	blood-brain barrier
BMP	bone morphogenetic protein
bp	base pairs
cDNA	complementary deoxyribonucleic acids molecule
CDS	coding sequence of a gene
CP	choroid plexus
CPH	candidate peptide hormone
CSF	cerebrospinal fluid
CVO	circumventricular organ
DEPC water	diethylpyrocarbonate treated water
DMSO	dymethyl sulfoxide
dNTP	deoxyribonucleic triphosphate
ECF	extracellular fluid
ECM	extracellular matrix
ER	endoplasmic reticulum
EST	expressed sequence tag(s)
GER	gastro-esophageal reflux
GERD	gastro-esophageal reflux disease
GPCR	G protein-coupled receptor
h	hour
HMM	hidden Markov model(s)
ICF	intracellular fluid
IGF1	insulin-like growth factor 1
IR	immunoreactive/immunoreactivity
kDa	kilodalton
KO	knock-out
LES	Lower oesophageal sphincter
MALDI-TOF	matrix assisted laser desorption/Ionisation-time of
ME	median eminence

mg	milligram
ml	milliliter
mM	millimolar
mRNA	messenger ribonucleic acid molecule
MW	molecular weight
n.s.	not specified
o.n.	overnight
OD	Optical density
PBS	phosphate buffered saline
PC	prohormone convertase(s)
PCR	polymerase chain reaction
PFA	paraformaldehyde
PH	peptide hormone(s)
PTM	post-translational modifications
PVN	paraventricular nucleus of the hypothalamus
PVT	paraventricular nucleus of the thalamus
RAAS	renin-angiotensin-aldosterone system
RF	Reissner's fibre
ROC	receiver operating curve
rpm	rounds per minutes
RT	room temperature
RT-PCR	reverse transcription polymerase chain reaction
s	second
SCO	subcommisural organ
SFO	subfornical organ
SGZ.	sub-granular zone of the hippocampus
SP	signal peptide
SVZ	subventricular zone of the brain
TEMED	N,N,N′,N′-tetramethylethylenediamine
Tris	2-amino-2-hydroxymethyl-propan-1,3-diol
UTR	untranslated region of mRNA.
UV	ultraviolet
ZG	zona glomerulosa of the adrenal gland
PPT	pedunculopontine nucleus
LDT	laterodorsal tegmental nucleus

List of common amino acids (aa)

common name	3-letter name	1-letter name	biochemical properties
alanine	Ala	A	hydrophobic
arginine	Arg	R	basic
asparagine	Asn	N	polar
aspartate	Asp	D	acidic
cysteine	Cys	C	hydrophobic
glutamate	Glu	E	acidic
glutamine	Gln	Q	polar
glycine	Gly	G	non-polar
histidine	His	H	weakly basic
isoleucine	Ile	I	hydrophobic
leucine	Leu	L	hydrophobic
lysine	Lys	K	basic
methionine	Met	M	hydrophobic
phenylalanine	Phe	F	hydrophobic
proline	Pro	P	non-polar
serine	Ser	S	polar
threonine	Thr	T	polar
tryptophan	Trp	W	non-polar
tyrosine	Tyr	Y	non-polar
valine	Val	V	hydrophobic

1 Introduction

1.1 Context of the project

When I was presented this project in January 2002, the draft genome of the mouse had just been released and I didn't know much about biological sequence analysis, let alone experimental biological work. Back then, an engineer by training, I had applied for the European Molecular Biology Laboratory (EMBL) PhD programme to learn molecular biology and I realise now how lucky I was when Nadia Rosenthal decided to supervise my thesis at the Mouse Biology Unit of the EMBL: the way was long and arduous, and all the more rewarding for it.

I have two researchers to thank for shaping up my project: Cornelius Gross, the neuroscientist/molecular biologist who supervised most of my work on a weekly (sometimes even daily) basis at the EMBL Monterotondo and Ewan Birney, the head of the Ensembl project who introduced me to databases and guided me in my early PhD years through the meanders of biological sequence analysis. The idea that novel peptide hormones (PH) could be found by looking in the newly sequenced vertebrate genomes reportedly came up during a group leading retreat in the winter of 2001. Cornelius and Ewan's rationale was based on the following observations: Firstly, several peptide hormone (PH) precursors had been recently discovered, notably ghrelin (Kojima et al. 1999) and orexin (Peyron et al. 1998) precursors. Secondly, PH had been notoriously difficult to identify using traditional biochemical purification techniques (mass spectrometry, chromatography techniques, etc.) due to their small size and chemical instability. Lastly, the large number of orphan G protein-coupled receptors awaiting ligands strongly suggested that many endogenous peptide hormones were yet to be discovered.

The recently sequenced vertebrate genomes, and in particular the mouse genome (Waterston et al. 2002) provided a unique resource to find novel PH. And by making this data accessible and usable, Ensembl and other publicly funded resources made it possible for me to embark on this adventure.

1.2 Importance of peptide hormones (PH)

PH are proteins encoded by the genome that are involved in the regulation of nearly all homeostatic processes in the body and fundamental behaviours. They modulate processes as diverse as blood pressure regulation, glucose homeostasis, sleep and feeding behaviour, fear, anxiety and stress, addiction and pleasure, sexual behaviour, and learning and memory.
One common role that they often share is that of a relay between the central nervous system and the periphery. They are typically messengers sent by the peripheral organs/tissues to inform the brain about the environment, the current stock of energy, and the organism's general metabolic state. The central nervous system also needs to send signals to the rest of the body so that it can, among other things, adjust best to the requirements imposed by the environment.

Basic organismal survival mechanisms including feeding and sexual behaviour, reproduction, energy metabolism, gastrointestinal function, response to stress, and even some immune functions are governed by hormones. Some of those PH regulate multiple processes, including orexin, leptin and pituitary adenylate cyclase activating peptide. The knowledge of all signalling molecules and their expression is a prerequisite for understanding how processes interact. Vertebrates have evolved biological processes which enable them to adapt to short and long term changes in their environment. Peripheral PH such as gastrointestinal and fat hormones inform the brain on nutritional status, and the brain integrates different signals, and implements an appropriate behaviour.

1.3 Examples of peptidergic systems

PH regulate virtually all physiological functions of the body. It is beyond the scope of this thesis to present in any detail the peptidergic systems that govern those regulatory processes. I shall however mention a few of the most prominent ones in the next paragraphs.

1.3.1 The hypophysiotropic PH

Certainly, the most important endocrine gland of the body is the anterior pituitary gland (adenohypophysis). For a long time it was believed to be the "master" gland because it produced hormones that controlled the metabolism of several other major glands and organs of the body including the gonads (follicle-stimulating hormone and luteinizing hormones), mammary glands (prolactin), adrenal glands (adrenocorticotropic hormone), thyroid gland

(thyroid-stimulating hormone), liver and fat (growth hormone). We now know that all of those essential trophic hormonal systems are controlled upstream by hormones produced by the hypothalamus, the hypophysiotropic peptides, or hypothalamic-releasing factors. Those include corticotropin-releasing factor (CRF), luteinizing hormone-releasing hormone (LHRH), Growth hormone releasing hormone (GRF), thyroid hormone-releasing hormone (TRH), Gonadotropin-releasing hormone (GRF), somatostatin (SOMA), vasoactive intestinal peptide (VIP), Arginine vasopressin protein(AVP). The hypothalamus is in turn regulated by those organs/glands to form negative feedback loops -or axes- such as the hypothalamus-pituitary-adrenal gland axis (HPA axis), hypothalamus-pituitary-thyroid gland axis (HPT axis) and hypothalamus-pituitary-gonad axis (HPG axis). Tight regulation of those systems is crucial since they all fulfil essential functions in the body: the HPA axis is known to modulate response to stress, ("fight or flight" mechanisms), the HPT axis is important for energy metabolism, while the HPG is responsible for the onset of reproduction.

1.3.2 Regulation of bodily fluids volume

Overall water and salt (Na+ is the most abundant in the body) concentration determines osmotic pressure in cells, and movement of fluids across membranes. Regulation of the body fluids volume, one of the most fundamental processes in organismal biology is orchestrated by the renin-angiotensin-aldosterone system (RAAS). This system is also responsible for regulating thirst and hence modulates behaviour. Apelin and vasopressin are two other PH important for regulation of fluids in the body.

1.3.3 Other systems

Calcium homeostasis is mainly regulated by the parathyroid hormone, an essential PH produced in the parathyroid gland. Insulin is involved in sugar homeostasis and hence energy balance. Some of the important PH involved in feeding behaviour are leptin, neuropeptide Y, agouti-related peptide (AGRP), members of the urocortin family, and ghrelin. The control of energy stores and energy use is controlled by many parameters at many different levels. For instance, the fact that our corporeal mass stays relatively constant is an example of the tight regulation of all the mechanisms involved. Another example of peptidergic system is the oxytocin system. Oxytocin is a PH released from the posterior pituitary (neurohypophysis) that affect social bonding, trust (Kosfeld et al. 2005), maternal and sexual behaviour (Pedersen et al. 1982). Finally, regulation of gastrointestinal and gastro-oesophageal motility

and acid secretion are mediated by the following PH: cholecystokinin, gastrin, intermedin, gastrin-releasing peptide, oxyntomedullin.

1.3.4 Pleiotropy of PH

Leptin (Flier et al. 2000), CART (Moffett et al. 2006; Vrang 2006), PACAP (Vaudry et al. 2000), orexins (Peyron et al. 1998) offer fascinating examples of gene pleiotropy. They open windows on how biological processes cooperate to yield coherent responses to environmental stimuli, stresses, or constraints.

1.3.5 Cell biology of PH

D. Steiner and colleagues in 1967 (Steiner et al. 1967) demonstrated that the bioactive insulin molecule was processed from a larger protein termed proinsulin, and M. Chrétien's team showed that bioactive melanotropins and endorphins were proteolytically processed products of the same precursor, propiomelanocortin. This mechanism of bio-activation has since then been verified for many PH, growth factors, and receptors (IGF1, CART), and is present in all eukaryotes. The essential milestone in the characterization of the proteases responsible for activating those proteins, the mammalian prohormone convertases (PC) family, was the discovery in (Julius et al. 1984) of a novel protease, Kex2, that was shown to cleave the yeast alpha-factor mating pheromone, a peptide-hormone-like molecule. Yeast kexin 2 was found to cleave correctly the POMC precursor (Thomas et al. 1988), and from then on it became likely that an ortholog of Kex2 in mammals would be the endoprotease responsible for conferring bioactivity to mammalian PH (Thomas et al. 1988). Indeed shortly after, a protein named furin, with sequence features analogous to Kex2 (50% identity in their catalytic domains) was shown to be the long sought endoprotease (Fuller et al. 1989). Since then, it has been demonstrated that cleavage by members of the prohormone convertase family (PC) of precursor proteins constitutes a general principle of control of activity of signalling molecules such as PH and GF. Furin does not cleave only PH/growth factors, it also cleaves numerous other precursors of bioactive molecules, such as receptors, cell-adhesion molecules, but also bacterial toxins such as diphtheria and anthrax and viral envelope glycoproteins such as AIDS and Ebola viral proteins (Cota et al. 2006; Thomas 2002).

There are six other known enzymes homologous to furin, which were cloned using PCR-based strategies. Together, they form a group called the prohormone convertases (PC) (Steiner 1998). PC1 and PC2 are the two prohormone convertases known to be responsible

for cleaving neuropeptides and PH, but the cleavage sequences specificity for each of those PC is not entirely understood.

1.4 Milestones

The term “hormone” was first used in 1905 when Ernest Starling discovered secretin, a gastrointestinal PH. Starling was the first researcher to describe a hormone a substance which is produced by organs and act at a distance on other organs. Soon after gastrin was isolated, that was shown to have an effect on gastric acid secretion (HCl) in the stomach. Some essential PH discoveries in the last century include the characterization of insulin by Banting and Macleod in 1923, oxytocin by Du Vigneaud in 1955, and the hypothalamic releasing hormones thyroliberin, gonadoliberin and somatostatin by Schally and Guillemin in 1977. In the 80s, chemical techniques based on the principle to find amidated peptide were developed and pancreastatin, a peptide from the chromogranin precursor, was identified this way and characterized as an inhibitor of insulin (Tatemoto et al. 1986). In the 1990s, reverse pharmacology (Civelli 1998) was developed that led to the isolation of several PH. The so-called orphan receptor strategy or “reverse pharmacology” central idea is to use orphan G-proteins-coupled receptors (GPCRs), i.e. GPCRs for which no ligand has been found, as baits to isolate and purify their corresponding cognate ligands (Civelli et al. 1998). This strategy has led to the isolation and characterization of several novel endogenous PH in the 1990s including orphaninFQ/Nociceptin (Reinscheid et al. 1995), Neuropeptide S (Xu et al. 2004) and NPB/NPW (Singh and Davenport 2006) . Lately, researchers have sought to uncover peptides using a combination of bioinformatics and biochemistry techniques (Chartrel et al. 2003; Shichiri et al. 2003; Zhang et al. 2005)

1.5 Precedents in genome-wide search for PH

There have been previous efforts made to mine the genome and try to survey the complete set of neuropeptides in genomes. I give here some examples of those studies.

1.5.1 Neuropeptides in the worm

Most C. Elegans neuropeptide-like genes have a defined structure: neuropeptides precursors in this organism typically contain multiple “copies” of the same short bioactive peptide, flanked by dibasic cleavage sites, a hallmark of furin/prohormone convertases cleavage sites. In (Nathoo et al. 2001), the authors attempt to survey the entire worm proteome in search of

repeated peptides flanked by dibasic sites. In mammals, only a few precursors have this structure, including thyroliberin which releases the peptide (pyro)Glu-His-Pro-NH_2 a multiple times and the proenkephalin and predynorphin precursors that encode several times a variation of the opioid peptide Tyr-Gly-Gly-Phe- NH_2. The authors took advantage of this peculiarity of neuropeptide-like precursors in the worm to find them systematically.

1.5.2 Neuropeptides in the fly

An attempt at surveying the genome to establish the complete repertoires of the fly D. Melanogaster neuropeptides and GPCRs can be found in the study by (Hewes and Taghert 2001) in which the authors report 22 bioactive peptide precursors and 44 GPCRs. However, the authors of this study came short of testing the biochemical properties/functions of those genes.

1.5.3 Neuropeptides in the mosquito

Neuropeptides of the mosquito A. Gambiae are essential for driving its feeding behaviour. Without this drive to have regular blood meals, it cannot reproduce, and the malaria parasites can not spread. Knowledge on how neuropeptides perform their task could prove instrumental in fighting against the scourge of its parasite Malaria. Using well-established bioinformatics methods, (Riehle et al. 2002) have mined the genome of A. Gambiae in the hope to inventoriate all PH and neuropeptides, and uncovered 35 PH genes.

1.5.4 Neuropeptides in the bee

Very recently, (Hummon et al. 2006) used a combination of molecular biology and biochemical techniques (mass-spectrometry, MALDI-TOF) and bioinformatics searches to derive the set of neuropeptides in the honey bee, a model for the study of social behaviour in insects.

Surprisingly, no study on the fish “neuropeptidome” has been published yet, even if many of the neuromodulators governing energy homeostasis in these animals are being discovered (Leder and Silverstein 2006). As most PH are of vertebrate origin the repertoire of neuropeptides/PH is likely to be very similar to the mammalian one.

1.6 Analysis of the problem

This section is an attempt at defining what PH are. A brief survey of PH and related proteins (growth factors, chemokines, etc.) is made and a list of PH is established, that will be used for building and testing the models.

1.6.1 Definition of PH

The largely intuitive notion of PH can be apprehended in a number of ways. From a cell biologist standpoint, a PH is a short protein which undergoes a series of chemical transformations in the cell (cf. Figure 1) and serves as a chemical messenger for other cells. The pharmacologist sees it chiefly as a ligand for receptors, which has the potential to affect the physiology and metabolism of organs and modulate currents in neuronal populations. A biochemist could define it as a biologically mature product of a PH precursor after proteolytic processing by prohormone convertases (cf. Figure progonadoliberin alignment in the bioinformatics part for an example of a peptide part of a larger precursor, progonadoliberin). In order to predict novel PH I needed to define them systematically, collect their sequences and find out the features they have in common. My starting point was the following: a PH is a short stretch of aa which has hormonal properties. As a logical consequence of that simple analysis, I needed to define what peptides and hormones are.

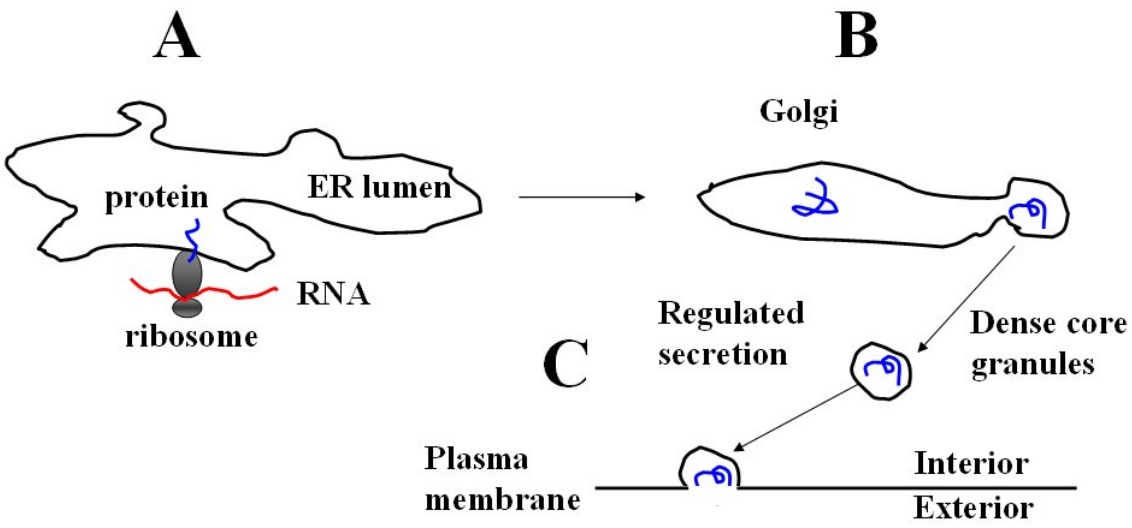

Figure 1: Cellular pathways for the synthesis of PH (adapted from the textbook Molecular Biology of the Cell (Alberts et al. 1989))
(A) Cotranslational transport of proteins across the endoplasmic reticulum membrane. Ribosomes are drawn in grey, the mRNA is drawn in red and the nascent protein is pictured in blue inside the ER lumen. (B) Folding of the protein, posttranslational modifications (n-glycosylation, amidation, etc.) and proteolytic processing occur in the Golgi. (C) PH are packaged into dense core granules and released in a regulated fashion into the extracellular milieu by exocytosis.

1.6.1.1 Peptides

The most general definition for a peptide is that it is a short protein. There is no clear consensus on the number of aa beyond which a polypeptide chain is no longer a peptide, and any threshold on the size of the protein is somewhat arbitrary. The Britannica encyclopaedia reads: “Peptide chains longer than a few dozen aa are called proteins.” This definition is too imprecise to be used to discriminate between peptides and non-peptides. The Columbia University Press and Oxford University Press dictionaries both agree on a threshold of about 50 aa beyond which a polypeptide chain is a protein. However, there are notable exceptions to that rule, like insulin-like growth factors, which is a PH of 70 aa.

Another convention states that a protein is a peptide, when it is short enough to be synthesised. However, this definition became flawed as advances made in peptide synthesis technologies made it possible to construct ever longer polypeptide chains. Wikipedia stresses the difficulty in assigning an objective definition to peptides and settles for the following definition: "a peptide is an aa molecule without secondary structure; on gaining defined structure, it is a protein." However this definition is also problematic since it would result in

the exclusion of all cysteine bonds-containing hormones (insulin, insulin-like growth factors) from the PH class. Along the same line, the divide could also lie in the fact that, for a peptide, only a few aa interact with the receptor. The majority of them are very short peptides, which are perhaps structurally more related to the classical monoamine neurotransmitters (dopamine, serotonin, glutamate, epinephrine and norepinephrine) than from proteins possessing globular domains. Characteristically, neurotransmitters as well as the great majority of PH are cognate ligands to G-protein coupled receptors, as opposed to growth factors which typically bind to tyrosine kinase receptors and other classes of receptors.

1.6.1.2 Hormones

A hormone is a chemical messenger that carries information from a population of cells or organ, to another group of cells or organ. It generally conveys its messages through binding of specific receptors on target cells that trigger signal transduction cascades and lead to physiological and morphological changes (e.g. intracellular concentration of cyclic AMP and calcium, modification of synapses in the central nervous system, etc.).

1.6.1.3 Endocrine vs. paracrine action of PH

Endocrinologists and animal physiologists generally distinguish between three types of hormonal mode of action. Substances which circulate in the bloodstream to act on organs at a distance are said to be endocrine (or exocrine depending on the type of glands that secrete them). Signalling molecules that directly affect neighbouring cells after being released are termed paracrine. In contrast to neuropeptides and neurotransmitters that typically act in a paracrine fashion, hormones are mostly endocrine (circulating). Some substances exhibit both paracrine and endocrine characteristics, including insulin-growth factors, which are produced by the liver and muscle and exert actions on peripheral tissues (muscle, bone, etc.). The third case is when a hormone binds receptors on the same cell population than the one which produced the hormone; they are then said to be autocrine.

1.6.1.4 PH vs. neuropeptides

Neuropeptides are secreted and processed proteins expressed in the brain that essentially modulate neuronal activity through binding to specific receptors, often coupled to G proteins.

Very schematically, binding of neuropeptides to G-protein coupled receptors (GPCRs) induce changes in the intracellular concentration of ions such as calcium, and modify the electrostatic state of the cell. All neuropeptides can be considered PH, in that they are peptides derived from larger precursors and function as chemical messengers of the brain. Conversely, many PH are expressed in the brain and give rise to peptides which exert neuromodulating functions. One classic example is the pro-opiomelanocortin precursor, which produces at the same time β-endorphins that function as neuropeptides and adrenocorticotropic hormones (ACTH), which are PH, produced in the anterior pituitary, that target the adrenal gland. Cholecystokinin, a gastrointestinal PH that was originally characterized as inducing contraction of the gallbladder (Ivy and Oldberg 1928), inhibits gastric acid secretion and emptying, and fulfils important functions as a neuropeptide to modulate, among other things, satiety, anxiety and stress behaviours (Bhatnagar et al. 2000; Crawley and Corwin 1994). Corticotropin-releasing factor (CRF), a hormone produced in the hypothalamus triggers pituitary adrenocorticotropic hormone (ACTH) release into the bloodstream (Vale et al. 1981) and also modulates feeding behaviour in hypothalamic neurons of the paraventricular nuclei (Heinrichs et al. 1992). Another hypophysiotropic hormone, prolactin-releasing peptide is a neuropeptide since it has been shown to modulate the activity of oxytocin neurons (Maruyama et al. 1999b) and is widely distributed in the mammalian brain (Maruyama et al. 1999a). Other notable peptides with ascertained dual roles as both hormone and neuropeptide include pituitary adenylate cyclase-activating peptide (Vaudry et al. 2000), gastrin-releasing peptide (Grimsholm et al. 2005; McDonald et al. 1979), vasoactive intestinal peptide and neuropeptide Y.

All those examples suggest that most PH expressed in the brain have neuromodulatory properties and hence are neuropeptides. Conversely, I could not find any example of a neuropeptide expressed elsewhere in the body which is not a PH. Those two observations lead to consider that, essentially, PH and neuropeptides are members of the same group that I will refer to throughout the thesis as “PH”.

1.6.1.5 Secreted and Signalling molecules

PH are particularly interesting molecules because they are chemical messengers of the body. They share this property with other groups of molecules, which are also secreted molecules and signal to cells through interaction with specific receptors. In the next paragraph I briefly

review the main groups of secreted molecules found in mammals, with a special emphasis on signalling molecules and properties they share with PH.

1.6.1.6 Hormones that are not PH

Some hormones are not cleaved nor give rise to products small enough to be considered peptides. This is the case for about 30 hormones. All classic anterior pituitary tropins belong to this group: Somatotropin (growth hormone), thyrotropins, follitropins, choriomammotropin, choriogonadotropin, but also thrombopoietin, transthyretin, thyroglobulin, erythropoietin and adiponectin. Those hormones are not classified as PH because they are not processed by prohormone convertases at dibasic-like residues and they are two large to be considered peptides (more than 100 aa long, with globular domains).

1.6.1.7 Growth factors

By definition, growth factors are secreted signalling molecules that affect the cell cycle to increase proliferation of cells and inhibit their differentiation programs. Examples of growth factors are vascular endothelial growth factor (VEGF), transforming-growth factor beta (TGF-β), granulocyte colony-stimulating factor (G-CSF), Insulin-like growth factors (IGFs), epidermal growth factor (EGF), bone morphogenetic proteins (BMPs), nerve growth factor (NGF), etc. Growth factor proteins all contain a signal peptide, which allows the molecule to go through the cell secretory pathway. Nearly all growth hormones are larger than PH (typically larger than 100 aa); they have globular domains, and their associated receptors are almost never G-protein coupled receptors (they are for instance tyrosine kinase receptors). Some of them do get processed at pairs of dibasic residues as PH are, but the resulting mature proteins are much larger than PH. Processed growth factors with their associated length include brain-derived neurotrophic factor (BDNF, 119 aa), transforming growth factor-beta 1 (TGF-β, 112 aa), bone morphogenetic protein 1 (BMP1, 866 aa), glial cell-line derived neurotrophic factor (GDNF, 134 aa), beta-nerve growth factor (β-NGF, 120 aa), platelet-derived growth factor A (PDGF-α, 125 aa), neurturin (102 aa), insulin-like growth factor 1A (IGF-1A, 70 aa) which can be considered as both a PH and a growth factor, etc... Note that IGF1A is one of the litigious cases and can also be considered to be a PH (I have included it in the final list of PH). As those representative examples suggest, the size of growth factors are in a range which differs radically from the size of PH. Hence, the model should be able to distinguish between growth factors and PH by using lengths/sizes arguments. It is noteworthy

that by tuning the parameter controlling for peptide length, one could in theory look for novel growth factors precursors instead of PH precursors.

1.6.1.8 Cytokines and chemokines

Cytokines are a class of secreted signalling molecules important for immune functions. Many of them also have, like growth factors, an effect on cell fate (proliferation, growth, differentiation and cell death) and there is a large overlap between cytokines and growth factors (e.g. Stromal cell-derived factor 1 precursor, Macrophage colony-stimulating factor 1, most of interleukins, bone morphogenetic proteins). Chemokines form a sub-group of cytokines. They are small immune molecules that have chemotactic properties. Chemokines and many other cytokines are recognizable by their short length (between 50 and 100 aa) and typical cysteine bonds motifs (Cys-Cys, Cys-X-Cys or Cys-X-X-X-Cys). Emerging data support claims that classical immunological molecules such as chemokines also have a role of neuromodulators and even neurotransmitters in the brain (Rostene et al. 2007), so the functional frontier between the two classes of molecules could be much more tenuous than previously thought. Very few non-growth factors cytokines are processed at pairs of basic residues. By querying the Swiss-Prot database (cf. Resources) I could only retrieve 5 such proteins (oncostatin M, interferon gamma, and tumor necrosis factor ligand superfamily members 12 and 13 and 13B). Hence a good model should be able to distinguish between these two types of molecules.

1.6.1.9 Other secreted proteins

Human signalling molecules, mainly constituted by human PH (80 proteins), growth factors (146 proteins) and cytokines (182), only represent a fraction of the 1598 secreted proteins. Other notable secreted proteins include histocompatibility antigens proteins, extracellular matrix proteins, Metalloproteinase-like proteins, serum albumin protein, and innate immune proteins such as defensins and antimicrobial peptides. Of those proteins, 71 are annotated in Swiss-Prot as having a dibasic processing site, including complement factors ADAMTS proteins (a disintegrin and metalloproteinase with thrombospondin motifs), Dickkopf-related protein 4 and von Willebrand factor. This reflects the fact that the mechanism of protein activation through cleavage by prohormone convertases is not restricted to PH.

1.6.1.10 List of PH

Making a definitive list of all PH was not easy because of the relative overlap between the group of growth factors/cytokines and the group of PH. The uncertainty over the size threshold for defining peptides added to this problem. However, this task was greatly facilitated by Swiss-Prot, the most comprehensive database repository of (mostly manually) annotated protein sequences. The aim was to query the Swiss-Prot database using the sequence retrieval system and produce a list of typical PH to build and test the model. Swiss-Prot provides annotations in the "Keywords" field for neuropeptide and hormone, and peptide annotations in the features key box. After analysis, the strategy I used to retrieve known PH was to form a set of sequences, which had the annotation "neuropeptide" or "hormone" in the "Keywords" field and intersect it with the set of proteins, which had an annotation "PEPTIDE" in "FeatureKey" of the feature table. Note that it was necessary to use the annotation "neuropeptide" to retrieve some PH; since Swiss-Prot considers those two keywords to be exclusive (only two proteins have both a "neuropeptide" and a "hormone" annotation in Swiss-Prot, galanin and melanin-concentrating hormone precursors). In Swiss-Prot, 49 proteins out of the 80 annotated hormones had a "PEPTIDE" annotation in the "FeatureKey" field and hence fell into the category of PH. 2 neuropeptides out of the 23 that were annotated as such were missing a "PEPTIDE" annotation (galanin-like peptide and prokineticin 2). Those two precursors were manually added in the list. This made for a total of 70 PH/neuropeptides. However after some exploratory work on the set of secreted proteins, I realised that several typical PH were missing from that list. I found out that by using the annotation "cleavage on pair of basic residues" in the "Keywords" field I could gather most of the missing sequences, including the KISS-1 PH precursor and the gastrin-releasing peptide precursor. By using the expression EXP:

EXP=((keywords="cleavage on pair of basic residues") OR (keywords="neuropeptide") OR (keywords="hormone")) AND (FeatureKey="PEPTIDE")

I obtained a list of 93 proteins, from which I then had to discard the 13 following sequences that had a dibasic cleavage site but were manifestly not PH: coatomer subunit alpha, beta defensins 118, 125, 126, 127, neuroendocrine secretory protein 55 precursor histatin 3 precursor, serine protease inhibitor Kazal-type 5 precursor, integral membrane protein 2B and 2C precursor, prothrombin precursor, plasminogen precursor and proSAAS precursor. I then manually added some proteins, which I deemed PH. Vitronectin can be considered a PH since

it releases somatomedin B at an arginine residue (dibasic-like site) but it has no annotation "neuropeptide", "hormone" or "cleavage on pair of basic residues" associated to it in Swiss-Prot. Parathyroid hormone precursor was also added to the list, although its peptide size of 84 aa was considered too large for Swiss-Prot to annotate it as a peptide precursor.

1.6.1.11 Difficult cases

One could argue on a number of decisions I made about exclusion/inclusion of problematic proteins in the list. Insulin-like growth factors, and parathyroid hormone were included in the list although their sizes of over 70 aa make them untypical peptides. The angiotensinogen precursor was included as well. It is a classic PH; although it is cleaved by the protease renin which is not a member of the prohormone convertases family. However, this protein was not predicted well by the models due to the atypical cleavage site sequence. Liver-expressed antimicrobial peptides 1 (hepcidin) and 2 are not considered as PH although their primary features suggest they are (they have a signal peptide, a dibasic cleavage site that defines a short peptide) and recent work suggest they could be signalling molecules that are important for iron homeostasis (Nemeth et al. 2003).

1.6.1.12 Outline of the strategy

PH all have some features in common: they possess a signal peptide at their n-terminus and exhibit at least one cleavage site recognized by members of the prohormone convertases family in order to release a peptide. The aim of the whole thesis was to find members from that family.

2 Hidden Markov models for PH precursors prediction

2.1 Introduction to the problem

2.1.1 Common features of PH precursors

PH precursors have some characteristics in common: they possess a signal peptide at their n-terminus and exhibit at least one cleavage site recognized by members of the prohormone convertases family (noted “cleavage site” in short) in order to release a (short) peptide. The aim of this chapter is to derive algorithms which can recognize proteins that exhibit those features.

Figure 2 shows an alignment of orthologous sequences of progonadoliberin, a PH precursor protein. It shows the primary structural features of the precursor, including the signal peptide, the 10 amino acid amidated peptide gonadoliberin, the perfectly conserved cleavage site (denoted CS) and the non-conserved flanking peptide region. Note that throughout this thesis in order to be consistent with SwissProt nomenclature flanking peptides are (inaccurately) referred to as propeptides, although the term “propeptide” generally refers to the precursor without the signal peptide.

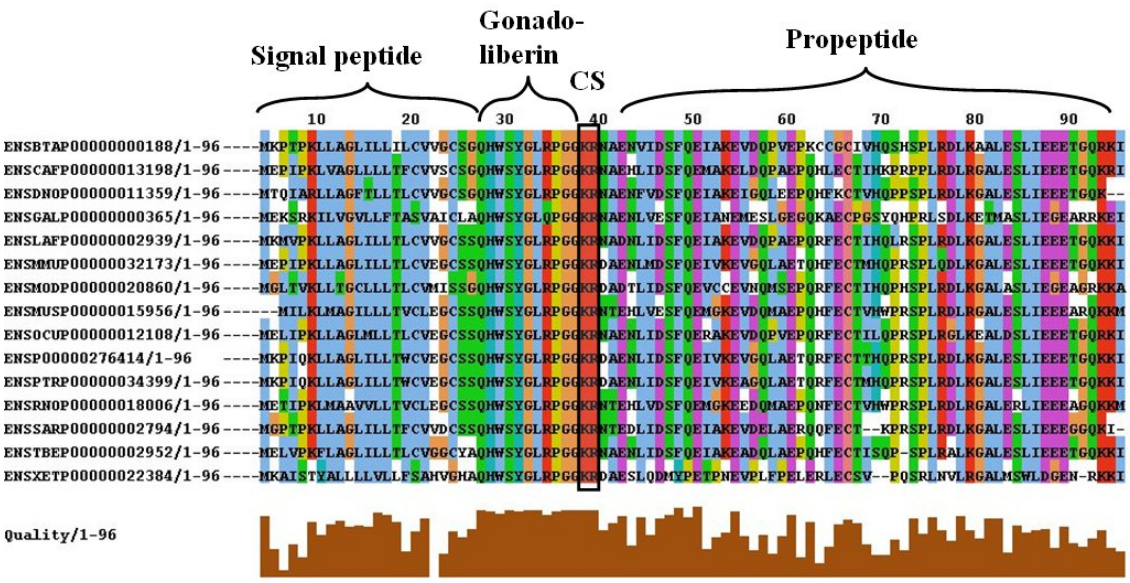

Figure 2: Alignment of progonadoliberin homologous sequences
Human gonadoliberin is a 10 aa-long amidated peptide ((pyro)Glu-His-Trp-Ser-Tyr-Gly-Leu-Arg-Pro-Gly-NH2) which is almost perfectly conserved in mammals. CS denotes a prohormone convertase cleavage site. Note the decrease of conservation in the alignment after the cleavage site CS. The multiple alignment was downloaded from the Ensembl website (www.ensembl.org, cf. Resources) and the Jalview software (Clamp et al. 2004) was used to visualise the alignment. The colouring of one letter code according to the corresponding aa biochemical properties (light blue colour for hydrophobic aa; red colour for basic aa, purple for acidic aa).

2.1.1.1 N-terminal signals

Much of the information about the subcellular localization of a protein is localized at the n-terminus of a protein (Emanuelsson et al. 2000). As a protein starts to be translated in the cytoplasm, a number of RNA/protein complexes binds to the nascent protein and target it into specific organelles. The best characterized N-terminal signal is the secretory signal peptide (noted SP, in short) which targets a precursor protein into the endoplasmic reticulum (ER) and determines its passage through the cellular secretory pathway. Other well-characterized N-terminal localization sequences include mitochondrial transit sequences which target a protein encoded by the nuclear genome to the mitochondria (Kumar et al. 2006). Those sequences can be mistaken for signal peptides and for this reason it is useful to model those signals to be able to discard mitochondrial proteins from the list of putative PH precursors.

2.1.1.2 Absence of transmembrane domains

By default, signal peptide-containing proteins are released in the extracellular milieu through exocytosis of secretory vesicles. However, if in addition to a signal peptide, a protein contains

stretches of 15-25 hydrophobic aa (called transmembrane domains) it finds itself incorporated into the plasma membrane of the cell (Krogh et al. 2001). Receptors and channels are examples of such membrane proteins. This means that for a protein to be secreted, it must not contain any transmembrane domains. There are some notable exception to this rule, as some membrane proteins release soluble ligands at the cell surface after cleavage by proteases, including fractalkine and members of the tumor necrosis factor family (TNF12 for instance).

2.1.1.3 Prohormone convertases (PC) cleavage sites

Mammalian PC proteases family form a group of seven proteases named PC1/3, PC2, furin, PC4, PC5, PACE4 and PC7 (Seidah and Chretien 1999). They are a phylogenetically old family of proteins also described as the subtilisin/kexin-like family after the name of their yeast (kexin) and bacterial counterparts (subtilisin) (Steiner 1998). Furin is the first PC to have been characterized as processing PH (Fuller et al. 1989; Nakayama 1997). Furin processes hormones and growth factors which undergo the constitutive secretory pathway, including IGF-1 (Duguay et al. 1995; Thomas 2002). However, the members of the family most important for the processing of PH (and neuropeptides) such as cocaine and amphetamine-related transcript (CART) are believed to be neuroendocrine-specific PC1/3 and PC2 (Stein et al. 2006; Steiner 1998). Substrate recognition sites for the PC family are generally described as dibasic sites Arg/Lys-Arg/Lys↓, where the arrow represents the site where cleavage occurs (the proteases recognize two basic residues just N-terminal of the cleavage site). Substrate recognition site motifs and specificity for those proteases will be further discussed in subsection 2.4.8 dedicated to building the PC cleavage site model.

2.1.2 Strategies for predicting PH precursors

Several strategies and formalisms were considered for modelling and predicting PH precursors. One strategy consisted in predicting each feature separately, and combining these independent predictions into a global PH predictor (cascade-like strategy). Modelling each feature separately makes it possible to choose different methods and formalisms according to the type of prediction. For instance, signal peptide and PC cleavage sites are localized signals that can be best predicted using non-linear methods such as artificial neuronal networks (Bendtsen et al. 2004; Duckert et al. 2004) or kernel-support vector machine (Vert 2002) that are methods able to capture dependencies between residues. One could envisage making one classifier for signal peptides and another one for short peptides flanked by either two cleavage sites or one cleavage site and a signal peptide cleavage site (or the end of the protein). A

protein would then be classified as a PH precursor if it is predicted as harbouring both features.

"Black box" approaches were also considered such as those consisting in summarizing information of parts of the sequence by translating it into other variables such as frequencies of aa pairs (Park and Kanehisa 2003) or more generally recoding the entire sequence (Xue et al. 2006) to enable use of powerful machine learning algorithms on those more amenable variables. I decided to employ the hidden Markov model (HMM) formalism over multi-layered prediction or "black box" machine learning approaches chiefly because it has an established record in solving a broad range of biological sequence problems (cf. subsection 2.2.3). More concretely, they provide an elegant and rigorous framework that can naturally handle variable sequence lengths signals, including short peptides, and implicitly integrate information from multiple heterogeneous and dispersed signals in the sequence, including aa compositional biases and localized signals such as cleavage sites. Furthermore, HMM framework allowed me to easily incorporate information about conservation of sites in a sequence (cf. 2.6). In the following sections 2.2 and 2.3 I give the necessary background on hidden Markov models.

2.2 Introductory notes on HMM

2.2.1 Foundations of HMM

HMM are mathematical objects, whose foundations were laid in the late 1960s by Baum and colleagues in a series of founding papers (Baum and Eagon 1967; Baum and Petrie 1966; Baum et al. 1970). HMM theoretical foundations lie in the field of inference theory, which itself is part of probability theory. HMM also belong to the artificial intelligence/machine learning field that aims to discover patterns in complex and noisy data in an automatic way. HMM owe much of their popularity to the computational feasibility of their algorithms, Viterbi, Forward-backward and Baum-Welch.

The founding papers mentioned above are not easily understandable by non-mathematicians, and concepts underlying HMM were made accessible to engineers and researchers from diverse fields through the landmark review (Rabiner 1989). In this report, Rabiner stresses the practical aspects of HMM, presents ready-to-use algorithms and puts the emphasis on their range of applications and limitations. This report certainly contributed to the dissemination of this method for solving many real-world problems and going in-depth into this report remains a sure route towards the understanding of HMM. In addition, this report enabled a certain standardization of HMM algorithms and notations in the field.

2.2.2 Early applications

Although Baum and colleagues had already found applications of their theory for modelling ecology systems (Baum and Eagon 1967), HMM only became popular in the late 1980s with the advent of a plethora of applications ranging from speech recognition (Kenny et al. 1990; Rabiner 1989) and digital handwriting recognition (Mohamed and Gader 1996), to linguistics and data compression. Notable applications to linguistics include word disambiguation/semantic tagging (the same word can have two different meaning depending on the context) (Abney 1996; De Loupy et al. 1998) text segmentation (Yamron et al. 1998) and information extraction (Freitag and McCallum. 1999). Algorithms derived from the HMM framework have given rise to commercial products such as speech recognition devices (Dragon Systems dictation system, (Gillick et al. 1998; Yamron et al. 1998), which have greatly facilitated the work of translators.

When Rabiner wrote in the opening paragraph of his report the following sentences, he was only foreseeing the future of HMM applications in molecular biology: “Although initially introduced and studied in the late 1960s and early 1970s statistical methods of Markov source or hidden Markov modelling has become increasingly popular in the last several years. There are strong reasons why this has occurred. First the models are very rich in mathematical structure and hence can form the basis for use in a wide range of applications. Second, the models, when applied properly, work very well in practice for several important applications.”

Since the early 90’s, applications of HMM are ubiquitous in molecular biology, protein science and bioinformatics/computational biology. Below are several examples of some pioneering applications of HMM in biology.

2.2.3 HMM in computational biology

Applications in the field of molecular biology and bioinformatics/computational biology have really started to emerge at the beginning of the 1990s. Early applications of HMM in molecular biology include the prediction of protein secondary structure (Asai et al. 1993), gene finding (Krogh et al. 1994b) and modelling of protein families including G-protein coupled receptors (Baldi and Chauvin 1994), Fibronectin type III domains in yeast (Bateman and Chothia 1996). HMM were also designed to predict three-dimensional structure of proteins (Camproux et al. 2004) and distinguish between prokaryotic and eukaryotic promoters (Pedersen et al. 1996), signal peptides and signal anchors (Nielsen and Krogh 1998), and between transmembrane domains and signal peptides (Kall et al. 2004).

They have been used to predict the subcellular localization of a protein, for instance mitochondrial (Kumar et al. 2006) and membrane-spanning (Krogh et al. 2001) proteins.

HMM are now established as the standard tool used for both whole genome annotations (Birney et al. 2001; Birney and Durbin 2000) and protein sequence analysis (Sonnhammer et al. 1998). They have been developed to model virtually every feature in nucleic acids and protein sequences.

Their relative ease-of-use and computational efficiency make them a fashionable and popular tool of the bioinformatician’s repertoire, on a par with other popular machine learning techniques such as neural networks and support vector machines.

In the end of the 1990s a flurry of HMM-based public domain algorithms and tools were made available to the larger community of life scientists, and this surely has contributed to the growing popularity of HMM.

2.2.4 Profile HMM

Krogh and colleagues were the first ones to implement Profile HMM (Krogh et al. 1994a). Profile HMM were developed as a generic tool to model any relatively conserved domain. It was designed to be flexible in allowing insertions and deletions at conserved sites (cf. profile HMM of Figure 3) and proved successful at modelling a broad range of groups of related sequences.

The original profile HMM introduced in (Krogh et al. 1994a) was seminal and, although some are more sophisticated, virtually all public domain implementations of Hidden Markov model for biology were largely inspired on the original idea of having match states for conserved sites and insertion/deletion states at every position.

Packages SAM (Hughey and Krogh 1996), PFTOOLS (Bucher et al. 1996), and HMMpro (Baldi et al. 1994) are all based on the original profile HMM (Eddy 1998). Libraries of protein domains profile HMM have been built, such as PFAM (Sonnhammer et al. 1998) and are now a standard way to describe protein domains and structural motifs. Figure 3 shows the architecture of different profile HMM implementations. Profile HMM are constrained by their underlying architecture, but this architecture is flexible, hence permitting to model a broad range of sequence objects. HMMER is a widely used package implementing profile HMM. It can do a great variety of tasks, in particular building profile HMM from multiple alignments. Typically HMMER takes as input a multiple alignment of a domain family, or protein family and learns the parameters that fit best the data. Then one can ask HMMER to rate how good the model is, (i.e. how homogeneous the set of sequences) and evaluate how likely a novel sequence belongs to that family.

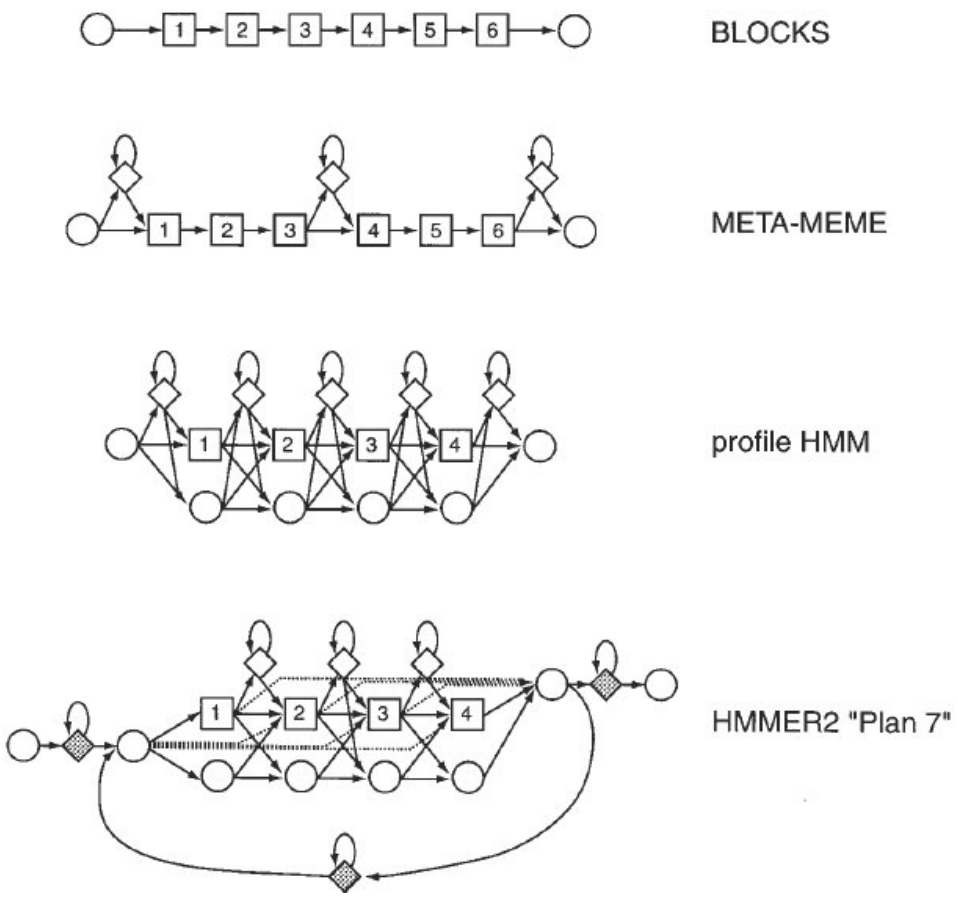

Figure 3: Different architectures of profile HMM, [extracted from (Eddy 1998)]
The BLOCKS architecture is equivalent to a weight-matrix. In this model, no insertions and no deletions are allowed. There is a one-to-one relationship between a site residue (that can be a nucleic acid, or an aa) and its corresponding emitting state (squares numbered 1, 2,…). In the META-MEME model, only insertions are allowed, and only between sites, which are three positions apart from each other. The original profile HMM as described by Krogh in 1994 is more elaborate and allows at each match state (squares numbered 1, 2,…) for both insertions (diamonds on top of match states) and deletions (circles at the bottom of match states) of residues. The more sophisticated HMMER2 "PLAN 7" used the same idea as in the original profile HMM, but has a slightly modified topology and allows via the passage through the shaded diamond state at the bottom to match the same motif several times for a given sequence.

This enables to automatically generate domains and sequence families, such as those in the PFAM databases. The HMMER package provides sophisticated algorithms, which can optimize the size of the profile HMM to explain the data best without overfitting (this is done by using mixture Dirichlet priors in a Bayesian framework).

The goal of this thesis was to find novel PH in the human genome/proteome. In the next chapter, I will present an HMM designed to model and ultimately predict PH precursors. So a legitimate question could come to the reader's mind at this point: Why not use the tools that a decade of research in HMM applications for molecular biology has fostered? Why build *de*

novo an HMM model while tools exist that have *a priori* been optimised to do it? There are two justifications for that: Firstly, profile HMM building packages such as HMMER work best when given a full alignment of the sequences to be modelled and the group of PH precursors is too heterogeneous to build an alignment of it. It is possible to group PH into different families such as insulin-like PH, glucagon-like family (VIP, GIP, intermedin, PACAP, secretin), and opioid peptides precursors (prodynorphin, proenkephalin and proopiomelanocortin) (Brownstein 1993; Noda et al. 1982), but making profile HMM of those groups of sequences only leads to finding PH from known families, and this has already been done. Secondly, as will become clear in the next chapters, those packages would not have given me the freedom that I needed to test different ideas and algorithms, including the incorporation of conservation scores in the HMM, a specific cyclic topology for the HMM, and non-standard scores that I've used to rank candidate PH precursors (cf. 2.7).

Understanding of the following section requires that the reader be familiar with some of the material introduced in Rabiner's tutorial which I present in this background section. Rabiner's report has certainly become the reference document for anyone wishing to learn quickly how to use HMM. The notations I have used throughout this thesis have been largely borrowed from this classic report.

2.3 Mathematical background on discrete HMM

2.3.1 HMM formal definition:

In the following section, I introduce to the reader the basics of discrete first order HMM. For that I borrowed Rabiner's notations and some of its prose: Let $S = \{S_1, S_2, ..., S_N\}$ be the N states describing the system, and let q_t be the actual state at *t*. The letter *t* stands for the *time*, as hidden Markov models were originally applied to continuous processes involving time. Note that for biological sequences analysis, *t* represents a *position* in the sequence and not a *time* point. For discrete first order homogeneous Markov chain, the transition probability to a state S_j at *t* only depends on the state at *t-1*: It is independent of the time point *t* (homogeneity property) and independent of the values of the states for time points prior to *t-1* (first-order property). One can say that a first-order Markov chain has a memory of only one state, all anterior states are "forgotten".

$$P(q_t = S_j / q_{t-1} = S_i, q_{t-2} = S_k, ...q_0 = S_l) = P(q_t = S_j / q_{t-1} = S_i),$$

and this probability is independent of *t*. Hence, we can define a set of state-to-state transition probabilities a_{ij} as:

$$a_{ij} = P(q_t = S_j / q_{t-1} = S_i), \quad t > 0, \quad 1 \le i, j \le N. \tag{0.1}$$

An HMM is characterized by:

1) Its hidden states $S = \{S_1, S_2, ..., S_N\}$.

Typically, in protein sequence analysis problems, the hidden states represent regions of homogeneity in the sequence, such as transmembrane domains, or conserved sites in a conserved domain such as cleavage sites.

2) Its observation symbols $V = \{v_1, v_2, ..., v_M\}$.

In protein sequence analysis, symbols are the 20 most common aa found in nature plus one "dummy" aa Z used for the implementation of the END state.

$V = \{\text{A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V,Z}\}$.

3) Its state transition probability distribution $A = \{(a_{ij})_{1 \le i, j \le N}\}$

$$a_{ij} = P(q_t = S_j / q_{t-1} = S_i), \quad t > 0, \quad 1 \le i, j \le N \tag{0.2}$$

Those correspond to the state transition probabilities introduced in equation (0.1).

4) Its observation symbol probability distribution $B = \{(b_{jk})_{1 \le j \le N, 1 \le k \le M}\}$

$$b_{jk} = P(v_k \text{ at } t / q_t = S_j), \quad 1 \le j \le N, \quad 1 \le k \le M \tag{0.3}$$

b_{jk} is the probability that state S_j emits symbol v_k. For convenience of notation, the probability of generating (or emitting) the symbol $O_t = v_k$ at *t*, which was previously noted b_{jk}, will also be written $b_j(O_t)$.

5) Its initial state distribution $\Pi = \{(\pi_i)_{1 \le i \le N}\}$

$$\pi_i = P(q_1 = S_i), \quad 1 \le i \le N. \tag{0.4}$$

6) Its end state probability distribution $\Omega = \{(\omega_i)_{1 \le i \le N}\}$

Let i_{END} be the index such that $S_{i_{END}}$ corresponds to the *END* state. ω_i is defined as:

$$\omega_i = P(q_{t+1} = S_{i_{END}} / q_t = S_i), \quad 1 \le i \le N. \tag{0.5}$$

ω_j are the probabilities of reaching the END state from state S_j.

Throughout this thesis, the model will be referred to as $\lambda = (A, B, \Pi, \Omega)$. An HMM can be viewed as a Markov chain on states, each of which emits symbols with probabilities $(b_{jk})_{1 \le j \le N, 1 \le k \le M}$. At every time point *t>1*, two things happen: First a state $q_t = S_j$ is determined according to the state $q_{t-1} = S_i$ at *t-1* and the state transition probability a_{ij}. Second a symbol $O_t = v_k$ is generated by the state $q_t = S_j$ with probability $b_{jk} = b_j(O_t)$.

2.3.2 The three problems

There are typically three questions that an HMM can solve:

1) Problem 1:

Given an observation sequence $O = O_1 O_2 ... O_T$ and a model $\lambda = (A, B, \pi)$, how does one compute efficiently $P(O / \lambda)$, the probability of the observation sequence given the model? Solving problem 1 allows ranking the proteins according to the model, i.e. to find the most likely PH in a list. It also provides an extended framework to rate models on how good the data fit to them (Maximum Likelihood approach).

2) Problem 2:

Given the observation sequence $O = O_1 O_2 ... O_T$ and a model $\lambda = (A, B, \pi)$, how does one compute a corresponding sequence of states $Q = Q_1 Q_2 ... Q_T$, which is optimal in some meaningful way, and that best explains the observations? Solving problem 2 will allow the labelling of a protein sequence according to the model. Solving problem 2 is typically referred to as decoding, as a reference to the HMM traditional fields of application (speech

recognition/processing). I will present in the next chapter different decoding algorithms and compare them according to how good the corresponding cleavage site predictions are.

3) Problem 3:

How do we adjust the model parameters $\lambda = (A, B, \pi)$ to maximise $P(O/\lambda)$? Knowing how to solve problem 3 will be useful to build the HMM of PC/furin cleavage sites and signal peptides.

2.3.3 Algorithms for solving the three problems

In this section I present the classical algorithms for solving those three problems: The Forward-backward, Viterbi and posterior-Viterbi decoding, and Baum-Welch algorithms.

2.3.3.1 Solving problem 1: The forward-backward algorithm

Solution to problem 1 is given by the forward algorithm. Let us define the forward variable $\alpha_t(i)$:

$$\alpha_t(i) = P(O_1 O_2 ... O_t, q_t = S_i / \lambda). \qquad (0.6)$$

We can solve $\alpha_t(i)$ inductively by the forward algorithm:

Forward algorithm

---

1) Initialization:

$$\alpha_1(i) = \pi(i) b_i(O_1), \quad 1 \le i \le N . \qquad (0.7)$$

2) Induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1}), \quad 1 \le t \le T-1, \quad 1 \le j \le N . \qquad (0.8)$$

3) Termination:

$$P(O/\lambda) = \sum_{i=1}^{N} \alpha_T(i)\omega_i \ . \qquad (0.9)$$

Using Landau's notation, where 'O' denotes the big-O symbol relative to T, we can see that the forward algorithm is achieved in $O\ (N^2T)$ operations. Note that problem 1 can also be solved by checking all possible state sequences and computing the corresponding probability. But this takes $O\ (TN^T)$ operations, which is computationally unfeasible, even for small values of T.

We define at this point the backward algorithm, which is usually implemented for solving problem 3 in the Baum-Welch procedure. It is not normally used for solving problems 1 or 2, but it is introduced here at this point because it is required for defining variables for other decoding algorithms (cf. posterior-Viterbi decoder paragraph 2.3.3.3).

Analogously to the forward variable we can define the backward variable:

$$\beta_t(i) = P(O_{t+1}O_{t+2}...O_T / q_t = S_i, \lambda). \qquad (0.10)$$

Again we solve for $\beta_t(i)$ inductively by the backward algorithm.

Backward algorithm

1) Initialization:

$$\beta_T(i) = \omega_i, \quad 1 \le i \le N \ . \qquad (0.11)$$

2) Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}.b_j(O_{t+1}).\beta_{t+1}(j), \quad t = T-1, T-2, ..., 1, \quad 1 \le i \le N . \tag{0.12}$$

2.3.3.2 Solving problem 2: The Viterbi algorithm

There is not one single optimal way to solve problem 2. One solution consists in choosing a path $Q = Q_1 Q_2 ... Q_T$ such that each Q_t verifies:

$Q_t = Arg \max_i [\gamma_t(i)]$, where $\gamma_t(i) = P(q_t = S_i / \lambda, O)$. $\gamma_t(i)$ can be written using the forward and backward variables :

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}. \tag{0.13}$$

Although we maximize the expected number of correct states by following this scheme, we cannot be sure that the resulting path does not violate the constraints dictated by the HMM architecture, i.e. choose a path which include transitions with zero probabilities. For instance, we could in certain cases have a protein labelled with two consecutive signal peptides!

The most widely used criterion for determining the "best" path is to find the most likely path through the sequence. Formally, we are looking for the state sequence Q which maximises the value of $P(O, Q / \lambda)$. This is solved using the Viterbi algorithm, which is a straightforward application of the dynamic programming technique. To find the single best path through the sequence $Q = Q_1 Q_2 ... Q_T$ for the given observation sequence $O = O_1 O_2 ... O_T$ we define the quantity:

$$\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} P(q_1, q_2, ..., q_{t-1}, q_t = S_i, O_1 O_2 ... O_t / \lambda). \tag{0.14}$$

By induction we have:

$$\delta_{t+1}(j) = \max_i \left(\delta_t(i) a_{ij} \right) b_j(O_{t+1}). \tag{0.15}$$

To retrieve the actual path, we need to keep track of the argument which maximises (0.15) for each *t* and *j*. This can be done through the variable $\psi_t(j)$.

Viterbi algorithm

1) Initialization:

$$\begin{cases} \delta_1(i) = \pi(i) b_i(O_1), & 1 \le i \le n. \\ \psi_1(i) = 0, & 1 \le i \le n. \end{cases} \tag{0.16}$$

2) Induction:

$$\begin{cases} \delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right] b_j(O_t), & 2 \le t \le T, \quad 1 \le j \le N. \\ \psi_t(j) = \arg\max_{1 \le i \le N} (\delta_{t-1}(i) a_{ij}), & 2 \le t \le T, \quad 1 \le j \le N. \end{cases} \tag{0.17}$$

3) Termination:

$$\begin{cases} P^* = \max_{1 \le i \le N} (\delta_T(i) \omega_i). \\ q_T^* = \arg\max_{1 \le i \le N} (\delta_T(i) \omega_i). \end{cases} \tag{0.18}$$

4) Path backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, ..., 1. \tag{0.19}$$

There are other ways to solve problem 2. For purposes of PH precursors (and signal peptides) prediction, I have implemented two algorithms, both close relatives of the Viterbi algorithm. Those modified versions of the Viterbi algorithm have been presented recently in the context of protein sequence analysis (Fariselli et al. 2005; Kall et al. 2005). I present first the decoding algorithm introduced by Fariselli and colleagues in the context of membrane

proteins prediction. This algorithm was named in their paper *posterior-Viterbi* (PV) algorithm. The second decoding algorithm is the *optimal accuracy* decoding algorithm described in (Kall et al. 2005). Other decoding algorithms have been described, including the *l-best* algorithm (Krogh et al. 1994a).

One way to find the "best" path through the sequence $(q_1, q_2, ...q_T)$ would be to maximise, for every residue t in the sequence the overall probability of being at t in the state q_t. Formally we are looking for:

$$\max_{1 \le i \le N} \left(P(q_t = S_i / O, \lambda) \right). \tag{0.20}$$

As mentioned previously, this can lead to paths which violate the grammar of the HMM, i.e. containing a path $Q = q_1 q_2 ... q_T$ such that there exists a t where $q_t q_{t+1} = S_r S_{r'}$ and $a_{rr'} = 0$. What we can do, is to incorporate those quantities $\gamma_t(i)$, which integrate information about different paths, in order to force the resulting best path to comply with the architecture constraints of the HMM. The two following algorithms, taken from (Fariselli et al. 2005) and (Kall et al. 2005), implement this idea.

2.3.3.3 Solving problem 2: Posterior-Viterbi (PV) decoding algorithm

The problem to be solved is the following:

$$Q^{PV} = \arg\max_{Q \in A_p} \left(\prod_{t=1}^{T} P(q_t / O, \lambda) \right), \tag{0.21}$$

where $Q = q_1 q_2 ... q_T$ and A_p is the set of allowed paths through the model.

Let us define the variable χ defined on the interval [0, 1] such that for the transition probability a_{ij} between two states *i* and *j*:

$$\begin{cases} \chi(a_{ij}) = 1 \ \ if \ \ a_{ij} > 0. \\ \chi(0) = 0. \end{cases} \tag{0.22}$$

χ is positive only for transitions which are allowed.

Assuming that the variable $\gamma_t(i)$ has been computed through the forward-backward algorithm we can write the posterior-Viterbi decoding algorithm:

Posterior-Viterbi decoding algorithm (Fariselli et al. 2005)

1) Initialization:

$$\begin{cases} \delta_1^{PV}(i) = \gamma_t(i), & 1 \le i \le N. \\ \psi_1^{PV}(i) = 0, & 1 \le i \le N. \end{cases} \tag{0.23}$$

2) Induction:

$$\begin{cases} \delta_{t+1}^{PV}(j) = \max_{1 \le i \le N}[\delta_t^{PV}(i)\gamma_t(j)\chi(a_{ij})], & 1 \le t \le T-1, \quad 1 \le j \le N. \\ \psi_{t+1}^{PV}(j) = \arg\max_{1 \le i \le N}[\delta_t^{PV}(i)\chi(a_{ij})], & 1 \le t \le T-1, \quad 1 \le j \le N. \end{cases} \tag{0.24}$$

3) Termination:

$$\begin{cases} P^{*PV} = \max_{1 \le i \le N}(\delta_T^{PV}(i)\omega_i). \\ q_T^{*PV} = \arg\max_{1 \le i \le N}(\delta_T^{PV}(i)\omega_i). \end{cases} \tag{0.25}$$

4) Path backtracking:

$$q_t^{*PV} = \psi_{t+1}^{PV}(q_{t+1}^{*PV}), \quad t = T-1, T-2, ..., 1. \tag{0.26}$$

The *Optimal Accuracy* (OA) decoder described in (Kall et al. 2005) is obtained from the PV decoder by replacing in equation (A) the product by a sum $\delta_{t+1}^{OA}(j) = \max_i[\delta_t^{OA}(i) + \gamma_t(j).\chi(a_{ij})]$. In the OA decoder the target function to be minimized is a sum:

$$Q^{OA} = \arg\max_{Q \in A_p} \sum_{t=1}^{T} P(q_t / O, \lambda). \tag{0.27}$$

Intuitively, those decoders should be able to increase the rate of positive cleavages found. Several PH precursors including gastrin and cholecystokinin precursors are sequentially processed at one end. The Viterbi procedure which attempts to find the best path, can only pull out the most likely cleavages among this set of alternative cleavages, and is thus fundamentally flawed for that type of predictions. I reasoned that the PV-decoder should do a better job than the Viterbi decoder in finding cleavages since the former takes into account all possible paths and not only the most likely one. However one can expect that this increase in sensitivity is accompanied by a decrease in specificity due to "overlabeling" of cleavage sites (cf. subsection 2.8.2.3). I will only present results pertaining to the PV decoder as it performed slightly better than the similar OA decoder for the problem at hand.

2.3.3.4 Solving problem 3: The Baum-Welch algorithm

Let $Q = q_1 ... q_T$ be a path of states through a given sequence O and P the set of possible paths through the sequence. Here we assume that the architecture of the HMM is known. For this given structure the Baum-Welch algorithm aims to estimate the parameters of the HMM so as to maximise the probability of the sequence O given the architecture α and parameters λ (formally $P(O/\lambda,\alpha)$).

The Baum-Welch algorithm relies on an expectation-maximisation procedure introduced in (Baum et al. 1970). At each step the parameters of the HMM are reestimated such that the posterior likelihood $P(O/\lambda,\alpha)$ is guaranteed to increase. We can define the Baum auxiliary function:

$$F(\lambda,\bar{\lambda}) = \sum_{Q \in P} P(O,Q/\lambda)\log(O,Q/\bar{\lambda}). \qquad (0.28)$$

We can show by using the concavity of the logarithmic function that:

$$F(\lambda,\bar{\lambda}) \geq F(\lambda,\lambda) \Rightarrow P(O/\bar{\lambda}) \geq P(O/\lambda). \qquad (0.29)$$

Proof:

Let $\Delta(\lambda,\bar{\lambda}) = F(\lambda,\bar{\lambda}) - F(\lambda,\lambda) = \sum_Q P(Q,O/\lambda)\log\left(\frac{P(Q,O/\bar{\lambda})}{P(Q,O/\lambda)}\right)$. $\Delta(\lambda,\bar{\lambda})$ is non-negative by hypothesis and can be re-written: $\Delta(\lambda,\bar{\lambda}) = A_Q \sum_Q \alpha_Q \log\left(\frac{P(Q,O/\bar{\lambda})}{P(Q,O/\lambda)}\right)$, with $A_Q = \sum_Q P(Q,O/\lambda) \geq 0$ and $\alpha_Q = \frac{P(Q,O/\lambda)}{\sum_Q P(Q,O/\lambda)} \geq 0$, with $\sum_Q \alpha_Q = 1$.

Since the logarithmic function is concave: $0 \leq \Delta(\lambda,\bar{\lambda}) \leq A_Q \log\left(\sum_Q \alpha_Q \frac{P(Q,O/\bar{\lambda})}{P(Q,O/\lambda)}\right)$

Hence: $1 \leq \sum_Q \alpha_Q \frac{P(Q,O/\bar{\lambda})}{P(Q,O/\lambda)} = \frac{\sum_Q P(Q,O/\bar{\lambda})}{\sum_Q P(Q,O/\lambda)} = \frac{P(O/\bar{\lambda})}{P(O/\lambda)}$ and $\boxed{P(O/\bar{\lambda}) \geq P(O/\lambda)}$.

We write the forward and backward variables that were introduced in 2.3.3.1:

Forward variable alpha:

$$\alpha_t(j) = P(O_1, O_2, ..., O_t, q_t = S_j / \lambda), \quad 1 \leq j \leq N, \quad 1 \leq t \leq T.$$

Backward variable beta:

$$\beta_t(j) = P(O_{t+1}, O_{t+2}, ...O_T / q_t = S_j, \lambda), \quad 1 \leq j \leq N, \quad 1 \leq t \leq T-1.$$

$\alpha_t(j)$ and $\beta_t(j)$ can be computed using the forward-backward algorithm, as shown previously. The reasons why those variables are used are twofold: first, because they can be efficiently computed ($O(2N^2T)$ operations) and second because all other relevant variables can be expressed in terms of the HMM parameters and those forward and backward variables. We can express the likelihood of the full sequence given the model in terms of $\alpha_t(j)$ and $\beta_t(j)$:

$$P(O/\lambda) = \sum_{j=1}^{N} P(O, q_t = S_j / \lambda) = \sum_{j=1}^{N} P(O_1, ...O_t, q_t = S_j / \lambda) P(O_{t+1}, ...O_T / O_1, ...O_t, q_t = S_j, \lambda) = \sum_{j=1}^{N} \alpha_t(j)\beta_t(j).$$

$$\Rightarrow\ P(O/\lambda)=\sum_{j=1}^{N}\alpha_t(j)\beta_t(j). \tag{0.30}$$

For every *t* and *i* the likelihood $\gamma_t(i)$ that state *i* is occupied at time *t* can also be expressed in terms of forward and backward variables:

$$\gamma_t(i)=P(q_t=S_i/O,\lambda)=\frac{P(q_t=S_i,O/\lambda)}{P(O/\lambda)}.$$

And the numerator of the fraction can be written:

$$P(q_t=S_i,O/\lambda)=P(O_1,...O_t,q_t=S_i/\lambda)P(O_{t+1},...O_T/q_t=S_i,\lambda)=\alpha_t(i)\beta_t(i).$$

And we have:

$$\Rightarrow\ \gamma_t(i)=\frac{\alpha_t(i)\beta_t(i)}{P(O/\lambda)}. \tag{0.31}$$

In the calculation of $\gamma_t(i)$, $P(O/\lambda)$ appears as a normalising factor. For every *t*, it is sufficient to calculate for each *i* the quantity $\gamma_t(i)P(O/\lambda)=\alpha_t(i)\beta_t(i)$ and then normalise it according to $\sum_{i=1}^{N}\gamma_t(i)=1$.

Likewise, for all *t*, *i* and *j*, the quantity $\xi_t(i,j)$, that represents the likelihood that the running state is S_i at *t* and S_j at *t*+1 given the model and the data, can be expressed simply with the parameters of the HMM and the forward/backward variables:

$$\xi_t(i,j)=P(q_t=S_i,q_{t+1}=S_j/O,\lambda)=\frac{P(q_t=S_i,q_{t+1}=S_j/O,\lambda)}{P(O/\lambda)}.$$

The numerator of this fraction can be expressed in terms of forward and backward variables:

$$\begin{aligned}P(q_t=S_i,q_{t+1}=S_j,O/\lambda)&=P(O_1,...,O_t,q_t=S_i/\lambda)\,P(O_{t+1},...,O_T,q_{t+1}=S_j/O_1,...,O_t,q_t=S_i,\lambda)\\&=\alpha_t(i)P(q_{t+1}=S_j/q_t=S_i)\,P(O_{t+1},...,O_T/O_1,...,O_t,q_t=S_i,q_{t+1}=S_j,\lambda)\\&=\alpha_t(i)a_{ij}P(O_{t+1}/q_{t+1}=S_j,\lambda)\,P(O_{t+2},...O_T/q_{t+1}=S_j,\lambda)\\&=\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j).\end{aligned}.$$

And we have:

$$\Rightarrow \ \xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O/\lambda)}. \tag{0.32}$$

In the calculation of $\xi_t(i,j)$, $P(O/\lambda)$ appears as a normalising factor. For every *t* and *i* it is sufficient to calculate for each *j* the quantity $\xi_t(i,j)P(O/\lambda) = \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$ and then normalise it according to $\sum_{j=1}^{N} \xi_t(i,j) = 1$.

Given the running model λ, we wish to find $\overline{\lambda}$ so that Baum's auxiliary function (0.28) is maximal. This guarantees an increase in the likelihood of the data by (0.29). The problem can then be re-stated as follows:

Maximise $F(\overline{\lambda}) = \sum_{Q \in P} P(O,Q/\lambda) \log(P(O,Q/\overline{\lambda}))$ under the constraints:

$$\begin{cases} \sum_{j=1}^{N} a_{ij} = 1, \quad 1 \le i \le N, \\ a_{ij} \ge 0, \quad 1 \le i, j \le N, \\ \sum_{s=1}^{S} b_i(s) = 1, \quad 1 \le i \le N, \\ b_i(s) \ge 0, \quad 1 \le i \le N, \quad 1 \le s \le S, \\ \sum_{i=1}^{N} \pi_i = 1, \quad 1 \le i \le N, \\ \pi_i \ge 0, \quad 1 \le i \le N. \end{cases}$$

This constrained optimization problem can be solved with the Lagrange multiplier technique and has the following analytical solution, known as the re-estimation formulae (proof not given):

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}. \tag{0.33}$$

$$\overline{b_j}(k) = \frac{\sum_{\substack{t=1 \\ s.tO_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}. \qquad (0.34)$$

$$\overline{\pi_i} = \gamma_1(i). \qquad (0.35)$$

Note that those formulae only involve the variables γ and ξ introduced previously. Note also those formulae can be interpreted intuitively. For instance, a_{ij}, the probability of transition from state S_i to state S_j, is re-estimated through the average value over *t* of probabilities of being in state S_i at *t* and state S_j at *t*+1 (term $\sum_{t=1}^{T-1} \xi_t(i, j)$) divided by the average likelihood of being in state *i* (term $\sum_{t=1}^{T-1} \gamma_t(i)$).

The idea behind the Baum-Welch algorithm, an expectation-maximisation algorithm, is to perform this operation iteratively until the likelihood of the data converges towards a local minimum. Each optimization iteration of the Baum-Welch training of an HMM involve two steps: in the expectation step of the algorithm, the forward-backward algorithm is performed and variables γ and ξ are computed using equations (0.31) and (0.32). In the maximisation step of the algorithm we look for parameters which increase the likelihood of the data by using the re-estimation formulae (0.33), (0.34) and (0.35). One of the most critical points that will decide for the success/failure of the algorithm, as in all expectation-maximisation algorithms, is the choice of the initial parameters. They must be chosen “not too far” from the global minimum in order to reach it.

Baum-Welch algorithm

1) Set a value for *threshold*, initialise matrices A, B and Π and compute $likelihood = P(O/\lambda)$. Set $\overline{likelihood} = likelihood + threshold$.

2) While $\left|likelihood - \overline{likelihood}\right| > threshold$ do:

- Compute variables $\alpha_t(i)$ and $\beta_t(i)$ using the forward-backward algorithm.

- Compute variables $\gamma_t(i)$ and $\xi_t(i, j)$ using formulae (0.31) and (0.32).

- Calculate $\overline{\lambda}$ using re-estimation formulae (0.33), (0.34) and (0.35).

- Set $likelihood = \overline{likelihood}$ and compute $\overline{likelihood} = P(O/\overline{\lambda})$, the likelihood of the data given the new model.

3) return model $\overline{\lambda}$.

2.3.4 Considerations on the implementation of the HMM algorithms

To implement the HMM algorithms, scaling of α and β variables in the forward-backward procedure as well as scaling of the δ variable in the decoding procedures were performed according to (Rabiner 1989), by taking the logarithms of quantities and transforming the products into sums.

To speed up the execution of the programs, logarithms of observation symbol probabilities ($\log\left(P(O_t/q_t = S_j)\right)$ for $1 \le t \le T$ and $1 \le j \le N$) were computed only once per sequence.

Some transitions are not allowed by the architecture of the HMM, and one can avoid to make unnecessary calculations corresponding to paths which contain such transitions, that give a trivial probability of 0. In order to do that, one can store in a table the possible transitions from and to each of the states. The two tables, condA and condAInv store that information:

For every $1 \le i \le N$ $CondA(i)$ is a vector containing all indexes $(j_k)_{1 \le k \le N_i}$ for which a transition $i \to j_k$ is allowed, if those indexes exist. Likewise, for every $1 \le j \le N$ $CondAInv(j)$ is a vector containing all indexes $(i_k)_{1 \le k \le N_j}$ for which a transition $i_k \to j$ is allowed, when those indexes exist. If most of the transitions are disallowed, as in our case, the

use of those functions considerably speed up the HMM algorithms. The recursion step of Viterbi, forward and Posterior-Viterbi decoding algorithms make use of the *CondA* table, while the recursion step of the backward algorithm involves the *CondAInv* table.

2.4 Building the PH HMM

2.4.1 Outline of the HMM strategy

Signals defining the characteristic PH sequence features (signal peptide, short peptides flanked by dibasic sites, absence of transmembrane domains) were analysed separately and combined into a single HMM (noted PHMM). To build PHMM sequences and annotations were retrieved from the Swiss-Prot and Ensembl databases (cf. Resources).

Parts of the HMM were directly estimated from Swiss-Prot annotations including state transition and observation symbol probabilities for the *peptide* and *propeptide* states. Those probabilities were inferred by looking at the length of those features (to infer transition matrix parameters cf. 2.4.4) or aa composition (to infer observation matrix parameters, cf. 2.4.5) in Swiss-Prot. Other parts of the model were learned in a semi-supervised way from the Baum-Welch procedure including some characteristics of the signal peptide model such as the structure of the hydrophobic region states (cf. 2.4.7.2) and size of the cleavage site model (cf. 2.4.8.3). All those parts of the model were then joined together according to the global PHMM structure described in the next paragraph. The main HMM programs were all written in java.

2.4.2 Overall architecture of PHMM

A PH precursor must contain a signal peptide at its n-terminus and at least one peptide and one cleavage site. It can contain two peptides following each other (separated by a cleavage site), but not two propeptides in a row. Furthermore a PH precursor can end either after a propeptide or a peptide region. It can also end on a cleavage site but only in the case that it follows a peptide region (e.g. atrial natriuretic factor precursor). The HMM structure depicted in Figure 4 translates those requirements in a parsimonious way.

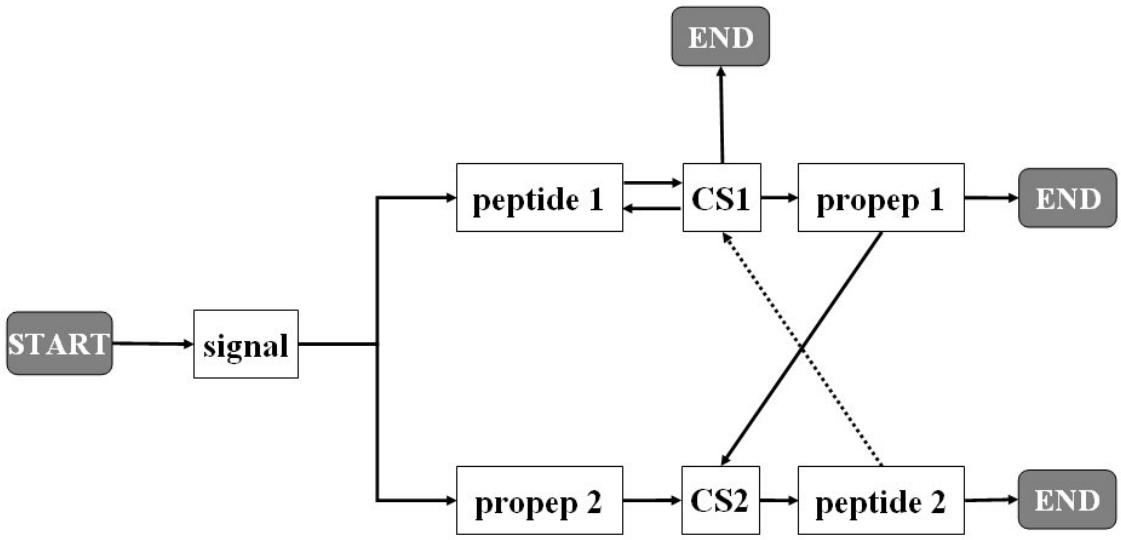

Figure 4: Architecture of the PH HMM
States corresponding to peptide (*peptide* 1 and 2), propeptide (*propep* 1 and 2) and cleavages sites (*CS1* and *CS1*) regions were duplicated as to force the HMM to label at least one peptide and a prohormone cleavage site. The *signal* box designates the signal peptide HMM which will be described in a later section (cf. 0). *CS1* and *CS2* denote two identical cleavage site HMM which will also be described in a later section (cf. 2.4.8.3).

Note that the architecture of Figure 4 does not integrate states useful for negative selection (transmembrane domains and mitochondrial transit peptides) and those states were only included when the HMM was used to screen the proteomes (2.9) but not for validating and comparing the models on the set of secreted proteins (section 2.8).

I now present the strategy I used to build each piece of the model.

1.1.1 Retrieving feature sequences using annotations from Swiss-Prot

Signal peptide, mitochondrial peptide, peptide, propeptide and transmembrane sequences were retrieved using the sequence retrieval system (SRS) of Swiss-Prot (http://expasy.org/srs5/, cf. Resources chapter). In order to retrieve signal, peptide and propeptide sequences, the field *FtKey* of the SRS interface was set to match respectively the values “SIGNAL”, “TRANSIT”, “PEPTIDE”, “PROPEP” and “TRANSMEM”, and the corresponding feature boundary indexes were retrieved. For each feature, an entry containing a reference to the human protein, its associated type (PEPTIDE, SIGNAL, etc…) and the corresponding boundary indexes was inserted into a table in a local MySQL database. Another table was then created in the database that contained Swiss-Prot/TrEMBL references

and sequences for all human proteins. For all features, sequences were then inferred from those two tables. Likewise, PC cleavage site sequences were logically deduced from the "PEPTIDE" boundary regions.

1.1.2 Logo motifs

Logo motifs are used to visualise position-specific distributions of aa. For each position in the alignment and for every aa the letter corresponding to that aa is added on top of a stack and drawn with a height proportional to the frequency of that aa in that position of the alignment. The total height of the stack of aa symbols is invsersely proportional to the entropy of $\sum_{s\in\{symbols\}} -P(s).\log\big(P(s)\big)$, where $P(s)$ is the probability of observing symbol *s* (Schuster-Bockler et al. 2004). Logo motifs give a faithful representation of both the conservation and the most frequent residues, at a given position in the alignment. However, for sequences of varying lengths, such as signal peptides, logo motifs do not capture the length variability of the alignment. A logo motif visualisation tool was implemented in java, based on modified classes from the bio.gui package of biojava. There are now publicly available softwares that produce logo motifs of sequences from multiple alignments (Schuster-Bockler et al. 2004).

2.4.3 Density estimations

In this subsection I present methods of distribution density approximations by geometric (2.4.3.1) and gaussian distributions (2.4.3.2), and the Parzen windows method (2.4.3.3).

2.4.3.1 Density estimation by a geometric distribution

A one-state HMM produces lengths following a geometric distribution (cf. Figure 5). Likewise, consider any state *s* of the PHMM. Let p_{stay} be the probability of transition from state *s* to the same state *s*.

Figure 5: Length distribution of sequences generated by a one-state HMM
A one-state HMM emits sequences of lengths that follow a geometric distribution.
The star (*) denote the end of the sequence (state *END*).

For an HMM containing that state *s*, the length distribution of any sequence of consecutive state *s* follows a geometric law of parameter p_{stay}:

$$P(length = n / state\, s) = p_{stay}^{n}(1 - p_{stay}) . \qquad (0.36)$$

Let $(x_1, x_2, ..., x_N)$ be a set of observed lengths of features corresponding to state *s*. We infer the value of p using the maximum likelihood principle:

$\log(Likelihood) = \sum_{i=1}^{N} [x_i \log(p_{stay}) + \log(1 - p_{stay})]$. A necessary condition is that:

$$\frac{d \log(Likelihood)}{dp_{stay}} = \frac{1}{p_{stay}} \sum_{i=1}^{N} x_i - \frac{N}{1 - p_{stay}} = 0 \text{ . Leading to: } p_{stay} = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i}{\frac{1}{N}\sum_{i=1}^{N} x_i + 1} .$$

The optimal transition probability of staying in state *s*, as defined by the maximum likelihood principle, is then:

$$p_{stay} = \frac{E(lengths)}{E(lengths) + 1}, \qquad (0.37)$$

where $E(lengths)$ is the mean value of observed feature lengths. p_{stay} is the probability which is used to estimate the transition probability from states *peptide* and *propeptide* to respectively states *peptide* and *propeptide*

2.4.3.2 Density approximation by a gaussian distribution

Let $(x_i)_{1\le i\le N}$ be an observed sample whose density we wish to estimate. Likewise we can show with the maximum likelihood principle that the best gaussian approximation of the density of x corresponds to a gaussian function $Gauss(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{x^2}{2\sigma^2})$ with parameters $m=E(x_i)$ and $\sigma=\left(1/N-1\right)E(E(x_i)-x_i)$.

2.4.3.3 Density approximation by the Parzen windows method.

Let $(x_i)_{1\le i\le N}$ be the observed sample (sizes or conservation scores). The approximated density function by this method is given by:

$$PW(x)=\sum_{i=1}^{N}F(x-x_i), \qquad (0.38)$$

where F is a smoothening function that verifies $\int_{-\infty}^{\infty}F(x)=1$ and $F(0)=\max_x F(x)$. Unless otherwise specified, the function F was chosen to be the Gaussian function centred on 0 and of variance $\sigma=0.5$,

$$F(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{x^2}{2\sigma^2}). \qquad (0.39)$$

In the final models all empirical sizes and conservation scores were approximated by the Parzen windows method.

2.4.4 Length distributions of propeptides, peptides, transmembrane domains and mitochondrial transit peptides

All human peptide, propeptide, transmembrane and mitochondrial transit peptides sequences were collected according to the strategy described in (1.1.1). Size distributions corresponding to each of the features were graphed in Figure 6 as histograms.

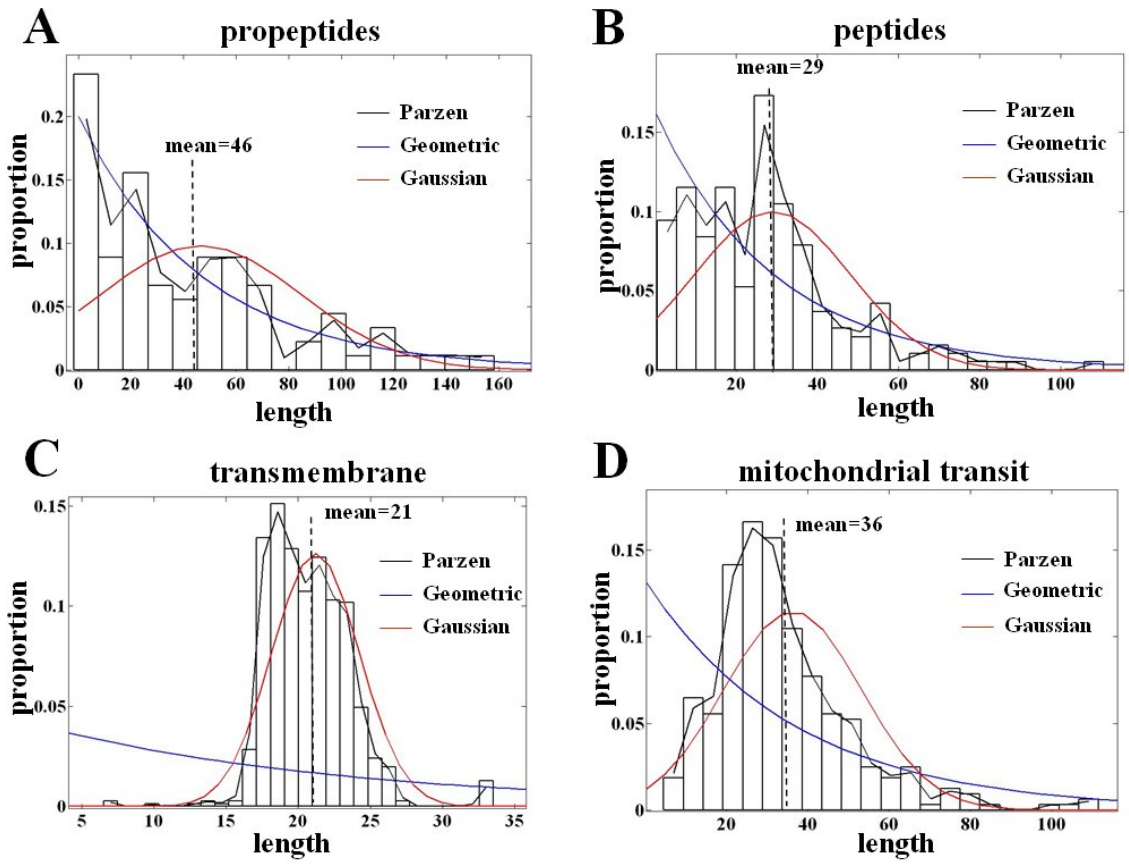

Figure 6: Distribution of propeptides, peptides, transmembrane domains and mitochondrial transit peptide lengths

Panels A, B, C and D show the histogram, Parzen windows (black curve), geometric (blue curve) and gaussian (red curve) approximations of the length distribution of respectively of 191 propeptides (A), 90 peptides (B), 709 transmembrane domains (C), 326 mitochondrial transit peptides (D) collected from the Swiss-Prot database.

All sequences were human and redundant sequences were removed from the sets.

Figure 6 shows that propeptides on average are clearly larger than peptides. Furthermore it shows that propeptides (panel A) and peptides (panel B) observed lengths (represented by a histogram) can be reasonably well approximated by a geometric distribution (blue curves). This size difference can readily be used in the HMM to help it distinguish between propeptide and peptide regions. This is the rationale behind modelling *propeptide* and *peptide* regions using only one cyclic state. Boxes representing those regions in the global HMM of Figure 4 were set as being one cyclic state. The looping probability p_{stay} was set according to formula (0.37). In contrast, transmembrane domains and mitochondrial transit peptides lengths follow more closely a gaussian (red curve) and for this reason those domains will be less accurately modelled using a single state. However, since the focus of my work was not on modelling these domains (which is only useful for PH discovery as a negative selection process), I decided to model those regions the simplest way possible by using only one HMM state for

transmembrane domains and 4 states for mitochondrial peptides (as residues -2 and -3 are described to be the important residues for cleavage, (Liu et al. 2003b)). In section 2.4.7 I present a more detailed study on signal peptide accurate length modelling using repetitive identical sate with a fixed architecture, as modelling of this structure proved much more critical to the success of my search algorithm.

2.4.5 Estimation of peptide and propeptide symbol observation probabilities

Peptide and propeptide symbol observation matrix probabilities were estimated by simple counting of the different aa frequencies for the two features. Given the large amount of data (more than 2700 residues for the *peptide* state) there was no need to make use of pseudo-counts as is often the case for profile HMM parameter estimation (Durbin et al. 1998).

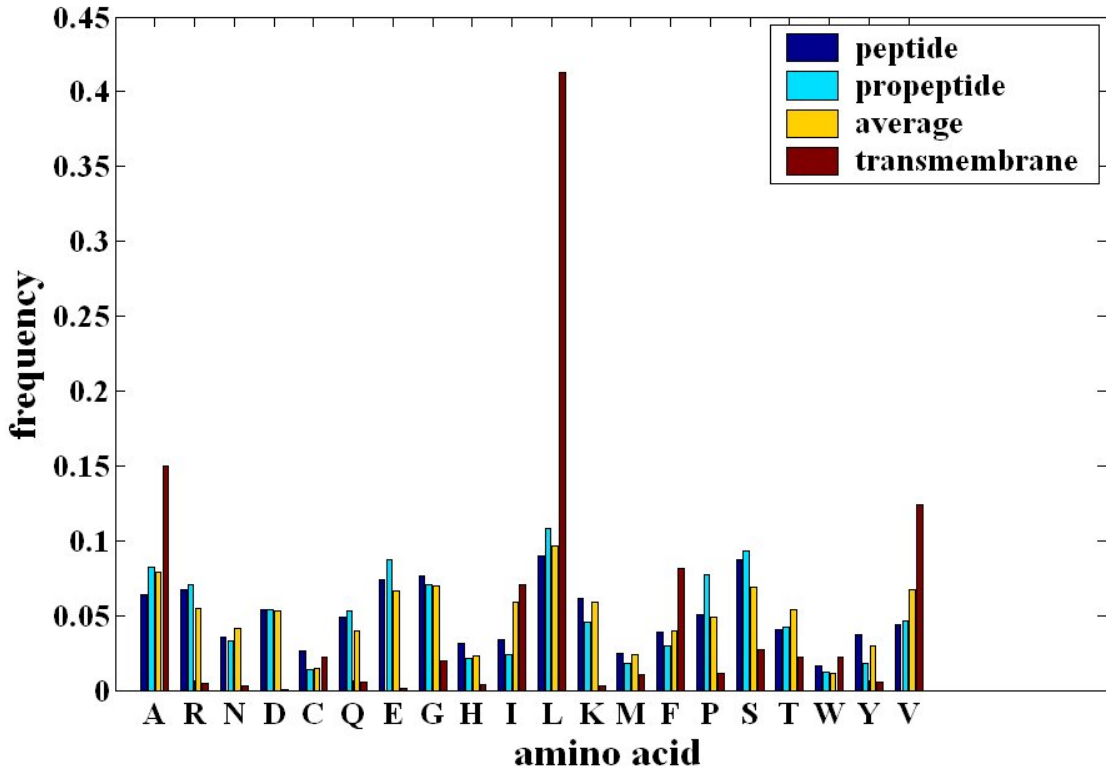

Figure 7: Estimated amino acid frequencies for PHMM
Estimated frequencies of amino acids for states *peptide* (dark blue), *propeptide* (turquoise) and *transmembrane* domains (brown). "average" denote average amino acid frequencies on the human proteome [source: Swiss-Prot]. The amino acids whose frequency significantly differs between peptides and propeptides appear to be cysteines (C) and tyrosines (Y).

2.4.6 Strategy used to refine models using sequence data from aligned features

A MySQL database was designed to build alignments of specific features of proteins. First, tables containing known protein sequences identified by their Ensembl peptide ID were built for human, mouse, opossum, frog, chicken, fugu, tetraodon, zebrafish, and stickleback fish. Second, tables linking every human protein sequence to orthologs from the 7 non-human organisms were built.

Third a script generated FASTA files containing orthologous sequences for every human protein. Those files were then used to generate alignments with the ClustalW program (with GAPOPEN=5, the other parameters set to their standard value). An annotation table from Swiss-Prot determining the indexes defining features in proteins that I was interested in were downloaded in Swiss-Prot. Those features included transmembrane domains, peptide,

propeptide, signal peptide and mitochondrial transit peptides. This table was then used to generate sub-alignments corresponding to features of interest (Figure 8).

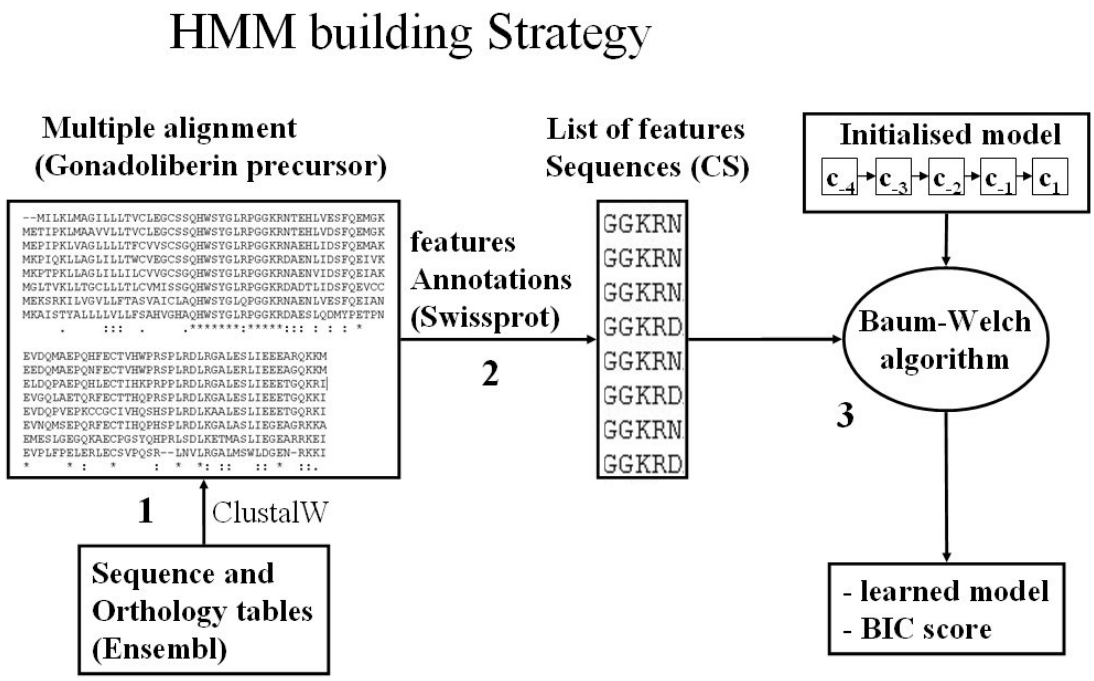

Figure 8: Strategy used for building HMM of features
Alignments of protein sequences from human, mouse, opossum, frog, chicken, tetraodon, zebrafish and stickleback fish were produced with the software ClustalW (1). For all PH, list of features sequences (cleavage site sequences for instance) were generated using annotations from Swiss-Prot (2). This list of sequences was then used in conjunction with an initial model to construct an HMM of that feature (3), and generate a BIC (Bayesian Information Criteria) score designed to rate how good the model is. The algorithm used is an expectation-maximisation procedure called the Baum-Welch algorithm.

I then used the Baum-Welch algorithm to refine the signal peptide and cleavage site models as described in the coming sections.

2.4.7 Modelling signal peptides using HMM

Signal peptides can be described in terms of primary sequence, as containing three parts: one N-terminal part (denoted –n part) of varying length containing positively charged residues, a core hydrophobic part (denoted –h part) of length which typically range between 6 and 15 aa and a more polar c-terminal end of varying length followed by the cleavage site itself, with residues at position -3 and -1 relative to the cleavage site being the most important ones (Gascuel and Danchin 1986; von Heijne 1986).

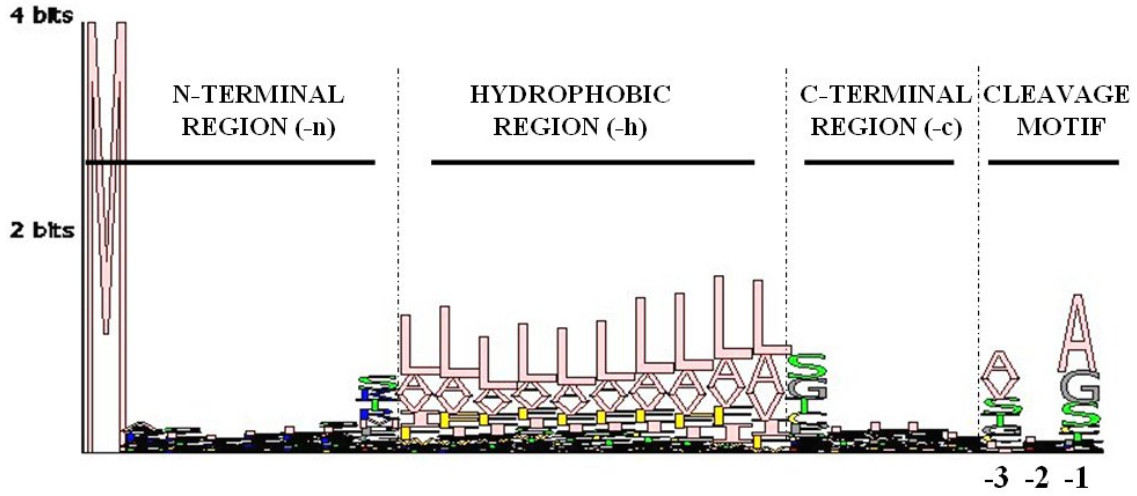

Figure 9: Logo motif of Eukaryotic signal peptides
The logo motif was generated with a java application built on modified classes from biojava packages. The program draws the logo motif of an alignment of sequences. This logo motif illustrates the basic structure of signal peptides: an N-terminal region of variable length and containing basic residues (Arginine/R and lysine/K, rendered as blue letters), a central hydrophobic region typically containing from 6 to 15 hydrophobic aa, a flexible c-terminal region and the signal peptide cleavage site typically containing small apolar residues at positions -1 and -3 relative to the cleavage site.

In order to generate the logo motifs of Eukaryotic signal peptides, a set of previously curated Eukaryotic non-redundant signal peptides was used, available at http://www.cbs.dtu.dk/ftp/signalp/euksig.red. Signal peptide hydrophobic regions were identified by a Perl script as the stretch of aa along the sequence with the highest percentage of hydrophobic aa. Known cleavage site indexes and predicted –h regions were sufficient to define the alignment, which was used as an input for the logo motif drawing application (Figure 9).

2.4.7.1 Building a preliminary signal peptide HMM

The same script was later used to derive symbol observation matrices (Figure 10) and length distributions (Figure 12) of signal peptide-related HMM states (of basic structure defined by Figure 11). Figure 10 shows the amino acid frequencies of the 4 regions associated to signal peptides (n and c- terminal and hydrophobic regions and cleavage site motif) making up for a total of 6 HMM states. Both Figure 9 and Figure 10 consistently highlight known characteristics of signal peptides: the N-terminal part of the signal peptide (Figure 10, dark blue) sees a larger number than expected amount of lysines and arginines, the core hydrophobic part of the signal peptide (Figure 10, light blue) is mainly composed of hydrophobic amino acids (leucines, alanines, valines and phenylalanines,…), at position -1

relative to the signal peptide cleavage site (SPCS-1, brown colour in Figure 10) there is a majority of alanines and glycines while at position -3 (Figure 10, yellow) there is a majority of alanines, valines and serines.

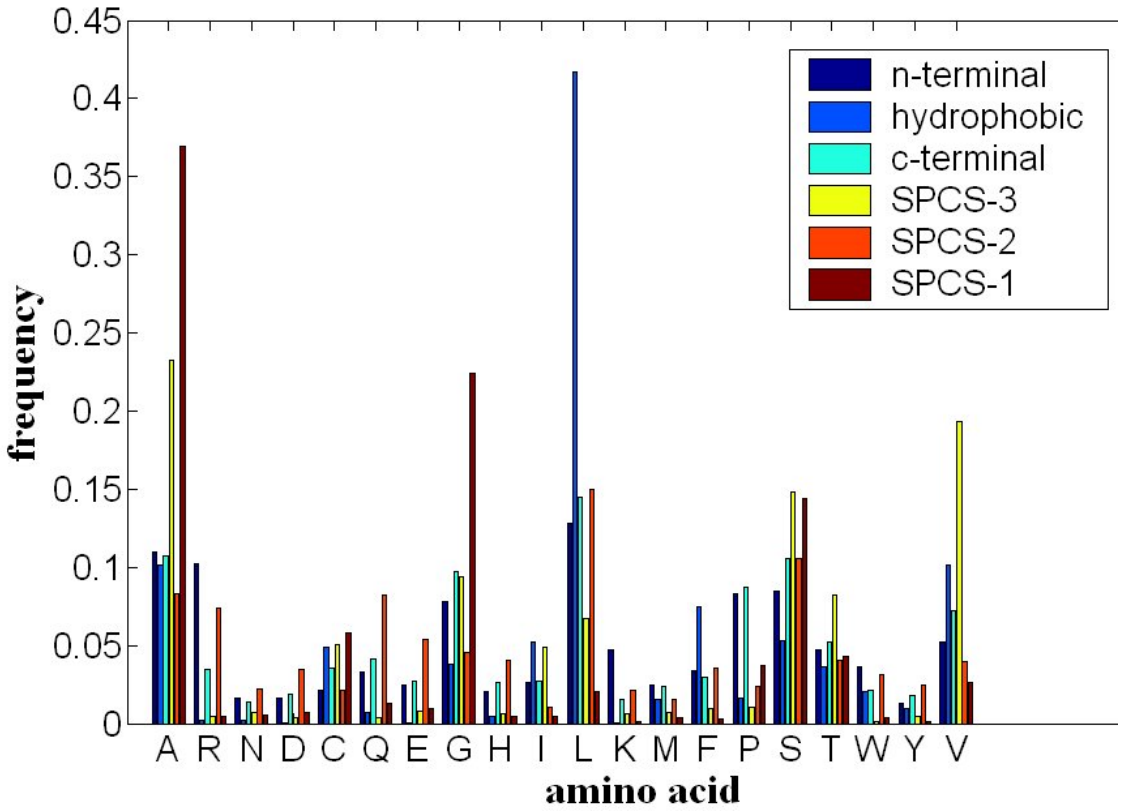

Figure 10: Estimated amino acid frequencies associated to signal peptides
Estimated amino acid composition for N-terminal, hydrophobic, c-terminal and cleavage site motif (SPCS-3, SPCS-2 and SPCS-1) of signal peptides.

A signal peptide HMM was designed to be integrated into the larger PH HMM. Figure 11 shows the architecture of a preliminary signal peptide HMM derived from what is known about signal peptide structure (von Heijne 1986) and inspired by the SignalP-HMM architecture as described in (Nielsen and Krogh 1998). The general architecture of the signal peptide HMM (referred to as signal-HMM from here on) largely followed the one described in (Nielsen and Krogh 1998). States modelling –n, -h, and –c terminal regions were included as well as the three states before the signal peptide cleavage site S_{-1}, S_{-2} and S_3. Since the "(-1,-3) rule" was originally formulated in (von Heijne 1986) we know that most of the information about substrate recognition by signal peptidases are found in residues at position -3 and -1 relative to the signal peptide cleavage site. -n, -h and -c regions symbol observation matrices and transition matrices were derived using the Baum-Welch algorithm with initial

transition and observation matrices corresponding to those determined previously with the Perl script that defined the core hydrophobic stretch of the signal peptide.

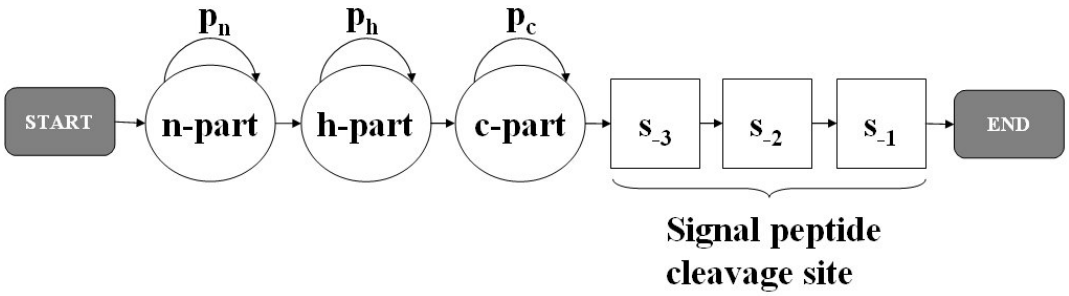

Figure 11: Preliminary architecture of signal peptide HMM
This preliminary signal peptide HMM is constructed using 6 states, plus the "start" and "end" states. The first state, "n-part" models the N-terminal part of signal peptides, as defined in (von Heijne 1986). The states "h-part" and "c-part" aim to model, respectively, the hydrophobic and c-terminal parts of signal peptides. The probabilities of staying in those states (termed p_n, p_h and p_c) control the length distribution of those regions. States c_{-3}, c_{-2} and c_{-1} model the conserved motif found before the signal peptide cleavage site. Note that only the c_{-3} and c_{-1} states are informative. States are represented by a square when they model a fixed residue on a functional motif such as in a proteolytic cleavage site. States are represented by a circle when they model a domain of varying length and given aa frequencies such as the hydrophobic core of a signal peptide.

Figure 12 shows the lengths distribution of –n, -h and –c regions of signal peptides, as labelled by the Viterbi algorithm for the preliminary signal peptide HMM of Figure 11. The –n and –c regions have length distributions which roughly match that of a geometric law and hence will be modelled reasonably well using only one cyclic state. However, a single cyclic state is likely to be a poor model of the –h region lengths empirical distribution, as the latter resembles more a bell-shaped Gaussian.

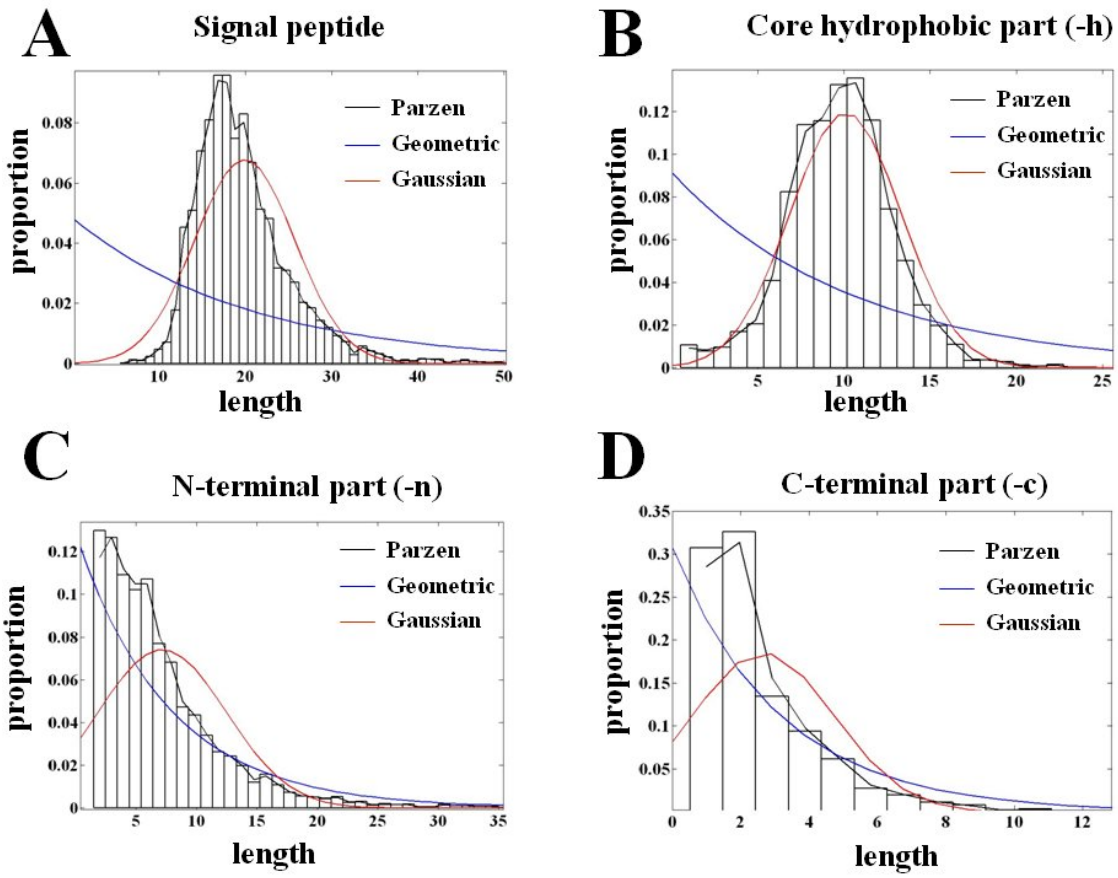

Figure 12: Length distribution of signal peptide features estimated with the HMM
Lengths distributions of the full signal peptide (A), hydrophobic (B), N-terminal (C) and c-terminal (D) parts of signal peptides for 4105 signal peptide sequences extracted from aligned signal peptides of human, mouse, opossum, chicken, frog, fugu, stickleback and zebrafish (cf. 2.4.6). Those features were estimated from the Viterbi labelling of the HMM. Black, red and blue curves are estimations of the underlying density function of those lengths distributions. The black curve was generated with a Parzen windows approximation, while the blue and red curves are respectively the geometric and Gaussian function which maximise the likelihood of the data. Note that lengths distributions of signal peptide and hydrophobic parts are similar to a Gaussian curve, while the n- and c-terminal parts have lengths which are better modelled with a geometric distribution.

For this reason, I decided to modify the architecture of this preliminary signal peptide HMM to better account for the length distribution of the core hydrophobic region of signal peptides.

2.4.7.2 Optimizing the signal peptide HMM

I have sought here to set the topology of the signal peptide HMM on a rigorous basis by using the BIC rating tool described in paragraph 2.4.8.4.

Figure 12 shows that a simple 6-state HMM is not able to accurately model the lengths distribution of signal peptides. In particular, the hydrophobic region of the signal peptide follows more closely a gaussian than a geometric distribution (Figure 12B). To address this

problem, the architecture of the signal peptide HMM was modified as to incorporate multiple identical hydrophobic states (Figure 13). Signal peptides all contain more than 6 residues (Figure 12A) because it must have a core hydrophobic part of minimal length (Nielsen and Krogh 1998). The first 4 "h" states were placed to translate the idea that a signal peptide hydrophobic core is always longer than 4 residues (Figure 12A). One could have chosen a larger threshold of 5 or even 6 residues for the hydrophobic region minimal length. The next $r+1$ states were introduced to approximate the Gaussian-like lengths distribution (Figure 13). Transition probabilities from state h_5 to the other h states were set identical one to another (bound to be so) and equal to $\alpha = 1/(r+1)$ (Figure 13).

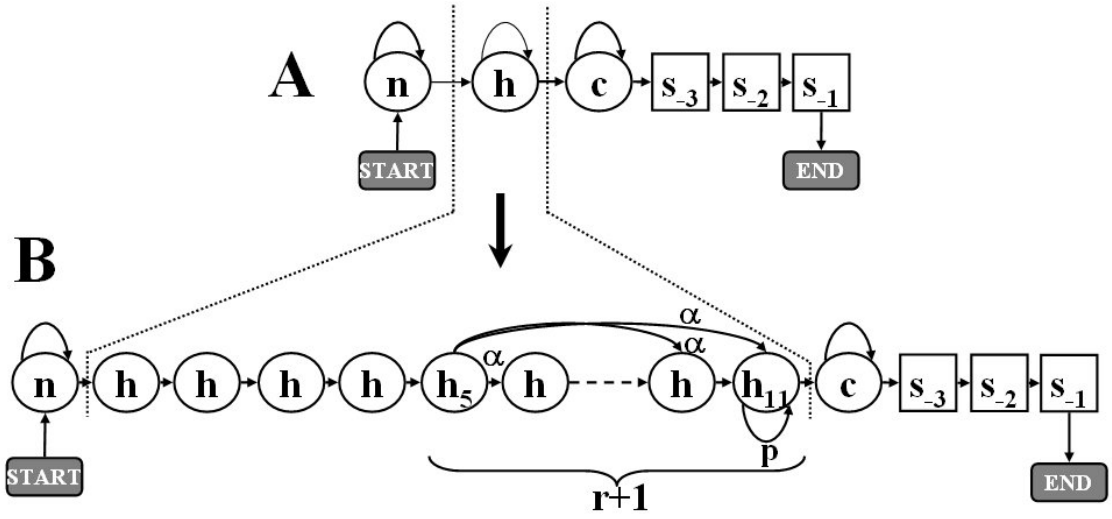

Figure 13: Architecture of the signal peptide HMM
To better model lengths of the core hydrophobic region of signal peptides, the single hydrophobic state from the initial model (A) was replaced by multiple identical hydrophobic states with a fixed architecture (B).
α is the transition probability from state 5 to the following hydrophobic states, and is equal to $1/(r+1)$. The "last" hydrophobic state h_{11} is a state which can label multiple consecutive residues. p denotes the associated probability of "staying" in state h_{11}.

The best number r and the corresponding probability p were estimated by Maximum Likelihood. Let us calculate the theoretical length distribution $P(x=n)$ generated by an HMM formed by the r+1 states between h_5 and h_{11} from the model shown in Figure 13B (where "start" and "end" states were added respectively before h_5 and after h_{11}). We have:

$$P(x=n) = p^{n-1}\alpha(1-p) + \ldots + p^0\alpha(1-p) = \alpha(1-p)\sum_{k=0}^{n} p^k = \alpha(1-p^{n+1}) \qquad if\ 1 \le n \le r.$$

$$P(x=n) = p^{n}\alpha(1-p) + \ldots + p^{n-r}\alpha(1-p) = \alpha(1-p)\sum_{k=n-r}^{n} p^k = \alpha p^{n-r}(1-p^{r+1})\ \ if\ n \ge r+1.$$

(0.40)

We can verify that $P(x=n)$ is continuous at the point $n=r$ and that $\sum_{n=1}^{\infty} P(x=n) = 1$.

Equations **(0.40)** were used to calculate the empirical likelihood of the observed hydrophobic region sizes. The number maximum(4,size) was subtracted from each of the observed sizes to account for the first 4 hydrophobic states of model Figure 13B. For r ranging from 1 to 20 the transition probability p was determined in $[0,1]$ as the value maximising the likelihood of sizes given the model. Figure 14 summarizes the results obtained by following this procedure. I find that the best HMM following the architecture described in Figure 13B corresponds to $r = 7$ and $p = 0.65$ (cf. Figure 14).

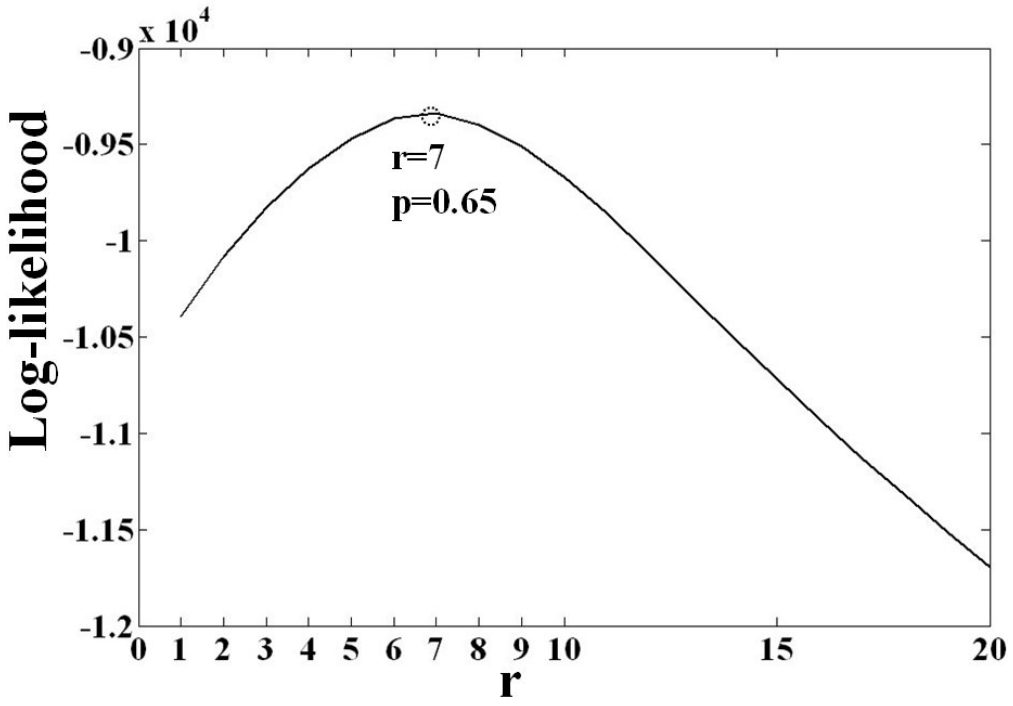

Figure 14: Determination of parameter r by the maximum likelihood method

The curve gives the log-likelihood of the data as a function of r. $r = 7$ was found to maximise the likelihood of the data (lengths of hydrophobic regions) given the model (of Figure 13B). The corresponding transition probability p of staying in state h_{11} was found to be 0.65.

The theoretical distribution defined by equations **(0.40)** for $r = 7$ and $p = 0.65$ was then graphed along with the histogram of sizes to verify that it constituted a good approximation of the underlying size probability density function (Figure 15).

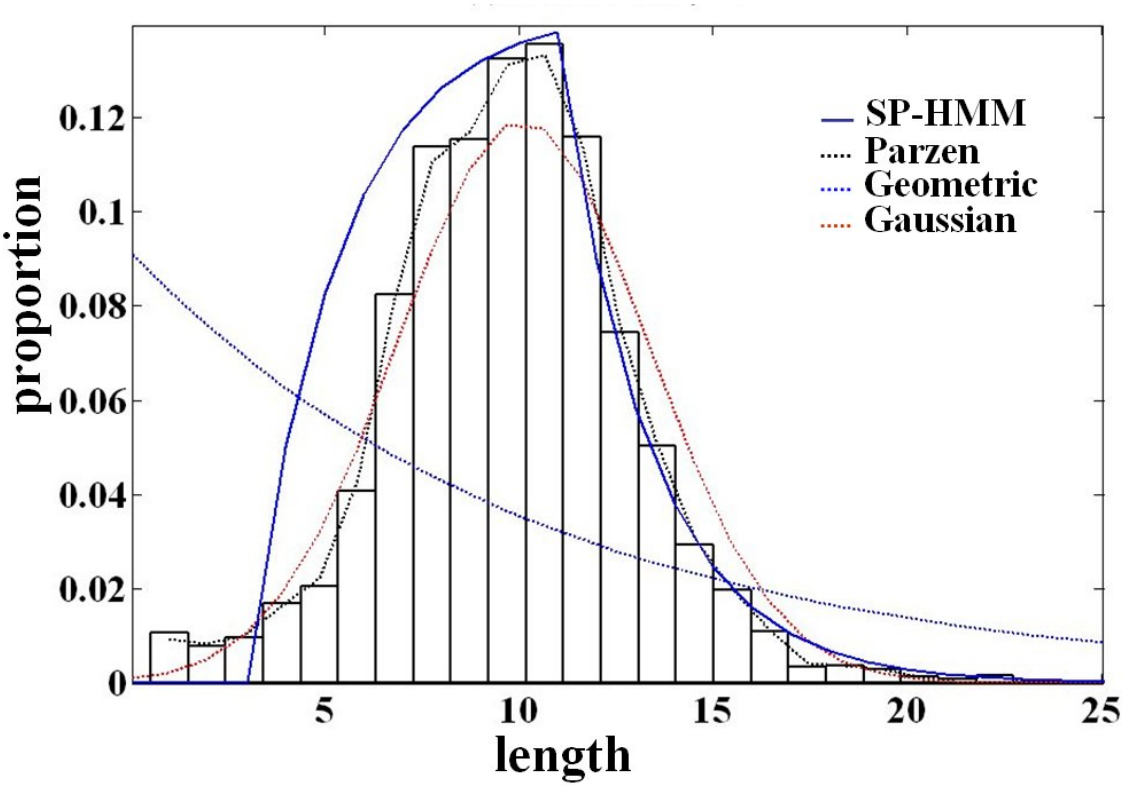

Figure 15: Estimation of the hydrophobic region lengths distribution
The histogram represents sizes of 4105 observed signal peptide hydrophobic regions. The dotted black line represents the approximated density of those sizes by the Parzen windows method. Red and blue dotted lines represent the best approximations of size distribution density by respectively Gaussian and geometric distributions. The blue line (SP-HMM) represent the theoretical density of hydrophobic region sizes generated by the HMM of Figure 13B (with $r = 7$ and $p = 0.65$). Note that this distribution is a relatively good approximation of the data and fits much better the observed sizes (histogram) than does the geometric distribution (dotted blue line) that corresponds to the initial model of Figure 13A.

2.4.8 Modelling PC cleavage sites

PC cleavage sites are generally defined as short motifs (4-6 aa) which contain a dibasic site. Furin and convertases form a group of 7 proteases belonging to the subtilisin/kexin-like family (cf. introduction 2.1.1.3).

2.4.8.1 Furin sites

Among known prohormone convertases, Furin is the best characterized in terms of its substrate recognition sites. Those furin cleavage sites have been described as sites following the rule Arg-X-Arg/Lys-Arg↓ with additional requirements at position -6 relative to the cleavage site. However not all Furin cleavage sites follow that rule. The minimal requirement for a motif to be recognized by furin seems to be Arg-X-X-Arg↓ (Thomas 2002). Another set

of rules states that at an arginine at position -1 relative to the cleavage site and a basic residue at least two of the three residues at -2, -4 and -6 relative to the cleavage site are required for efficient cleavage by Furin, and that hydrophobic aliphatic aa are ruled out at position +1 relative to the cleavage site (Nakayama 1997).

2.4.8.2 Prohormone convertase sites

Traditionally, substrate recognition sites for the larger prohormone convertases family is described as (Lys/Arg-X-Lys/Arg-Arg-↓-X) (Steiner 1998). However, the diversity of cleavage sites from that family of proteases has recently become more evident with so-called atypical prohormone cleavage site possessing only one basic residue at position -1, such as those found in proghrelin (Pro-Arg↓), and proapelin (Ser-Arg↓). The task of finding a clear consensus sequence for cleavage sites of the larger prohormone convertases family has proved much harder than for the more homogeneous furin cleavage sites (Duckert et al. 2004; Steiner 1998). Some bioinformatics tools have been developed to predict such motifs including ProP (cf. Resources) where researchers tried to predict separately furin and PC1/2 motifs. Figure 16 shows the logo motif of eukaryotic PC cleavage sites.

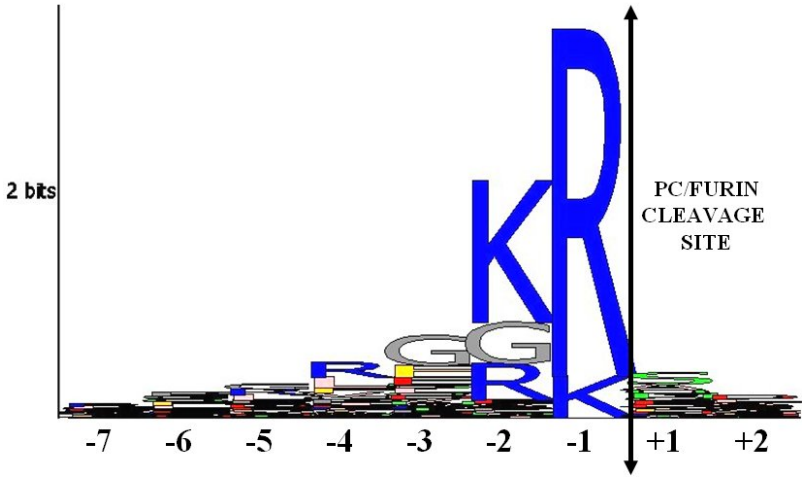

Figure 16: Logo motif of non-redundant eukaryotic PC/furin cleavage sites
PC/furin cleavage sites were collected from boundary regions of known peptides, aligned with ClustalW and visualised as a logo motif using a program written in java. Motifs that did not contain a basic residue at position -1 of the cleavage site were discarded.

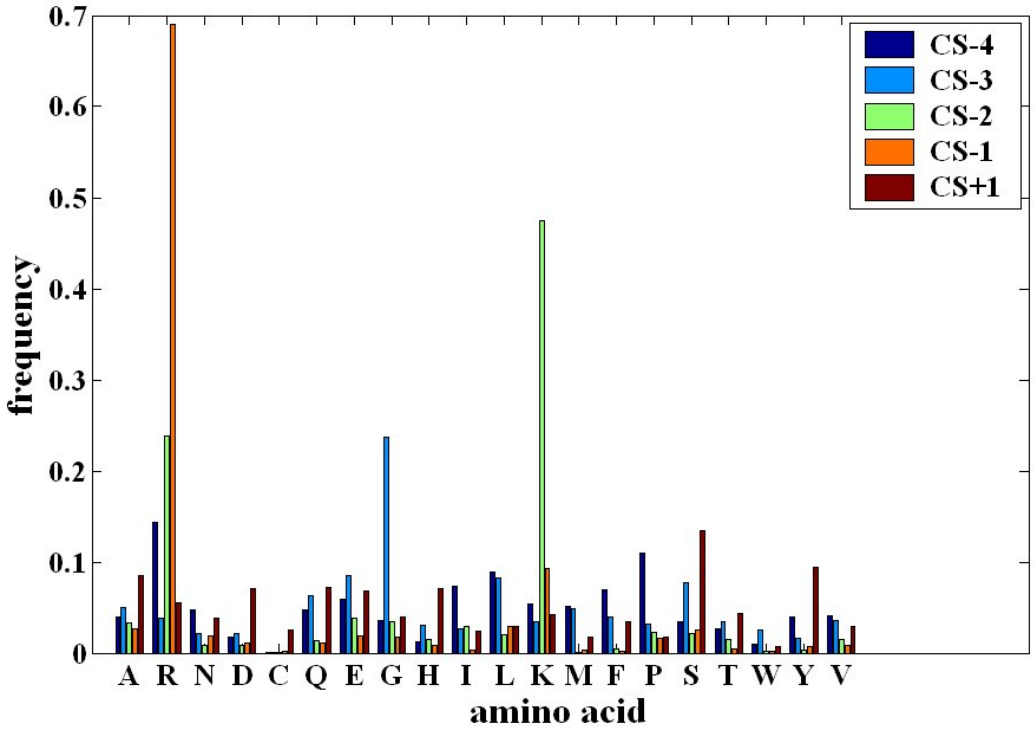

Figure 17: Estimated amino acid frequencies associated to PC cleavage sites
Estimated amino acid composition in positions -1 to -4 N-terminal of PC cleavage sites (CS-4, CS-3, CS-2 and CS-1) and position +1 C-termnal of PC cleavage sites (CS+1).

2.4.8.3 Selection of PC/furin cleavage site size

I wanted to determine which motif around the PC/furin cleavage site was most informative. For this I considered two parameters *n* and *p* corresponding to the number of aa respectively before and after cleavage (cf. Figure 18). I rated each of the (*n*,*p*) models using a modified likelihood which measures how likely the sequence is to be generated by the model and penalises for the complexity of the model (cf. next paragraph 2.4.8.4).

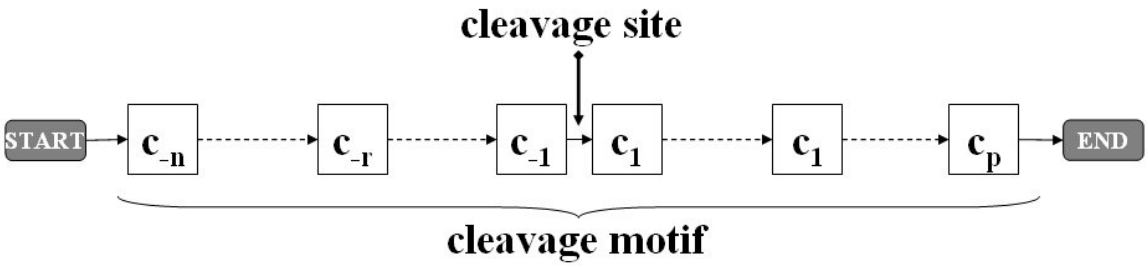

Figure 18: Simple weight matrix-like HMM of prohormone convertase cleavage sites
prohormone convertase cleavage sites are modelled with motifs of sizes (n,p) corresponding to the number of residues respectively before and after cleavage. Cleavage happens between sites c_{-1} and c_1 and residues (typically dibasic) corresponding to states c_{-2} and c_{-1} are removed by proteases.

From the literature (2.4.8.1 and 2.4.8.2) and the logo motif Figure 16, the residues N-terminal of PC cleavage sites (positions "-") seems more important than those c-terminal of cleavage sites (positions "+"). More precisely *n* was expected to be comprised between 4 and 7 and *p* was expected to be 1 or 2.

When models with growing *p* and *n* were tested, the size of the motif grew as *n*+*p* and the likelihood of the sequences given the data then decreased mechanically. The likelihood must then be normalised to cancel this unwanted size effect. A one state HMM with emission probabilities equal to the average frequencies of aa found in Swiss-Prot was used as a normalising null model (Figure 19). The effective log-likelihood that was used in the BIC formula of paragraph 2.4.8.4 was thus equal to the log-likelihood of sequences given the model to be tested minus the log-likelihood of the data given the null model.

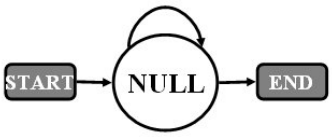

Figure 19: HMM null model
The HMM null model consists in a cyclic one-state HMM with observation symbols probability distribution corresponding to the average frequencies of aa observed in the human proteome

2.4.8.4 Selection of models through penalisation of complexity

This paragraph presents tools that answer the following question: how does one choose the best architecture of HMM In general choosing the best model within an increasing family of

models leads to an increase in the likelihood of the data (given this best model). For instance one has more chances to find a function that fits the data in a large family of functions than in a small family of functions. In order to overcome this problem, one must find a way to penalise for the complexity of the HMM architecture. This is a very common issue in the field of artificial intelligence/machine learning and is known as the model selection problem. Essentially, there are two types of strategies to address this issue: algorithms which penalise for complexity of the model directly, and strategies which leave out some data for testing the models, and evaluate the errors of the model on that data. In order to select the best architecture and size of cleavage sites HMM and signal peptide HMM, the Bayesian Information criterion (BIC) was used (Hastie et al. 2001). The BIC states that the "best model" is the one which minimises the following *BIC* quantity:

$$BIC = -2\log Likelihood + d\log(p). \tag{0.41}$$

where p is the number of sequences and d the number of free parameters in the family of models.

In the case of HMM we can calculate d:

$$d = \left[r^{\pi} - 1\right] + \left[N(M-1)\right] + \left[\sum_{i=1}^{N}(r_i^a - 1)\right]. \tag{0.42}$$

where the first term $\left[r^{\pi} - 1\right]$ translates the contribution of independent non-zero initial states probabilities π_i, the second term $\left[N(M-1)\right]$, where M is the number of symbols, makes up for emission probabilities a_{ij} and the last term $\left[\sum_{i=1}^{N}(r_i^a - 1)\right]$, where r_i are the number of non-zero transition probabilities from state i to the other states, represent the contribution of transition probabilities. Model selection is then performed by choosing the model with the lowest BIC score. Other information criteria include the Aikake Information criterion (AIC) and vary only slightly from the BIC.

2.4.8.5 Determination of the optimal PC/furin motif through BIC score minimization

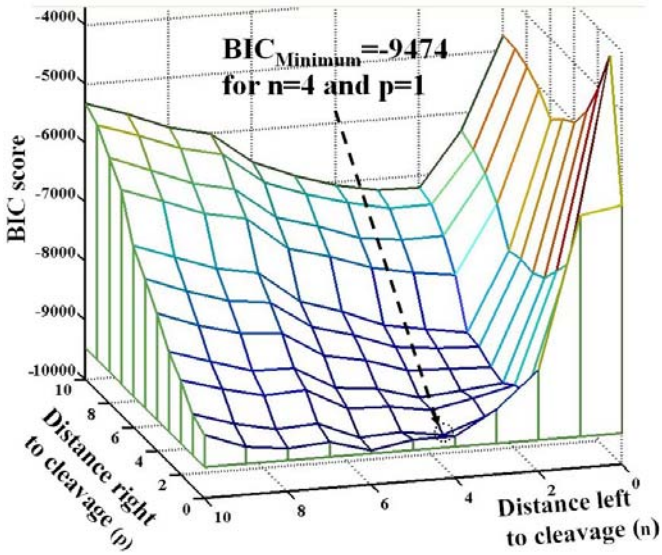

Figure 20: BIC scores of cleavage site-HMM for different positions/sizes of the motif
For HMM representing motifs characterized by their distances left and right of the cleavage site, BIC scores were computed using the formula introduced in paragraph 2.4.8.4. As expected motifs including several residues left of the cleavage site make better HMM (blue region of low BIC scores). And, as literature suggests, residues right are largely uninformative. The "best" motif is 5 residues-long and starts 4 aa N-terminal of the cleavage site (*n*,*p*)=(4,1). The cleavage site HMM corresponding to this motif was the one implemented in the global PH-HMM

Sequences containing gaps were discarded altogether. Considering the "gap" as a symbol was considered, but when that was tried, the models likelihood and BIC score kept increasing for large values of *p*, due to the fact that the model was simply fitting the increasing gap proportions that were found when considering residues remote from cleavage sites.

Unsurprisingly, the most informative motif, i.e. the one which minimized the *BIC* (0.41), was found to be (*n*,*p*)=(4,1) (Figure 20).

The best model is the one corresponding to a distance N-terminal to the cleavage site equal to 4 and a distance c-terminal from the cleavage equal to 1 (Figure 20) and this is the model that was implemented in the global PH HMM .

Other models with a more connected structure (feedforward and clique topologies) were tested and those modifications lead to a general increase in the BIC scores. From this point on we will consider that the cleavage site HMM is the one pictured in Figure 21, termed (4,1).

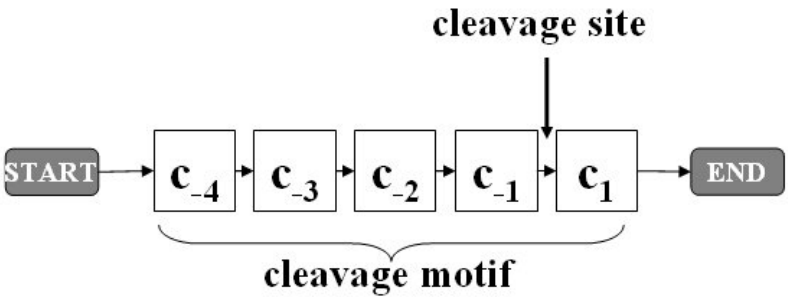

Figure 21: Cleavage site HMM (4,1)
The cleavage site HMM has 5 consecutive states which model the 4 residues directly N-terminal to the cleavage site and one residue at its c-terminal side.

Another question that could have been addressed using this CS model selection strategy was that of finding multiple independent motifs in the set of CS. This is an interesting question to follow up on because finding those non linear features could greatly improve the global HMM and also potentially provide hypotheses regarding PC substrate specificities.

At this point we consider the architecture of PHMM to be fixed (as in Figure 4) with its signal peptide HMM and cleavage site HMM topologies set respectively as in Figure 13B and Figure 21. Furthermore, all transition and observation symbol probabilities have been estimated from respectively feature lengths and aa frequencies, as explained in subsections 2.4.3, 2.4.4 and 2.4.5. I know present two types of slight variations on standard discrete HMM to take into account orthology information from alignments of human proteins with vertebrate orthologous proteins.

2.5 Alignment-based HMM (A-HMM)

With the growing availability of annotated genome sequences, techniques integrating homology information for improving prediction accuracy in computational biology have become commonplace (Kall et al. 2005; Majoros et al. 2005; Pedersen and Hein 2003; Siepel et al. 2005; Wu and Haussler 2006).

Residues from other organisms which are aligned to human sequences carry important information which can be advantageously used. Intuitively by integrating homology information we do not expect to increase the rate of true positive found, but hope to lower the rate of false positives. For instance PH have dibasic cleavage sites in human sequences which are aligned with dibasic residues from orthologous sequences, hence those true cleavage sites

are likely to be recognized by the A-HMM and the sequence is likely to score well. In contrast, those human sequences which "by chance" happen to have dibasic cleavage sites which are just part of a charged stretch of positively charged residues or alpha-helix for instance, will not necessarily align to dibasic residues in the alignment. Hence we expect the dibasic residue site not to be recognized by the HMM, and the protein to score poorly.

2.5.1 Notations and algorithms

Let $(O_{st})_{\substack{1\le s\le S\\ 1\le t\le T}}$ be a multiple alignment of length T, and of size S, constructed as described previously.

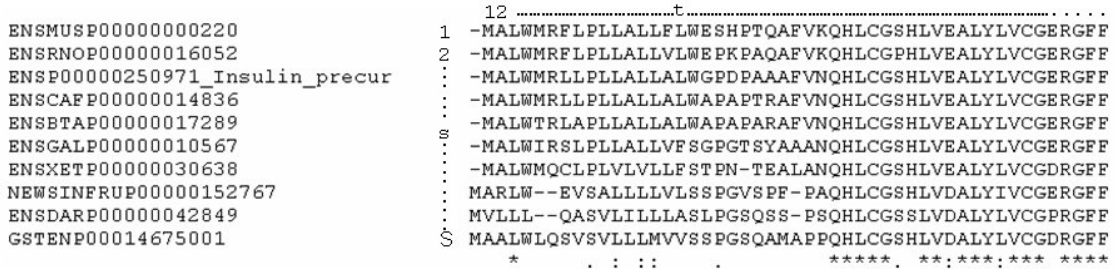

Figure 22: Truncated multiple alignment of insulin sequence orthologs.

Labelling and scoring of a multiple alignment using the A-HMM method involves the replacement of the scalar human residue O_t^{human} at each position t a by a vector $O_t^{alignment} = (O_{1t},...,O_{S_t t})$ representing all non-gap aligned residues at position t of the human sequence. The observation probabilities $b_j(O_t^{human})$ at position t for the human sequence are then replaced by the geometric mean $\left[\prod_{s=1}^{S_t} b_j(O_{st})\right]^{1/S_t}$ in the Viterbi and forward-backward procedures.

A simple way to take into account all aligned residue is as follows:

We replace in the decoding algorithms each single residue observation probabilities

$P(O_t / q_t = S_j) = b_j(O_t)$ by the geometric mean of non-gap residue observation probabilities at the given aligned position t $\left[\prod_{s=1}^{S_t} b_j(O_{st})\right]^{1/S_t}$, where S_t is the number of non-gap residues at position t. Assuming that $t = t_1$ corresponds to the first column from the alignment where the human residue is not a gap, the Viterbi procedure can be written:

Viterbi for alignment-based HMM (A-HMM)

Initialization:

$$\begin{cases} \delta_1(i) = \dfrac{\pi_i}{S_1}\left[\prod_{s=1}^{S_1} b_j(O_{s1})\right]^{1/S_1}, & 1 \le i \le n. \\ \psi_1(i) = 0 \end{cases} \tag{0.43}$$

Induction:

$$\begin{cases} \delta_t(j) = [\max_{1\le i\le N}(\delta_{t-1}(i)a_{ij})]\left[\prod_{s=1}^{S_t} b_j(O_{st})\right]^{1/S_t}, & 2 \le t \le T, \quad 1 \le j \le N. \\ \psi_t(j) = \arg\max_{1\le i\le N}(\delta_{t-1}(i)a_{ij}), \end{cases} \tag{0.44}$$

Termination:

$$\begin{cases} P^* = \max_{1\le i\le N}(\delta_T(i)\omega_i). \\ q_T^* = \arg\max_{1\le i\le N}(\delta_T(i)\omega_i). \end{cases} \tag{0.45}$$

Path backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, ..., 1. \tag{0.46}$$

The same modifications apply for writing the forward-backward algorithm in the context of the alignment-based HMM, i.e. observation probabilities $P(O_t / q_t = S_j) = b_j(O_t)$ are replaced by the geometric mean of non-gap residue observation probabilities at the given aligned position t $\left[\prod_{s=1}^{S_t} b_j(O_{st})\right]^{1/S_t}$.

2.5.2 Issues and remarks

The gain of information from using sequences which are far apart is balanced by the difficulty of aligning the sequences accurately in an automated way.

One should also envisage using a weighted sum of orthologs. If organisms are close, using sequences from both organisms will bias the calculations toward them. There are strategies to weight sequences in alignments so as to avoid this caveat (Valdar 2002a).

One might also want to weight organisms according to the overall quality of their annotations. Generally, one should put more trust in sequences from organisms where expressed sequence tags (ESTs) databanks are larger, and of better quality. One could take into account those different criteria and devise a rigorous method to pick the organisms which should be used, in order to maximise the information gained while keeping the computational cost low.

Handling gaps in the sequence is not a trivial issue. I chose to discard the parts of the alignment where residues from other organisms were aligned with gaps in the reference sequence (the human sequence in this case). However, gaps can be considered as symbols in the HMM and different states can have different probabilities of generating gaps. I did not follow this strategy because I reasoned that the probable resulting gain in sensitivity of the A-HMM would be outweighed by a likely rise in its sensitivity to annotation and alignment errors.

2.6 Conservation score and alignment-based HMM (CSA-HMM)

In this paragraph I present the development of HMM algorithms that use a conservation score index along the sequence. Aligned genomic sequences have already been used to predict genes structure more accurately (Chu et al. 2006) and phylogenetic HMM already provide a sound framework (Felsenstein and Churchill 1996; Siepel et al. 2005) for using orthology information in HMM. However, I did not follow those methods and surely the phylo-HMM framework presented in (Siepel et al. 2005) could be well adapted for modelling alignments of peptide hormone precursors. Instead I followed a more simple and straightforward idea: by integrating a conservation index along the alignment the HMM would be much more apt at detecting boundaries between peptide and propeptides. My hope was that conservation arguments would make up for the lack of sequence information relevance around the site of

cleavage (untypical non-dibasic cleavage sites) and make it possible for the HMM to detect untypical non-dibasic cleavage sites.

So the question at this point is the following: how to derive a conservation score on an alignment? Good estimation of conservation along alignments is not trivial and has been the focus of an entire PhD thesis (Valdar 2002b).

2.6.1 Calculation of a conservation score along the sequence

There is a rather large body of literature concerning quantification of functional sites conservation. The first step for the calculation of the conservation score is to normalise the substitution matrix.

2.6.1.1 Choice of substitution matrix

There are several matrices which are available. Among those matrices are the PAM matrices, the Dayhoff matrices, PAM and BLOSUM (Altschul 1991). They are constructed using alignments of known homologous regions of proteins (domains for examples). The matrix I used, BLOSUM62, is built using alignments with at least 62% homology and is the matrix used in the BLAST software package.

2.6.1.2 Normalisation of substitution matrix

The goal is to transform a substitution matrix B such that the distances $B(s_i, s_j)$ are nonnegative numbers, and smaller than 1. In addition, all $B(s_i, s_i)$ are required to be equal to 1:

$$\begin{cases} 0 \le B(s_i, s_j) \le 1, \\ B(s_i, s_i) = 1, \end{cases} \qquad 1 \le i, j \le S. \tag{0.47}$$

Option 1:

One way to achieve this is to make the following linear transformation:

$$\forall\ 1 \le i, j \le N \qquad \hat{B}_{linear}(s_i, s_j) = \frac{B(s_i, s_j) - \min_{s \in S} B(s_i, s)}{\max_{s \in S} B(s_i, s) - \min_{s \in S} B(s_i, s)}. \tag{0.48}$$

However, this matrix is not symmetrical, a property which is required for distance matrices. One way to make it symmetrical is to construct a positive definite matrix $\hat{B}_{positive}$ such that:

$\hat{B}_{positive} = \hat{B}_{linear}.(\hat{B}_{linear})'$, where $(\hat{B}_{linear})'$ is the transposed matrix of $\hat{B}_{linear}$.

Option 2:

Another option would be to first make the transformation:

$$\forall\ 1 \le i, j \le N \qquad \hat{B}_{symmetrical}(s_i, s_j) = \frac{B(s_i, s_j)}{\sqrt{B(s_i, s_i).B(s_j, s_j)}}. \tag{0.49}$$

followed by the same linear transformation as described by equation (0.48):

$$\forall\ 0 \le i, j \le 1 \qquad \hat{B}_{Karlinke}(s_i, s_j) = \frac{\hat{B}_{symmetrical}(s_i, s_j) - \min_{s \in S} \hat{B}_{symmetrical}(s_i, s)}{\max_{s \in S} \hat{B}_{symmetrical}(s_i, s) - \min_{s \in S} \hat{B}_{symmetrical}(s_i, s)}. \tag{0.50}$$

(Karlin & Brocchieri, 1996).

Both strategies were tried and I opted for the second option since it already been proposed and seemed to perform well.

2.6.1.3 Defining a conservation score

The problem is the following: we wish to calculate a conservation score on a column of aligned residues $S = \{s_k\}_{1 \le k \le N}$. One of the simplest ideas is to take the normalised similarity matrix and calculate an average distance between any two aa:

$$C_{average} = \frac{2}{N_{NOTGAP}(N_{NOTGAP} - 1)} \sum_{\substack{i \prec j \\ s_i, s_j \neq gap}} \hat{B}(s_i, s_j). \tag{0.51}$$

where N_{NOTGAP} is the number of non-gapped residues in the column.

$C_{average}$ is the conservation measure that was implemented for the PH CSA-HMM.

2.6.1.4 Improvements/Issues

There are a number of issues that have not been addressed in this thesis regarding the calculation of a conservation score on a column of aligned residues.

First, weighing residues contribution according to how much the sequence they belong to differs from the other sequences will surely improve the quality of the conservation score (Valdar 2002a). Simply put, if two of the aligned sequences are very close from each other, only one of the two should contribute in the actual calculation of conservation scores. Hence the importance in my case of not considering organisms too "close" from each other (human and monkey for instance).

The entropy is a good natural measure to account for the diversity of aa. However, classical entropy calculated on a distribution is blind to knowledge we have on structural/chemical relatedness of aa among each other. Mixture of entropy and similarity-based measures have been proposed to measure conservation (Valdar 2002a).

Gaps were not accounted for in the calculation of conservation scores. Gaps may give us information about conservation of a site, but it may also just account for other things like alternative splicing or missing exons. This alignment-based method is already quite sensitive to alternative splicing mispairings and annotation errors and it is therefore unclear whether or not the model would have been better if the HMM had included gaps in the computation of conservation scores.

2.6.2 Estimation of feature conservation score distributions

2.6.2.1 *Peptide* vs. *propeptide* conservation score distributions

Figure 23 shows the conservation score distributions of peptide and propeptide regions in known human PH precursors, using the score defined in subsection 2.6.1. A clear difference can be seen between the two distributions, justifying the initial idea of CSA-HMM. As expected, peptide regions are more conserved as a whole (mean of peptide conservation=0.78>0.70= mean of propeptide conservation). There are nearly twice as many residues which are found perfectly conserved (conservation=1) in peptide regions (proportion=35%) compared to propeptide regions (proportion=18%). There are also more residues in proportion which are found poorly conserved in propeptide regions (23% have conservation scores less than 0.5) compared to peptide regions (16% have conservation scores

less than 0.5). Those observations support the claim that conservation is an important discriminating variable between peptide and propeptide regions. Incorporating this variable into the HMM should definitely improve overall HMM performance.

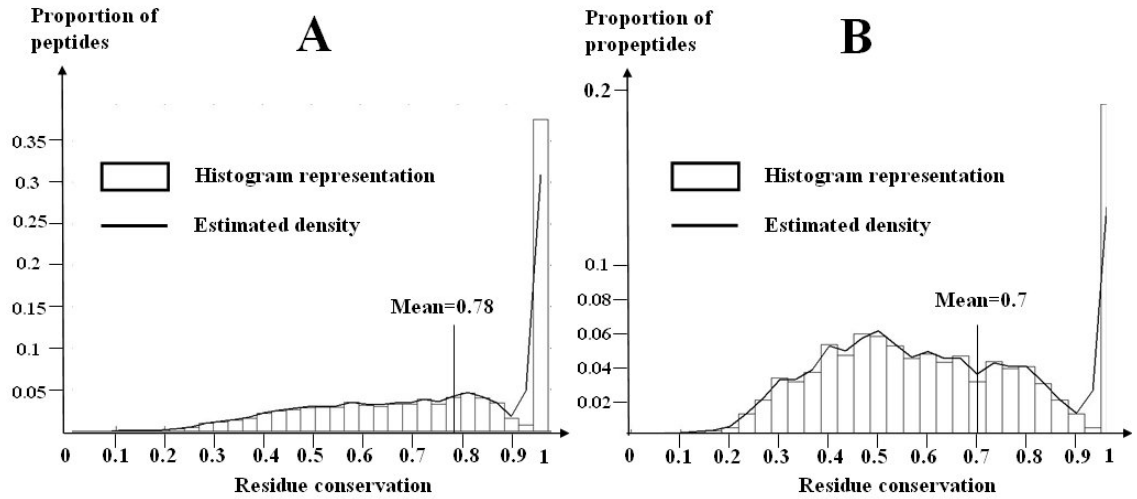

Figure 23: Conservation score distributions for peptide and propeptide regions.
Conservation scores were computed along portions of human PH precursors alignments which were annotated as peptide (A) and propeptide (B) domains. Distributions of peptide and propeptide conservation scores were drawn as histograms and estimated density distributions were computed using the Parzen windows method (Gaussian kernels, with a variance sigma=0.5). Conservation data from 191 human peptides and 90 human propeptides were used to generate those graphs. This corresponded to 5523 conservation scores for peptide regions and 4195 conservation scores for propeptide regions.

2.6.2.2 Estimation of all feature conservation score distributions

All the other states had their conservation score distribution estimated using alignments of features derived from the strategy described in subsection 2.4.6. Conservation score distributions were approximated by the Parzen windows method (cf. 2.4.3.3), they were discretized into 20 intervals ([0,0.05],[0.05,0.1],…[0.95,1]) and included as additional model parameters in the HMM.

2.6.3 Incorporation of a conservation score in the HMM

One can readily incorporate this new variable (a conservation score computed according to equation (0.51)) into the HMM algorithms: let us consider that the space of observed variables instead of being composed by the 20 aa $V = \{v_1, v_2, ..., v_M\}$ defined in 2.3.1, is now the space $V \times X_{cons}$ where X_{cons} is a set of intervals spanning [0,1] (conservation scores have

their values in this interval) and × is the canonical Cartesian product between two sets. Let us define $X_{cons} = \{[0, 0.05], ..., [0.95, 1]\}$, since this is how CSA-HMM was implemented in practice (dividing the interval [0,1] into 20 parts). If we consider as a first approximation that observation probabilities and conservation score probabilities are independent from each other we can write:

$P(O_t, x_t / q = S_i) = P(O_t / q = S_i) P(x_t / q = S_i)$. More generally we can write, $P(O_t, x_t / q = S_i) = \rho\left(P(O_t / q = S_i) P(x_t / q = S_i)\right)$ where $\rho : [0,1]^2 \mapsto [0,1]$. Then for each *t* and *i* the $P(O, x / q = S_i)$ need to get normalised so that: $\sum_{O,x} P(O, x / q = S_i) = 1$.

The difficulty at this point lies in choosing an appropriate function ρ. For making our predictions we now rely on two types of information: information about the primary sequence (aa composition and feature length distributions) and information about residue conservation among orthologs. The function ρ may set the relative importance of those two criteria for making our prediction calls (scoring and labelling). It may control the trade-off between the importance of primary sequence information and conservation information in the HMM. For instance, if the importance of conservation increases in the function ρ, then the HMM will b better at finding less-canonical cleavage sites such as those found in ghrelin, cholecystokinin, neuropeptide RF-amide and urocortin 2 precursors. However, it will also typically tend to label as cleavage sites regions of the protein corresponding to nucleic acid splicing sites.

2.6.4 CSA-HMM algorithms

In the next pages I give the modified Viterbi and forward-backward algorithms in the case where a conservation score along the alignment is explicitly used (CSA-HMM). Those are nearly identical to the standard ones, with the only difference being that the observation symbol likelihood at *t* $P(O_t / S_i)$ is replaced with the quantity $\rho\left(P(O_t / S_i), P(x_t / S_i)\right)$.

Viterbi algorithm for the CSA-HMM

1) Initialisation:

$$\begin{cases} \delta_1(i) = \rho\left(P(O_1 / S_i), P(x_1 / S_i)\right), \\ \psi_1(i) = 0, \end{cases} \quad 1 \le \mathrm{i} \le \mathrm{N}. \tag{0.52}$$

2) Recursion:

$$\begin{cases} \delta_t(j) = [\max_{1 \le i \le N}(\delta_{t-1}(i)a_{ij})]\rho\left(P(O_t / S_i), P(x_t / S_i)\right), \\ \psi_t(j) = \arg\max_{1 \le i \le N}(\delta_{t-1}(i)a_{ij}), \end{cases} \quad 2 \le t \le T, \ 1 \le j \le N. \tag{0.53}$$

3) Termination:

$$\begin{cases} P^* = \max_{1 \le i \le N}(\delta_t(i)\omega_i). \\ q_T^* = \arg\max_{1 \le i \le N}(\delta_t(i)\omega_i). \end{cases} \tag{0.54}$$

4) Path backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \ \ t = T-1, T-2, ..., 1. \tag{0.55}$$

Likewise, we introduce the conservation score in the Forward-Backward algorithm by replacing the observation scores $P(O_t / S_i)$ with $\rho\left(P(O_t / S_i), P(x_t / S_i)\right)$:

Forward algorithm for the CSA-HMM

Initialization:

$$\alpha_1(i) = \pi(i)\rho\left(P(O_1 / S_i), P(x_1 / S_i)\right), \quad 1 \le i \le N. \tag{0.56}$$

Induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] \rho\left(P(O_{t+1} / S_j), P(x_{t+1} / S_j)\right), \quad 1 \le t \le T-1, \quad 1 \le j \le N . \qquad (0.57)$$

Termination:

$$P(O / \lambda) = \sum_{i=1}^{N} \alpha_T(i) \omega_i . \qquad (0.58)$$

Backward algorithm for the CSA-HMM

Initialization:

$$\beta_T(i) = \omega_i, \quad 1 \le i \le N . \qquad (0.59)$$

Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} \rho\left(P(O_{t+1} / S_j), P(x_{t+1} / S_j)\right) \beta_{t+1}(j), \quad t = T-1, T-2, ..., 1, \quad 1 \le i \le N . \qquad (0.60)$$

I looked for the best function ρ in the family $\rho = xy^k$. I tried several values of k in the range of [0,1], and found that $k = 0.25$ seemed to give the best results. Values of k around or above 1 worsens the predictor as the conservation information takes on all the importance in the HMM. This value, set in a very heuristic manner, translates the fact that the HMM is putting more weight on the aa distribution information than the conservation information, at a ratio of 4:1.

2.6.5 Issues with the CSA-HMM method

The quality of alignments greatly influences performance of the predictor. There are two cases where the result is biased. First, when a large portion of the alignment only has the

human sequence. Then by defaults the conservation score along the sequence is equal to 1, which will bias the classifier towards labelling peptide and cleavage sites. The second case is when, because of alternative splicing events, a portion of human sequence is missing in the alignment. This is the case for instance for ENSP00000350269/FGF5 where a peptide is labelled by chance because it is only a small portion of a larger secreted protein.

2.7 Scoring PH precursor candidate sequences

The focus of this section is to address the problem of finding a suitable scoring system that allows strong discrimination between PH precursors and non-PH hormone precursors.

2.7.1 The overall probability score

One seemingly valid approach is to use the likelihood of the sequence under the model (given by the forward-backward algorithm) divided by the length of the sequence (for normalisation purposes): $Score_{overall} = \frac{\log(P(O/\lambda))}{length}$. Other normalisation schemes were tried, including dividing by the likelihood of the sequence under a null model (instead of the sequence length), that did not improve on this *overall* score

The *overall* score is a natural choice, but it is nevertheless not entirely satisfying. For instance, it will tend to give short signal peptide-containing sequences a very good score. In the end, there are entire pieces of the sequence which we do not wish to give too much importance to (like *propeptide* regions). What we are looking for are sequences with a "good" signal peptide, and at least one "good" peptide.

2.7.2 The best peptide score

The best peptide is not necessarily one of the peptide labelled by the Viterbi procedure. To increase the chances to label accurately the best peptide and to give it an appropriate score, the following empirical procedure was done. To compute the *bestpeptide* score (called $Score_{bestpeptide}$) corresponding to the best peptide, we use the precomputed gamma score from the forward-backward algorithm ($\gamma_t(j)$, cf. equation (0.13)). We then select the r sites t_k with the highest $\gamma_{t_k}(i = CLEAVAGE)$ corresponding to the positions where cleavage sites have more chances to occur. Those sites are considered to be potential cleavage sites. *r=8* was chosen. Three factors/scores are then taking into account: how high the numbers

$\gamma_{t_k}(i = PEPTIDE)$ are on average in the region between any two consecutive cleavage sites taken from the list of potential cleavage site generated in 2). How the length of this region fits the empirical peptide length distribution, and last, how likely this peptide is to be flanked at least on one side by a propeptide region.

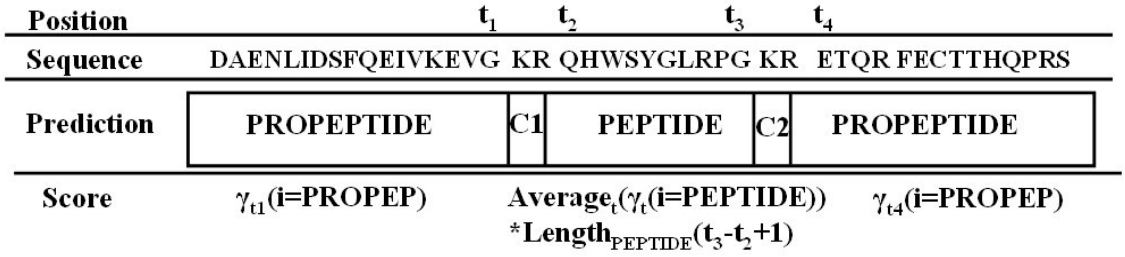

Figure 24: *Best peptide* scoring scheme diagram
This diagram showing how the hybrid *bestpeptide* score is constructed. A "good peptide" must have the right aa composition, the right length, and at least one propeptide region flanking it. The peptide is scored by averaging along its region the gamma score corresponding to the peptide state ($Average_t(\gamma_t(i = PEPTIDE))$) and multiplying it by the empirical lengths distribution of peptides applied to the length of the region ($Length_{PEPTIDE}(t_3 - t_2 + 1)$). For a good *bestpeptide* score the peptide is required to be flanked by at least one propeptide region. One can translate this requirement by adding in the bestpeptide score the quantity $Max(\gamma_{t_1}(i = PROPEP), \gamma_{t4}(i = PROPEP))$. t_2 and t_3 are position in the sequence flanking the peptide region, while t_1 and t_4 are positions inside the propeptide region (chosen to border the peptide region).

The *best peptide* score is then constructed as follows:

$$score_{bestpeptide} = \log(Average_t(\gamma_t(i = PEPTIDE))) + \log(Length_{PEPTIDE}(t_3 - t_2 + 1))$$
$$+ \log(Max(\gamma_{t_1}(i = PROPEP), \gamma_{t4}(i = PROPEP)))$$

2.7.3 Expectation of the number of cleavages

A "good" PH precursor can be considered to have either one "good" peptide or a high probability to have several cleavages, potentially giving rise to several peptides. Some PH precursors contain many different peptides, including prothyroliberin, proopiomelanocortin, proglucagon, proenkephalin and prodynorphin precursors, to name only a few. It is reasonable to assume that those proteins should have a higher score than precursors which contain only one peptide. The example of the prothyroliberin precursor which yields several identical peptides (pyro)Glu-His-Pro-NH_2 (cf. appendix C, "A survey of common vertebrate peptide hormones") is eloquent. By only looking at how well the precursor primary structure matches that of the HMM architecture described in Figure 4, one misses the information that

precursors potentially contains several peptides, and are as such more likely to be PH precursors. I introduce here an algorithm that allows for the computation of the number of cleavages expectation in a sequence.

ComputeExpNumCleav algorithm

In order to compute this score, we consider the modified alpha variable:

$$\begin{aligned}&\alpha_t^n(j) = P(O_1, O_2, ..., O_t, q_t = S_j, numcleav = n/\lambda), \quad 0 \le n \prec n_{\max}.\\&\alpha_t^{n_{\max}}(j) = P(O_1, O_2, ..., O_t, q_t = S_j, numcleav \ge n_{\max}/\lambda).\end{aligned} \tag{0.61}$$

where variable *numcleav* denotes the number of cleavages occurring before t (t included), and $n_{\max}$ is the maximal number of cleavages, set beforehand. In practice this number will be set to 8, so as not to make this algorithm too computationally expensive. PH that harbour more than 8 cleavages are rare, and this number should allow for a sufficiently accurate approximation of number of cleavages expectation. Let us define the set of indexes K corresponding to cleavage states:

$$\mathrm{K} = \{ j \in [1, N] / S_j = CLEAVAGE\ \ STATE \}$$

1) Initialisation:

$$\begin{cases}\alpha_1^0(j) = \pi_j b_j(O_1), \quad \forall\, j.\\ \alpha_1^n(j) = 0, \quad (\forall\, n \succ 1)\ \ or\ \ (n = 1,\ j \in K).\\ \alpha_1^1(j) = \pi_j b_j(O_1), \quad for\ \ j \in K.\end{cases} \tag{0.62}$$

2) Induction:

for $j \in K$:

$$\begin{cases}\alpha_t^n(j) = \sum_{k=1}^{N} a_{kj}\alpha_{t-1}^{n-1}(k) b_j(O_t), \quad 0 \prec n \prec n_{\max},\ \ t \succ 1.\\ \alpha_t^{n_{\max}}(j) = \sum_{k=1}^{N} a_{kj}\left[\alpha_{t-1}^{n_{\max}-1}(k) + \alpha_{t-1}^{n_{\max}}(k)\right] b_j(O_t).\end{cases} . \tag{0.63}$$

for $j \notin K$:

$$\alpha_t^n(j) = \sum_{k=1}^{N} a_{kj}\alpha_{t-1}^n(k)b_j(O_t), \qquad 0 \le n \le n_{\max}, \quad t \succ 1. \tag{0.64}$$

Note that for $t \in [1,T]$, the calculation of the $\alpha_t^n(j)$ for $j \in [1,N]$ and $n \in [0,n_{\max}]$ can be done in any order. Loops on *n* and *j* are independent from each other.

3) Termination:

We have:

$$P(O, numcleav = n/\lambda) = \alpha_T^n = \sum_{j=1}^{N} \alpha_T^n(j)\omega_j \,. \tag{0.65}$$

The expectation of the number of cleavages is then approximated by:

$$Score_{numcleav} = \mathop{\mathrm{E}}_{n}\left[\alpha_T^n\right] \approx \sum_{n=0}^{n_{\max}} n\alpha_T^n \,. \tag{0.66}$$

The time complexity of this algorithm is $n_{\max}$ times larger than the standard forward algorithm. If $n_{\max}$ is large, the approximation of the expectation of the number of cleavages will be better, but it will also be more costly in time.

We cannot calculate directly the numbers $\alpha_T^n(j)$ because for real-world values of T (around 300 on average) those are too small for computers to deal with. We can use a trick analogous to the scaling trick used to implement the forward algorithm (Rabiner 1989). At every step we divide all $\alpha_t^n(j)$ by C_t:

$$\begin{cases} C_t = \sum_{n=0}^{n_{\max}} \sum_{j=1}^{N} \alpha_t^n(j), \quad t \in [1,T-1]. \\ C_T = \sum_{n=0}^{n_{\max}} \left[\sum_{j=1}^{N} \alpha_T^n(j)\omega_j \right]. \end{cases} \tag{0.67}$$

Note that C_t only depends on the position t in the sequence, but not on j or n, and that equations (0.62), (0.63) and (0.64) are linear in the $\alpha_{t-1}^n(k)$.

The trick is to compute some $\hat{\alpha}_t^n$ using the same algorithm as for the α_t^n, with the only difference being that the $\hat{\alpha}_t^n$ are normalised by $\frac{1}{C_t}$:

Initialisation:

$$\begin{cases} \hat{\alpha}_1^n(j) = 0, \quad (\forall n \succ 1) \ \ or \ \ (n = 1, \ j \in K). \\ \hat{\alpha}_1^1(j) = \frac{1}{C_1}\pi_j b_j(O_1), \quad for \ \ j \in K. \end{cases} \tag{0.68}$$

Induction:

for $j \in K$:

$$\begin{cases} \hat{\alpha}_t^0(j) = 0, \quad n = 0, \quad t \succ 1. \\ \hat{\alpha}_t^n(j) = \frac{1}{C_t}\sum_{k=1}^{N} a_{kj}\hat{\alpha}_{t-1}^{n-1}(k) b_j(O_t), \quad 0 \prec n \prec n_{\max}, \ \ t \succ 1. \\ \hat{\alpha}_t^{n_{\max}}(j) = \frac{1}{C_t}\sum_{k=1}^{N} a_{kj}\left[\hat{\alpha}_{t-1}^{n_{\max}-1}(k) + \hat{\alpha}_{t-1}^{n_{\max}}(k)\right] b_j(O_t). \end{cases} \tag{0.69}$$

for $j \notin K$:

$$\hat{\alpha}_t^n = \frac{1}{C_t}\sum_{k=1}^{N} a_{kj}\hat{\alpha}_{t-1}^{n}(k) b_j(O_t), \qquad 0 \le n \le n_{\max}, \ \ t \succ 1 \ , \tag{0.70}$$

where the C_t were computed according to equations (0.67).

From equations (0.62), (0.63), (0.64) and (0.68), (0.69) and (0.70), we deduce the link between $\alpha_t^n(j)$ and $\hat{\alpha}_t^n(j)$:

$$\forall t, n, j \quad \alpha_t^n(j) = \hat{\alpha}_t^n(j)\left[\prod_{r=1}^{t} C_r\right]. \tag{0.71}$$

By definition we have: $P(O/\lambda) = \sum_{n=0}^{n_{\max}}\left[\sum_{j=1}^{N}\alpha_T^n(j)\omega_j\right] = \left[\prod_{r=1}^{T} C_r\right]\sum_{n=0}^{n_{\max}}\left[\sum_{j=1}^{N}\hat{\alpha}_T^n(j)\omega_j\right]$.

And, since $\sum_{n=0}^{n_{\max}}\left[\sum_{j=1}^{N}\hat{\alpha}_T^n(j)\omega_j\right]=1$:

$$P(O/\lambda)=\prod_{r=1}^{T}C_r. \qquad (0.72)$$

From (0.65), (0.71) and (0.72):

$$\sum_{j=0}^{N}\alpha_T^n(j).\omega_j = P(O,numcleav=n/\lambda)=P(O/\lambda)\sum_{j=0}^{N}\hat{\alpha}_T^n(j)\omega_j.$$

We can also express $P(O,numcleav=n/\lambda)$ as:

$$P(O,numcleav=n/\lambda)=P(O/\lambda)\,P(numcleav=n/O,\lambda).$$

By identification we obtain:

$$P(numcleav=n/O,\lambda)=\sum_{j=0}^{N}\hat{\alpha}_T^n(j)\omega_j. \qquad (0.73)$$

This last variable can be used to derive a score $score_{numcleav}$ measuring the chances that there are many peptides in a given precursor protein:

$$score_{numcleav}=\log\left[P(numcleav=n/O,\lambda)\right]. \qquad (0.74)$$

However, the score $score_{cleavage\,density}=score_{numcleav}/length$ has been found to be much more powerful in discriminating between PH precursors and non-PH secreted proteins.

2.7.4 Correlation between scores

Correlations of the scores presented in the previous section were computed on a set of 300 high scoring sequences, in order to see if it made sense to take a combination of the three scores (cf. next subsection 2.7.5).

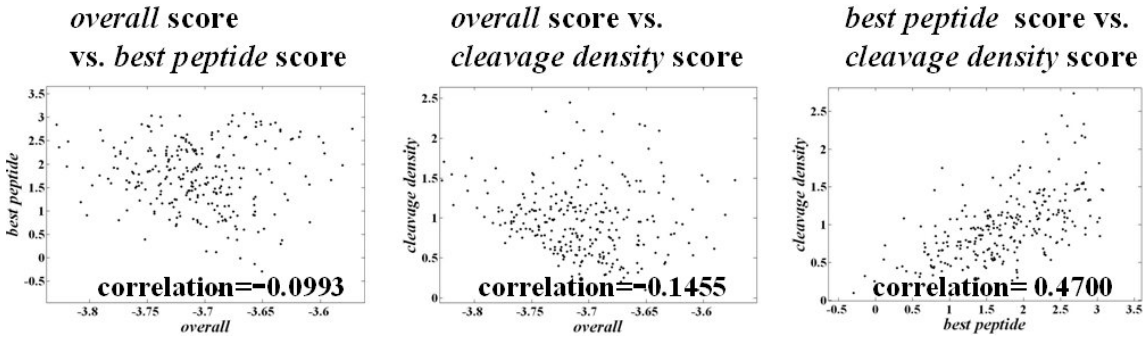

Figure 25: correlations of three scoring systems, two by two.
Correlations of *overall* and *best peptide* scores, *overall* vs. *best peptide* scores, and *best peptide* vs. *cleavage density* scores on the set composed by the 300 highest scoring sequences by CSA-HMM algorithms. Note that best peptide and cleavage density are highly correlated, probably due to their priming of high cleavage site probabilities in sequences. The other two correaltaions are relatively low suggesting that a linear combination of those discriminant variables will improve prediction performance.

Figure 25 shows that on that set, *overall* scores are not highly correlated with the other two scores, *best peptide* and *cleavage density* suggesting that (given that those scores are discriminant in the first place) we should improve on the prediction performance by forming a linear combination of those scores.

2.7.5 The *mixed* score

Since all three scores are to a certain extent uncorrelated it is reasonable to think that a linear combination of those scores will increase will produce as score with increased predictive power.

$$Score_{mixed} = \alpha Score_{best\ peptide} + \beta Score_{cleavage\ density} + \gamma Score_{overall}\ . \qquad (0.75)$$

α, β and γ were chosen so that all numbers added were in the same range. Values for α, β and γ were set respectively to 0.25, 100 and 10. No real optimization was made on those parameters to increase the predictive power of this *mixed* score and there is certainly room for improvement in this area.

2.8 Validation and evaluation of models

This section presents the results obtained on the evaluation of the models presented in sections 2.4, 2.5 and 2.6, by cross-validation on the set of secreted proteins.

2.8.1 Training and testing sets

It is known that we can predict rather well signal peptides using HMM (Bendtsen et al. 2004). Thus, the global PHMM will likely perform well if it is possible to detect PH precursors among secreted proteins.

For this reason, a cross-validation experiment was set up on a set of 1355 alignments of secreted proteins in which 90% of randomly chosen PH precursors and 90% of randomly chosen secreted non PH precursors were selected to learn the HMM and the remaining 10 % of PH precursors and secreted non PH precursors were set aside for testing purposes. Alignments of human, mouse, rat, cow, dog, opossum tetraodon, fugu and zebrafish secreted protein sequences were generated using Ensembl annotations and the ClustalW multiple alignment program.

Three algorithms (standard HMM on human sequences, A-HMM and CSA-HMM on alignments), two types of decoding procedures (Viterbi vs posterior decoding), and four scoring systems (*overall*, *best peptide*, *cleavage density* and *mixed* scores) were compared.

A database of known PH was created, including known cleavage sites positions and precursor sequences. A python script generated model files and testing sequences files for 50X cross-validation of the HMM. All evaluation scores were averaged through the 50 testing procedures. Architecture of the model was set as described in Figure 4 and only transition and observation matrices were estimated from the training set. More precisely, for each training set, the 5 PC cleavage sites states (corresponding to positions -4 to+1 relative to the cleavage site) aa frequencies were estimated. Annotations pertaining to the lengths of *peptide*, *propeptide* and *chain* states were translated into transition probabilities by the formula: $p_{stay} = \frac{avg(lengths)}{avg(lengths)+1}$ (cf. 2.4.3.1).

Model parameters derived from training set sequences, including (whenever applicable) state conservation score distributions, were stored in files of XML format. This format allowed me to change many times the models structure and variables with only minor changes on the document type definition (DTD), and benefit from efficient document parsers. Predictions were also generated as XML files, and standard XML parsers could be used to retrieve the data. To avoid unnecessarily overloading of the cross-validation script, the signal peptide model was not re-learned every time, and was left unchanged from the beginning on for each

of the 50 training sets. Parameters of the signal peptide HMM were estimated by the Baum-Welch training procedure (cf. subsection 2.4.7).

2.8.2 Results on the validation of models

2.8.2.1 Receiver Operating Characteristics (ROC) curves

For any given classifier, there exist a trade-off between the sensitivity and specificity it can achieve. For a given true positive rate (sensitivity or recall), a ROC curve maps the corresponding false positive rate (1-specificity). The area under a ROC curve, called the ROC score can be used to synthesise the overall performance of classifiers. The greater this number, the more sensitive it is for a given level of errors (false positive rate). A ROC score of 0.5 means that the classifier performs as good as random. A ROC score of 1 is achieved only by perfect classifiers.

Figure 26 shows that the best scoring system is the *mixed* score (red line) and the best performance is obtained for the CS-HMM. It also shows that taking into account information from alignments improves predictive power of the HMM. Averaging the contribution of orthologous aa aligned to those of a human sequence substantially improves performance of the HMM, as it is done by the naive HMM on alignments (blue line). In addition, the incorporation of a conservation score along the human sequence in the HMM further improves prediction power, as is implemented by the conservation-score HMM method (red curve).

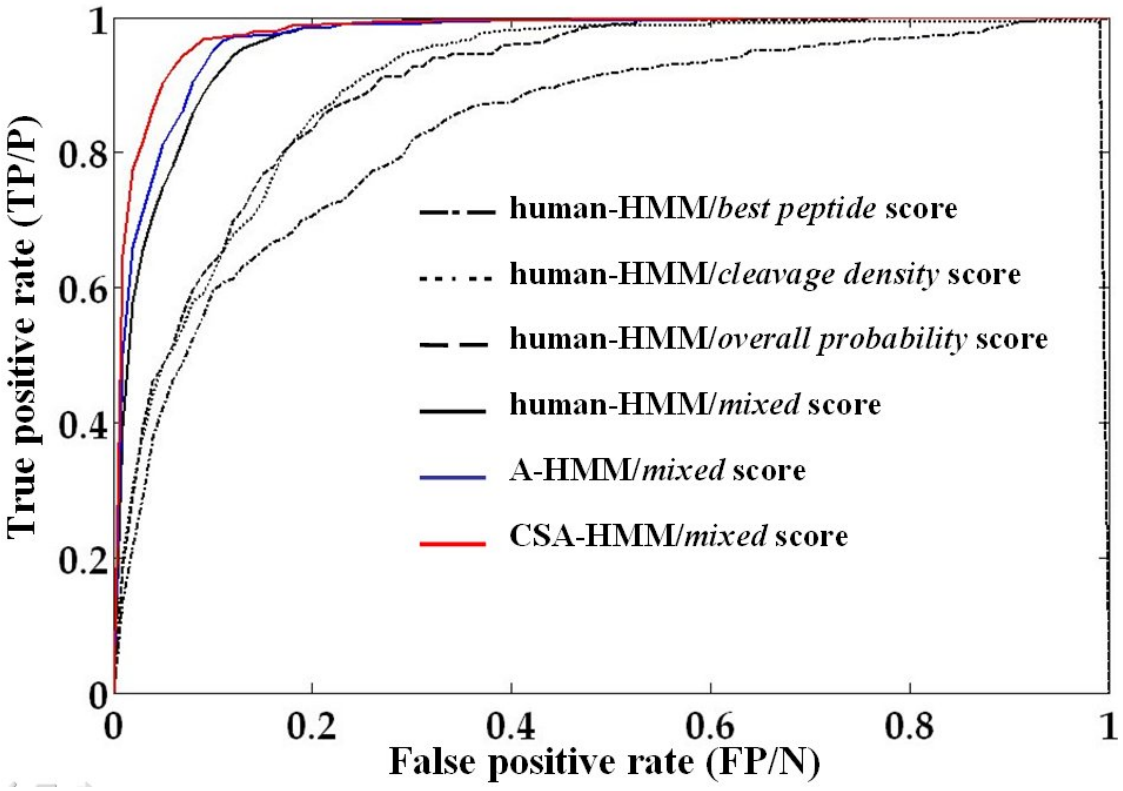

Figure 26: Receiver Operating Characteristics (ROC curves) for different HMM and scoring systems

The x axis defines the percentage of false positives found (FP) on the overall number of negative examples (N) (also called the false positive rate, equal to 1-specificity) while the y axis defines the fraction of true positives found (TP) on the overall number of positive examples (P) (the true positive rate). Black lines are ROC curves for four scoring systems, the *best peptide* score (cf. 2.7.2), the *cleavage density* score (cf. 2.7.3), the *overall probability* score (cf. subsection 2.7.1) and the *mixed* score (cf. subsection 2.7.4) for a standard HMM that was run only on human protein sequences (human-HMM). The blue line is the ROC curve corresponding to the *mixed* score for the alignment-HMM (A-HMM, cf. section 2.5). The red line is the ROC curve corresponding to the *mixed* score for the conservation score-based HMM on alignments (CSA-HMM, cf. section 2.6)

2.8.2.2 Table of ROC values for the different models/scoring systems:

The ROC numbers of all classifiers tested are relatively close, however the difference in terms of performance can be important in the region of low false positive rate. This cross-validation data shows that for all of three types of HMM (standard, A-HMM, CSA-HMM) the *mixed* score is the most discriminating score. It also shows that, whatever the score, CSA-HMM performs better than A-HMM, which itself performs better than the standard HMM.

	H-HMM	A-HMM	CSA-HMM
overall	0.9081	0.9284	0.9374
cleavage density	0.8983	0.8994	0.9169
best peptide	0.8369	0.8880	0.9025
mixed	0.9594	0.9657	0.9743

Table 1: ROC scores
associated to four scoring systems (*overall*, *best peptide*, *cleavage density* and *mixed* scores) and the three types of HMM procedures, standard HMM on human proteins ("H-HMM"), and the two alignment-based HMM (A-HMM and CSA-HMM).

2.8.2.3 Accuracy of cleavage site predictions

The HMM ranks well PH, but how accurately can it label PC cleavage sites? The problem of PC cleavage sites prediction is a challenging problem. The signal around the cleavage site seems not to be strong, and to date no rule has been derived to this point, other than the well-known and poorly-discriminating rule K/R-X_n-K/R-K/R where n is an odd number (Steiner 1998). This rule is not powerful enough to use it systematically for a proteome screen, because too many candidates would qualify as containing one of those sites. One needs to find other criteriae for identifying candidate PH. To my knowledge, only a few studies have addressed this problem, and they reported poor accuracy of general PC cleavage sites by classical machine learning techniques such as neural networks (Duckert et al. 2004). The reason for that may be that too little data is available on cleavage sites by this class of proteases. Furthermore, few studies have shown that certain dibasic sites cannot be cleaved by any PC/Furin cleavage sites. As is often the case in biochemistry studies this sort of negative data that the bioinformatician would need to derive sound machine learning predictors is often missing.

I contrasted the prediction performance of the cleavage sites predictions induced by the HMM for two different decoders with other commonly used descriptors of PC cleavage sites (such as dibasic motifs K/R-K/R). The performance of PC/Furin cleavage sites classifiers were assessed using specificity and sensitivity measures. The specificity of a classifier is defined by the probability that it produces a true negative result when it is indeed negative.

$$specificity = \frac{TN}{TN + FP}, \tag{0.76}$$

where TN is the number of true negatives in the test, and FP the number of false positives. In our case, a high specificity means that when the HMM labels a cleavage site, it is likely to be a real one. In contrast, the sensitivity of the PC/Furin cleavage site classifier is defined as the likelihood that a given known cleavage site is identified as such.

$$sensitivity = \frac{TP}{TP + FN}, \tag{0.77}$$

where TP is the number of true positives in the test, and FN the number of false negatives. Figure 27 summarizes some of those definitions.

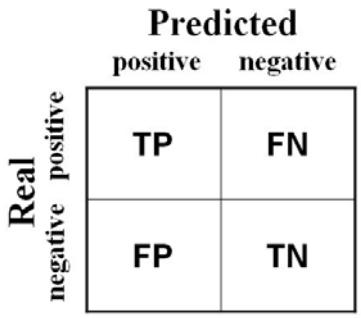

Figure 27: diagram of predicted call vs. real labels
Diagram showing the definitions of true positive (TP), false negative (FN), false positive (FP), true negative (TN) classifications depending on the prediction call (positive or negative) and the real label (positive or negative).

Accuracy takes into account both specificity and sensitivity:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{0.78}$$

Table 2 and Table 3 show that the HMM provide better predictions than any other of the classical PC/Furin cleavage site models, including the dibasic motif K/R-K/R. The best "intuitive" classifiers K/R-R and K/R-K/R have a sensitivity which is comparable to that of the posterior-Viterbi (PV) decoder, for all HMM types (about 62-65%).

classifier	K/R-K/R	K/R	K/R-X-X-R	K-R	(K/R)R	K/R-X-K/R-R
sensitivity	0.67929	0.81839	0.05564	0.39181	0.60046	0.14297
specificity	0.44461	0.10992	0.23841	0.73478	0.57684	0.59871

Table 2: Sensitivity and specificity of different PC/Furin cleavage sites predictors:
Performances of common descriptors of PC/Furin cleavage sites, ordered by increasing specificity.

However, they are much less specific (44% for K/R-K/R and 57% for K/R-R) than the HMM PV decoders (specificity of more than 96%). Table 3 also shows that, generally the CSA-HMM predicts CS more accurately than the other HMM. There is one exception to that, the Viterbi specificity is lower for CSA-HMM than the other two. This is certainly due to alignment errors which lead the more sensitive CSA-HMM Viterbi decoder to wrongly attribute CS (typically in regions corresponding to splicing sites).

decoding	Viterbi			Posterior-Viterbi		
type of HMM	H-HMM	A-HMM	CSA-	H-HMM	A-HMM	CSA-
sensitivity/recall	0.52859	0.52318	0.55719	0.62442	0.62751	0.65920
specificity	0.98357	0.98993	0.97518	0.96451	0.97674	0.97638
accuracy	0.92246	0.92724	0.91904	0.91883	0.92983	0.93378

Table 3: comparative performances of the Viterbi and the posterior-Viterbi decoders
Sensitivity, specificity and accuracy of the three types of HMM, standard HMM (H-HMM), alignment-based HMM (A-HMM) and conservation scores and alignment-based HMM (CSA-HMM)

Comparison of the two decoders does not lead to a clear advantage of one over the other. As expected (cf. 2.3.3.3), the PV decoder is more sensitive and detects more cleavage sites than Viterbi does (62-65% sensitivity for the PV decoder vs. 52%-55% for the Viterbi decoder). However, the PV decoder also makes more wrong calls and this translates into a lower specificity compared to the Viterbi decoder. The accuracy, as a sort of synthesis between these two measures gives comparable values for both decoders.

Note that for the purpose of calculating specificities I needed to define a set of possible cleavage sites. I restricted myself to sites where the position -1 was a basic residue.

2.8.3 Conclusions on the evaluation of models

Firstly, 50X cross-validation ROC curves in Figure 26, along with the high ROC scores shown in Table 1 clearly demonstrate that the HMM can distinguish PH precursor sequences among secreted protein sequences.

Secondly, Figure 26 shows that the mixed score is the best scoring system and that the most natural score, the *overall* score, is not sufficient to discriminate well PH precursors vs. non PH secreted protein sequences.

Thirdly, inclusion of orthologous sequences improves prediction quality both in terms of PH precursor scoring and accuracy of CS prediction. Incorporation of a conservation score in the HMM (2.6.1.3) further helps the algorithms to distinguish PH precursors from non-PH precursors (ROC curves of Figure 26).

Lastly, the comparison for our problem of the two decoders, Viterbi and posterior-Viterbi (cf. 2.3.3.2 and 2.3.3.3), was inconclusive in terms of overall performance.

2.9 Screening the proteome with PHMM

In this section, I summarize the results I obtained when I ran the CSA-HMM version of the PH HMM on a large set of aligned protein sequences.

2.9.1 Building the protein alignments

The first thing that was done was to create for each human protein a file containing orthologous sequences from the following vertebrate organisms: *Mus musculus* (Mouse), *Monodelphis domestica* (opossum), *Gallus gallus* (chicken), *Xenopus laevis* (frog), *Danio rerio* (zebrafish), *Gasterosteus aculeatus* (stickleback fish), and *Takifugu rubripes* (puffer fish). For each of those files multiple alignments were done using the ClustalW software with settings –GAPOPEN=5 and default settings otherwise.

Briefly, here is how I collected those orthologous sequences: for each organism I gathered through the Ensembl biomart interface (http://www.ensembl.org/biomart/martview) the orthology table linking all human protein identifiers (ids) with orthologous sequences ids from the organism in question. Those tables were then inserted into a local MySQL database. Then for each organism (including human) tables containing protein ids and their corresponding sequences were inserted into MySQL tables. A large table was then created (using sql join commands on all the previously mentioned tables) that contained for each human protein the reference and corresponding sequence its orthologous proteins. Orthology relationship in Ensembl is defined at the level of the gene. In order to avoid combinatorial explosion of files containing spliced variant proteins I had to restrict myself, for each organism and each human protein, to picking at most one orthologous protein.

A script was then written in java which created, for each human protein, one file containing its orthologous sequences. After the execution of ClustalW on those sequence files, a total of 23344 alignment files were made available as input for the CSA-HMM algorithm.

2.9.2 Screening PHMM architecture

The architecture of the HMM largely follows that of the PHMM in the previous section (cf. section 2.8 and Figure 28). A mitochondrial HMM (composed of four states, one cyclic state plus three corresponding to cleavage of the transit peptide) was added to discard mitochondrial proteins. A *null* state designed to model the average protein in terms of length

and amino acid composition (cf. Figure 19) was added to the global HMM. In addition to that, a possibility to go from *START* to *peptide1* and *propeptide2* was allowed to discard those sequences which did not contain predicted signal peptides (after running the decoding algorithm). The HMM presented here does not contain a transmembrane state because the transmembrane-predicting HMM had not been optimized (as it had been optimized for SP length modelling). I chose nonetheless to run the HMM without it, expecting to find more false positive due to it, but at least being certain not to discard potentially interesting proteins.

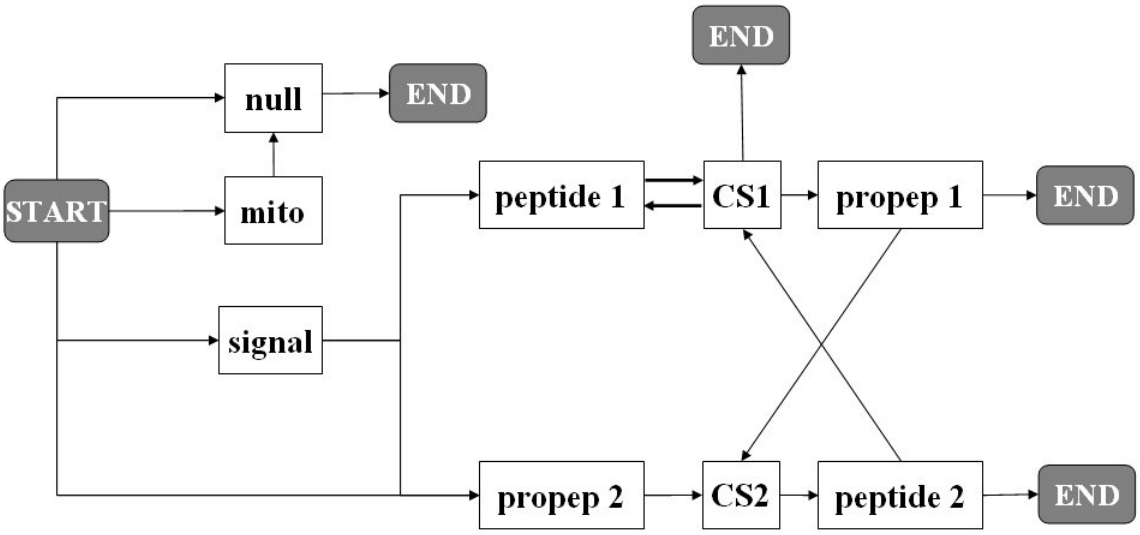

Figure 28: PHMM architecture for screening proteome databases
mito, mitochondrial transit peptide HMM; *signal*, signal peptide HMM.

2.9.3 Results

The results obtained with the architecture of Figure 28 and the CSA-HMM algorithms compare favourably to those initially obtained when I first screened the databases in search of novel PH (in March 2004, cf. paper at the end of the appendix section). In particular, nearly 75% of the top 100 proteins were found to be peptide hormone precursors. Furthermore, I found spexin (ranked number 12 as opposed to 41) and augurin (ranked number 76) higher in the list than previously.

The composition of the top 200 proteins shown in Figure 29 reveals that more than 50% of those highly scoring proteins are PH precursors and that the vast majority of the remaining 50% of characterized proteins are either growth factors/ cytokines (that are most of the time

secreted and processed signalling molecules, cf. 1.6.1.7 and 1.6.1.8) or other secreted proteins. This result demonstrates the power of the CSA-HMM algorithms at singling out PH, and secreted and processed signalling proteins. The majority of proteins which were wrongly predicted as PH were membrane proteins (about 10%, Figure 29). Surely, most of them could have been disqualified had the transmembrane HMM been included in the search algorithm. However, their presence could be meaningful, as some membrane proteins are known to be processed into mature soluble ligands, including tumor necrosis factor ligand superfamily 12 (TNF12) (found in the list with ranking 212, cf. appendix C) and fractalkine (found in the list with a ranking of 274).

2.9.3.1 Composition of top 200 candidates

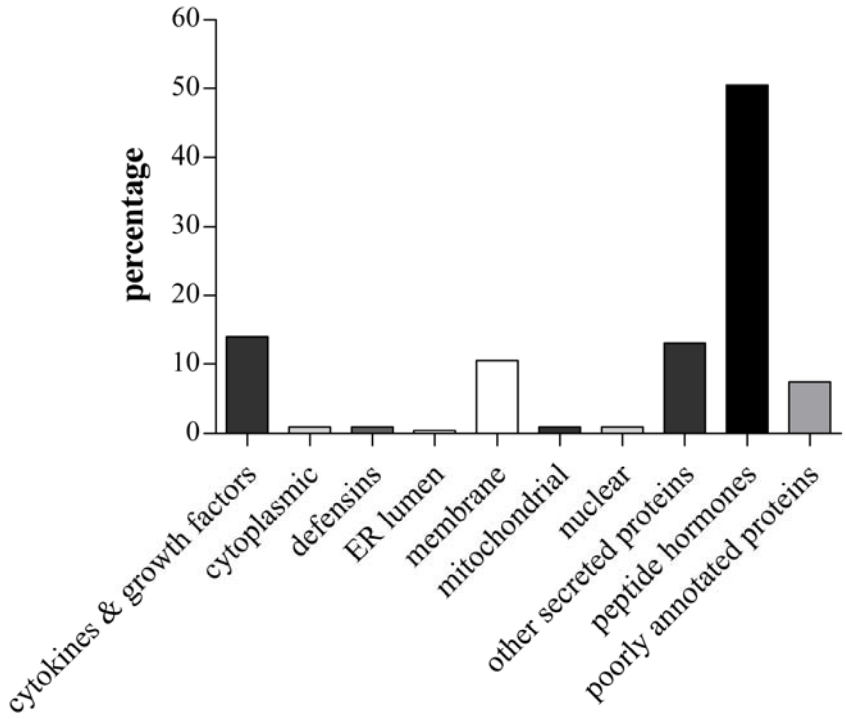

Figure 29: Composition of top 200 proteins
The top 200 proteins, ranked by the *mixed* score of the CSA-HMM, were annotated according to the following categories: cytokines and growth factors cytoplasmic, defensins, ER lumen, membrane, mitochondrial, nuclear, peptide hormones, poorly annotated proteins and other secreted proteins. Note that the great majority of proteins in this top 200 set received the "peptide hormone" annotation (there are more about 4 times many more than any other category). The composition is given in percentage of the total number of proteins (=200)

2.9.3.2 Candidate PH that were studied

In the next chapters I present data on some of the candidates that were on candidate PH (CPH) lists generated during the early years of my PhD (March 2004-January 2005). Most of them are not present in this "final" list (see appendix A) which has been generated using the algorithms that I have presented in 2.4, 2.5, 2.6 and 2.7. Spexin (ENSP00000256969 for Ensembl, Q9BT56 for Swiss-Prot/TrEMBL) was immediately singled out as a very promising candidate due to its striking alignment features (cf. Chapter 2: Characterization of spexin). Augurin as a CPH was found nearly one year later, after I had made some adjustments on the HMM and came back to screening protein sequence lists. I originally detected augurin after I had manually screened candidates "far" from the top of the candidate list (its best rank was 276) and nearly a full year later after I started studying spexin and other candidates I

individuated augurin as a very promising candidate. After looking at full alignments of augurin I convinced myself that the augurin protein would surely contain peptide hormones/growth factors and I immediately started working on it. I present the data I have on augurin in the third chapter. In this thesis I also present data about proteins uncharacterized at the time of the initial search that were chosen from the list because their sequence showed interesting features (cf. Chapter 4: Preliminary characterization of four other candidates)

2.9.3.3 CPH52, a novel candidate

I have very recently come across one other protein (Ensembl peptide ID:ENSP00000367952, named CPH52) that was ranked number 90 in the list of CPH (cf. appendix A) and which was not found at the time of the initial screen among the top 500 proteins. I had originally found this candidate when I screened the fish sequences for possible fish-specific PH. In the latest screen, this protein ranks second highest among proteins which were labelled by Swiss-Prot as uncharacterized proteins (augurin is not in that category but spexin is) and looks very promising as one can judge from the multiple alignment of Figure 30.

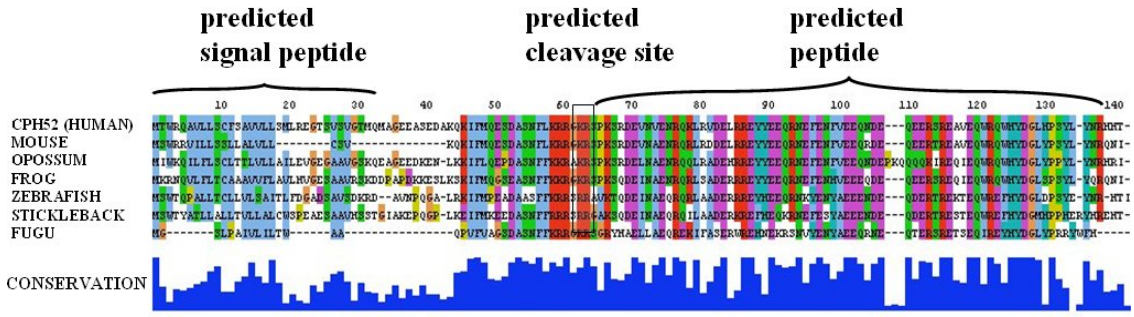

Figure 30: multiple alignment of CPH52 orthologous sequences

Note the predicted features identifiable with aa colour codes: the N-terminal signal peptide (core hydrophobic part in light blue), the perfectly conserved cleavage site (region in red) and the highly conserved region highly charged (negatively charged "purple aa" glutamate and aspartate) peptide region at the c-terminus end of the protein a drop of conservation in the propeptide region (conservation graph).

The alignment was visualised using the Jalview software (cf. Resources)

2.10 Feature prediction visualisation tool

In order to explore best features and scores associated to the CPH, a visualising tool was developed in java. This tool takes as input a prediction file in XML format and generates a table of relevant information about sequences and predictions. Each line of the table refers to a given sequence of the prediction file. The columns contain the following information: sequence id (Ensembl protein ID, or Swiss-Prot accession number), description of the protein, primary sequence, different scores, and boundaries of the predicted most likely peptide (cf. Figure 31A). This table allows for the ranking of proteins according to any of the scores that was stored, and for a visualisation of predictions on any sequence. A diagram displaying the predicted PH features appears upon clicking on a candidate from the list. (cf. Figure 31B)

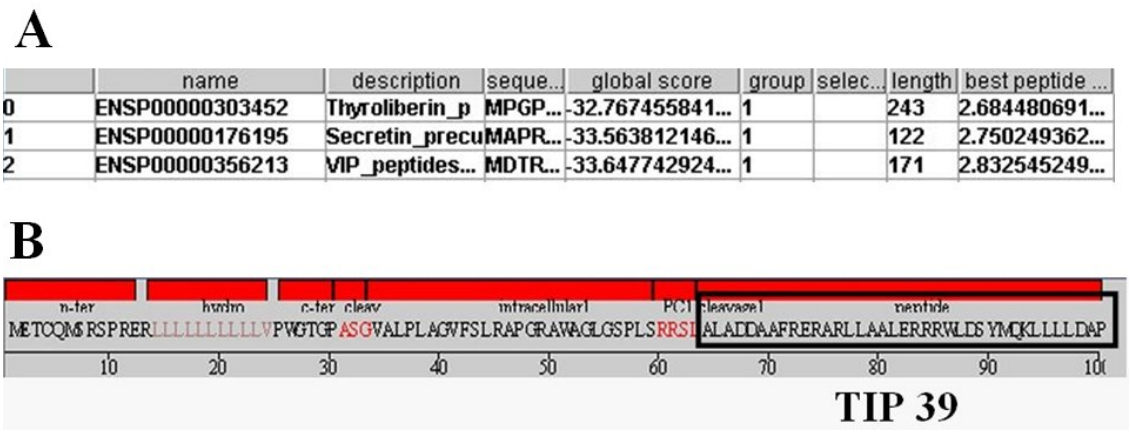

Figure 31: Java-based visualisation tool

(B) This tool was developed using biojava (http://bopjava.org) objects and was part of java-based table interface (A) that allowed exploration of candidate lists.

In (A), information on the three highest ranked protein sequences (thyroliberin, secretin and vasoactive intestinal peptide precursors) is displayed. In (B) predicted features of PH tuberoinfundibular protein 39 (TIP39) are displayed.

2.11 Additional useful HMM algorithms

In this section I introduce two modified versions of standard HMM algorithms (Viterbi and forward-backward) that were not essential to arrive at the most important results of my work, but which could be of some general interest for biological sequence analysis problems.

2.11.1 Accurate lengths modelling

I present a modified version of the Viterbi algorithm to better model feature length distributions. The idea of this algorithm is to keep track of the current length duration for each state, and use the approximate conditional distribution to correct for any current length. In other words, at each transition, instead of multiplying by a constant probability determined by the transition matrix, we multiply by a probability which depends on the current duration of the state. State transition probabilities are not fixed like in the Viterbi procedure but depend now on the current state durations. Note that this approximate inhomogeneous HMM procedure can only be applied to solve problem 2, i.e. the decoding problem, and not to assign a score to proteins. The implementation is as efficient (we just have to add one variable) as the original Viterbi algorithm and we can reasonably hope to gain in length modelling accuracy.

Let us remind the definition of the delta variable:

$$\delta_t(i) = \max_{q_1,q_2,...,q_{t-1}} P(q_1, q_2, ..., q_{t-1}, q_t = S_i, O_1 O_2 ... O_t / \lambda). \quad (0.79)$$

Let $d_t(j)$ be the number of consecutive times the process has stayed at state j to get to position t. $d_t(j)$ is referred to as the Viterbi best current duration of state *j* at *t*.

$$d_t(j) = \max_d (q_{t-d+1} = S_j, q_{t-d+2} = S_j, ..., q_t = S_j / \lambda). \quad (0.80)$$

Let F be the estimated empirical cumulative distribution of state lengths.

$$F_i(l) = P(length(S_i) \le l). \quad (0.81)$$

where $length(S_i)$ is the random variable describing the possible state duration lengths of state S_i. F_i can be constructed empirically from biological databases: Let us call $\{sequence\}_i$ the

set of sequences we wish to model (with a single state HMM). For each *n* and *i* the values $F_i^{approx}(l) = \frac{|\, seq \in \{sequence\}_i / length(seq) < l \,|}{|\, \{sequence\}_i \,|}$ are computed.

If the training set is not large enough, such estimation will lead to some of the terms $F_i^{approx}(l)$ being equal to zero. To overcome this problem, a smoothening procedure was applied where a Gaussian function was convoluted with F_i^{approx}: $F_i = F_i^{approx} * H(\alpha, \sigma)$. To lighten expressions let us define the random variable $X_i = length(S_i)$.

We can define pseudo-transition probabilities $\tilde{a}_{ij}(l)$ that depend on the current state duration as:

If $i \neq j$:

$$\tilde{a}_{ij}(l) = P(q_t = S_j / q_{t-1} = S_i, i \neq j, d_i(t-1) = l) = k_{ij} P(X_i = l / X_i \geq l) = k_{ij} \frac{P(X_i \leq l) - P(X_i \leq l-1)}{P(X_i \geq l)}.$$

$$\Rightarrow \tilde{a}_{ij}(l) = k_{ij} \frac{F_i(l) - F_i(l-1)}{1 - F_i(l-1)}, \quad l \geq 1. \tag{0.82}$$

And if $i = j$:

$$\tilde{a}_{ii}(l) = P(q_t = S_i / q_{t-1} = S_i, d_i(t-1) = l) = kP(X_i \geq l+1 / X_i \geq l) = \frac{P(\{X_i \geq l+1\} \cap \{X_i \geq l\})}{P(X_i \geq l)}.$$

$$\Rightarrow \tilde{a}_{ii}(l) = \frac{1 - F_i(l)}{1 - F_i(l-1)}, \; l \geq 1. \tag{0.83}$$

where k_{ij} is set to $k_{ij} = \frac{a_{ij}}{1 - a_{ii}}$ to translate the fact that the relative transition probability from state *i* to *j* and from state *i* to *j'* is unchanged: $\frac{\tilde{a}_{ij}(l)}{\tilde{a}_{ij'}(l)} = \frac{a_{ij}}{a_{ij'}}$. With this value for k_{ij} we can also verify that $\sum_{j=1}^{N} \tilde{a}_{ij}(l) = 1$.

We can now rewrite the Viterbi decoding algorithm by replacing transition probabilities a_{ij} by $\tilde{a}_{ij}(l)$ while keeping track of the variable $d_t(j)$.

Inhomogeneous Viterbi algorithm

1) Initialization:

$$\begin{cases} \delta_1^{IH}(i) = \pi(i) b_i(O_1) \\ \psi_t^{IH}(i) = 0 \\ d_1(i) = 1 \end{cases}, \quad 1 \le i \le n.$$

2) Recursion:

$$\begin{cases} \delta_t^{IH}(j) = \max\limits_{1 \le i \le N} \left[\delta_{t-1}^{IH}(i) \tilde{a}_{ij} \left(d_{t-1}(j) \right) b_j(O_t) \right] \\ \psi_t^{IH}(j) = \arg\max\limits_{1 \le i \le N} \left[\delta_{t-1}^{IH}(i) \tilde{a}_{ij} \left(d_{t-1}(j) \right) \right] \end{cases}, \quad 2 \le t \le T, \quad 1 \le j \ne i \le N.$$

$$\begin{cases} d_t(j) = 1 & if \quad i \ne j. \\ d_t(j) = d_{t-1}(j) + 1 & if \quad i = j. \end{cases}$$

3) Termination:

$$P^* = \max_{1 \le i \le N} \left(\delta_T^{IH}(i) \omega_i \right).$$

$$q_T^* = \arg\max_{1 \le i \le N} \left(\delta_T^{IH}(i) \omega_i \right).$$

4) Path backtracking:

$$q_t^{IH} = \psi_{t+1}^{IH}(q_{t+1}^{IH}), \qquad t = T-1, T-2, ..., 1.$$

2.11.2 Constrained forward-backward algorithm

The following algorithm may be used when some states are predefined, for instance in the case where we already know certain PC cleavage sites or SP CS positions and we want to include this information in the HMM.

We assume here a unique constraint at time/position k_0: $q_{t_0} = S_{k_0}$. This algorithm is easily extendable to cases where there are constraints at several distinct time points. Furthermore, this algorithm is still applicable for cases where instead of a fixed constraint we have a distribution of probabilities over the likelihood of states at t_0. Note that the idea used here can be applied to the Viterbi algorithm (or posterior-Viterbi decoder) to translate hard-wired constraints.

Constrained forward-backward algorithm

Initialisation:

$$\alpha_1^{\inf}(j) = \pi_j b_j(O_1), \quad 1 \le j \le N \tag{0.84}$$

We first apply the forward-backward algorithm between 0 and t_0:

Induction step 1:

$$\begin{cases} \alpha_{t+1}^{\inf}(j) = [\sum_{i=1}^{N} \alpha_t^{\inf}(i) a_{ij}] b_j(O_{t+1}), \quad 1 \le j \le N, \quad 1 \le t \le t_0 - 1 \\ \beta_{t_0}^{\inf}(j) = \delta_{k_0}(j), \quad 1 \le j \le N \\ \beta_t^{\inf} = \sum_{k=1}^{N} \beta_{t+1}^{\inf}(k) a_{jk} b_k(O_{t+1}), \quad 1 \le j \le N, \quad 1 \le t \le t_0 - 1 \end{cases} \tag{0.85}$$

Where δ_{k_0} is the Kronecker function of parameter k_0 ($\delta_{k_0}(j) = 1$ if $k_0 = j$ and 0 otherwise). By definition we have: $P(O_1,...,O_{t_0}, q_{t_0} = S_j / \lambda) = \alpha_j(k_0)\beta_j(k_0) = \alpha_j(k_0)\delta_{k_0}(j)$ which verifies the constraints $P(O_1,...,O_{t_0}, q_{t_0} = S_j / \lambda) = 0, \quad if \quad j \ne k_0$. We now apply the forward-backward algorithm between t_0 and T:

Induction step 2:

$$\begin{cases} \alpha_{t_0}^{\sup}(j) = \alpha_{t_0}^{\inf}(j)\delta_{k_0}(j) \quad 1 \le j \le N \\ \alpha_{t+1}^{\sup}(j) = [\sum_{i=1}^{N} \alpha_t^{\sup}(i) a_{ij}] b_j(O_{t+1}), \quad 1 \le j \le N, \quad 1 \le t \le t_0 - 1 \end{cases}$$
$$\begin{cases} \beta_T^{\sup}(j) = \omega_j, \quad 1 \le j \le N \\ \beta_t(j) = \sum_{k=1}^{N} \beta_{t+1}(k).a_{jk}.b_k(O_{t+1}), \quad 1 \le j \le N, \quad t_0 \le t \le T - 1 \end{cases} \tag{0.86}$$

3 Characterization of spexin

In this chapter I present and discuss data that has been obtained on spexin, the first candidate PH.

In the first two sections (3.1) and (3.2) I present what is known about its gene structure and the conserved elements of the protein primary structure that the spexin gene encodes. In sections (3.3) and (3.4), I discuss the data I have on the secretion, processing and subcellular localization of FLAG-tagged spexin proteins in a cell expression system. In section (3.5) data obtained on spexin mRNA expression in mouse tissues is presented. In section (3.6) I show data supporting that the predicted spexin peptide is a biologically active molecule. The purpose of the last two sections (3.8) and (3.9) was to propose various hypotheses for spexin's function in the body and speculate on disease it may be involved in.

3.1 Structure of the spexin gene

Spexin was annotated by Ensembl (cf. Resources) as a gene containing 6 exons in human and mouse, all of which are coding (Figure 32). The first exon encodes only two aa, and its annotation was missing for most of the organisms available in Ensembl, possibly due to the difficulty in sequencing and mapping 5' ends of transcripts. Some of the missing first exons were retrieved manually using those first exons which were available, and blasting those sequences against genomic regions upstream of exon 2 for organisms for which exon 1 was missing.

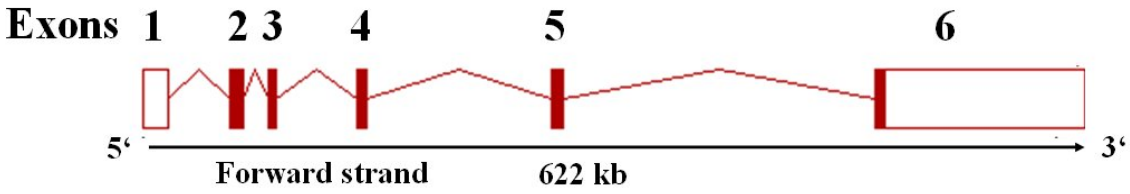

Figure 32: Structure of the human spexin gene. [modified screenshot from Ensembl]
Human spexin has six exons all of which are coding. Exon 1 only codes for two aa residues, and exon 6 is mostly non-coding. Filled red rectangles represent coding-regions of exons, while empty red rectangles define boundaries of non-coding regions of exons. The gene is found in chromosome 12 and spans 622 kb.

No alternative splicing has been reported for any of the 20 vertebrate sequences retrieved from Ensembl including (using Linnæus taxonomy) *Gallus gallus* (noted chicken), lizard, *Bos taurus* (cow), *Myotis lucifugus* (bat), *Loxodonta africana* (elephant), *Canis familiaris* (dog), *Felis catus* (cat), lemur *Otolemur garnettii* (lemur), *Oryctolagus cuniculus (rabbit)*, *Homo*

sapiens (human), *Macaca mulatta* (macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Echinops telfairi* (lesser hedgehog), *Spermophilus tridecemlineatus* (squirrel), *Monodelphis domestica* (opossum), *Takifugu Rubripes* (fugu), *Tetraodon nigroviridis* (tetraodon), *Gasterosteus aculeatus* (stickleback fish) and *Danio Rerio* (zebrafish).

3.2 Primary sequence features of spexin protein

In this section, I present an analysis of primary sequence features (signal peptide, cleavage site, putative peptides) and their conservation using spexin orthologous sequences that were made available in the Ensembl database.

In order to get a good sense of spexin protein primary features conservation, attempts were made at retrieving spexin orthologous sequences from as many organisms as possible using Ensembl annotations. While those annotations were generally accurate, I found that for some sequenced organisms the spexin gene was present in the genome but not annotated as a gene (e.g. macaque, tetraodon, stickleback fish), and for other organisms some exons were either missing (e.g. opossum sequence) or erroneous (for instance, the rat sequence). In order to retrieve the missing sequences/exons, I used the Ensembl BLASTP tool on translated genomic sequences; in combination with genome vs. genome alignments provided by the UCSC genome browser (cf. Resources). The first and last exons (exon 6) are both difficult to identify using standard BLAST methods because they have diverged more than the other 4 exons. In lizard and fishes, the last exon is likely to encode part of the c-terminal spexin propeptide, and has probably diverged too much from mammalian sequences to be identified based solely on homology arguments at the level of the protein. Furthermore, the length of intron 5 is highly variable in mammals (1.2 kb in bat, 5kb in the cow and 2kb in mouse and human), making the search for exon 6 difficult in distant vertebrate organisms. Exon 1 is missing in many sequences but only codes for two aa and is not critical for the study of spexin conservation. Missing aa in sequences where those two exons were not retrieved are replaced with the letter X in Figure 33. No spexin gene was found in the genomes of platypus and frog, and a BLASTP search in those genomes of the translated sequence of human exon 3 (Arg-Pro-Leu-Glu-Arg-Arg-Asn-Trp-Thr-Pro-Gln-Ala-Met-Leu-Tyr-Leu-Lys-Gly-Ala-Gln) did not yield any match.

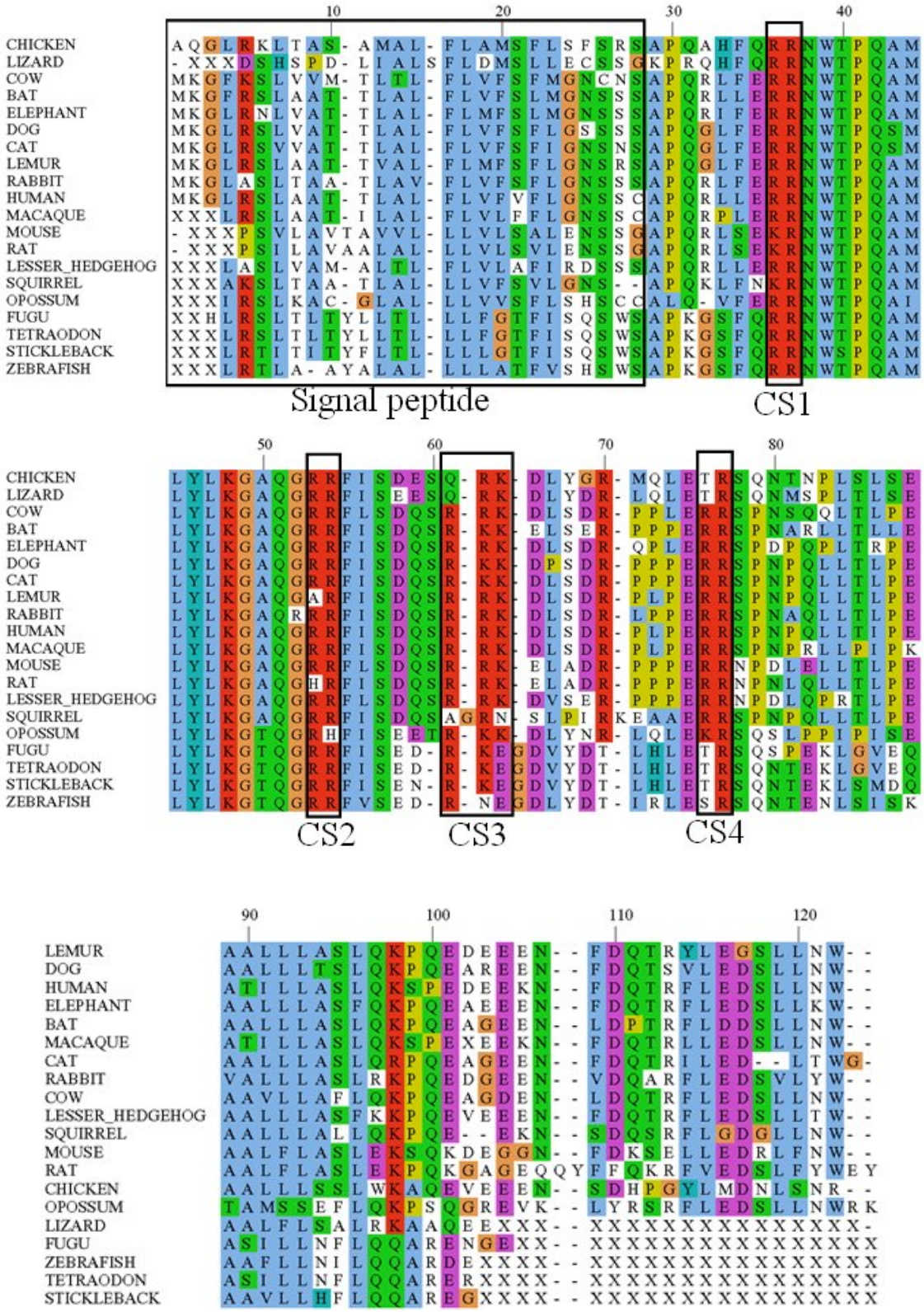

Figure 33: alignment of spexin orthologs.

Spexin orthologs were collected from the Ensembl database, aligned with ClustalW and visualised with the Jalview software. The 4 putative PC/Furin cleavage sites are denoted CS. The putative signal peptide sequence is indicated.

The alignment of Figure 33 shows that spexin contains an obvious signal peptide, and 4 somewhat conserved putative PC/Furin cleavage sites, at pairs of basic residues (denoted CS1-4).

3.2.1 Signal peptide

The hydrophobic region (residues boxed in blue) is about 15 residues long, a number which lies within the range of known signal peptide –h lengths (Nielsen and Krogh 1998). One can note the presence of basic residues (arginines and lysines) before the hydrophobic region, a feature characteristic of –n regions of signal peptides and which contribute to defining the intracellular/extracellular orientation of the peptide (Martoglio and Dobberstein 1998). It is quite straightforward to predict that the putative signal peptide cleavage site is likely to occur just before the conserved Ala-Pro at positions 28-29. Firstly, because it is common that the region before cleavage occurs is less conserved than the region after, since, in most cases, only the latter will be used in living cells. Secondly, because if we postulate cleavage before Pro-28, we see that, for all organisms, the motif before the signal peptide cleavage site matches the expression (Gly/Ser/Cys)-X-(Gly/Ser/Cys)↑ , where the arrow ↑ marks the position of the cleavage, and "X" stands for any aa. And all those residues found at positions -1 and -3 relative to the cleavage site are aa frequently observed in known Eukaryotic signal peptides (cf. logo motif of signal peptides).

3.2.2 PC/Furin cleavage sites

All four dibasic sites highlighted in the alignment show some degree of homology between organisms, making all of them potential PC/Furin cleavage sites. Using both human and mouse sequences I verified that all 4 sites are predicted with a probability of 0.808 to be PC/Furin cleavage sites by the Neuropred algorithm (cf. Resources). However, only residues corresponding to CS1 contain dibasic residues in all organisms. Hence, CS1 is more likely to be used by specific pro-hormone convertases than the other 3 sites; even if we know that perfect conservation is not an absolute requirement for cleavage to occur (e.g. pancreatic hormone and CART are cleaved at non-conserved dibasic sites). Furthermore, only CS1 and CS3 delimitate regions of high and low homology: clearly the region between the end of the putative signal peptide and CS1 (aligned residues 28-34) and c-terminal of CS3 (aligned residues beyond position 63) are poorly conserved, while the region between CS1 and CS3 is a region of high homology (>90%, cf. Figure 33). Those observations suggest that CS1 is the most likely PC/Furin cleavage site, with CS3 being also a likely candidate. On the other hand,

processing at site 2 (CS2) would release in most organisms a peptide ending with a glycine residue. Peptides ending with a glycine residue get amidated (Eipper et al. 1992), and amidation is a common post-translational modification among PH/neuropeptides (Eipper et al. 1992). A the time of my initial HMM search (Spring of 2004) it was therefore tempting to hypothesize that this cleavage was likely to be used, since I found the Gly-Arg-Arg motif (CS2/residues 51-53) present in the sequence of all organisms which were available. Thanks to the recent sequencing and annotation of many more mammalian and fish organisms it is now clear that it if this site (CS2) is used, the amidation is probably not required for the activity of the peptide (cf. alignment Figure 33), since it is not present in all organisms. It also weakens my former hypothesis that this site is used to yield a 15 aa product that we named spexin peptide (Mirabeau et al. 2007). I also noted the presence of a perfectly conserved Asn-Trp-Thr motif (at position 36 of Figure 34 sequence) which stood out as a potential n-glycosylation site, since it followed the canonical n-glycosylation site pattern Asn-X-(Ser/Thr) with X not being a tryptophane residue (Apweiler et al. 1999; Nishikawa and Mizuno 2001).

3.3 *In vitro* secretion and processing of spexin

In this section I present the results of experiments I conducted to test if spexin is secreted and processed by cells in culture.

3.3.1 FLAG-tagging strategy

To study intracellular trafficking of spexin, the eight aa FLAG antigen sequence (Glu-Tyr-Lys-Glu-Glu-Glu-Glu-Lys) was inserted at different locations of the putative precursor.

This FLAG-tagging strategy has been used before to study processing of other PH, including IGF1 (Duguay et al. 1995), CART (Dey et al. 2003), and to purify peptides, such as INSL7 (Liu et al. 2003a). As a control, the FLAG antigen was also inserted just upstream of neuropeptide K (NPK) in the human betapreprotachykinin (TAC1) gene. Previous studies have shown that FLAG sequences are compatible with proteolytic cleavage just N-terminal to the FLAG sequence (Duguay et al. 1995). The FLAG antigen sequence was inserted at three different locations of the spexin protein (cf. Figure 34): just after the predicted signal peptide cleavage site (S-FLAG-spexin), right after the putative PC/Furin cleavage site CS1 (N-FLAG-spexin) and at the c-terminus of spexin (C-FLAG-spexin).

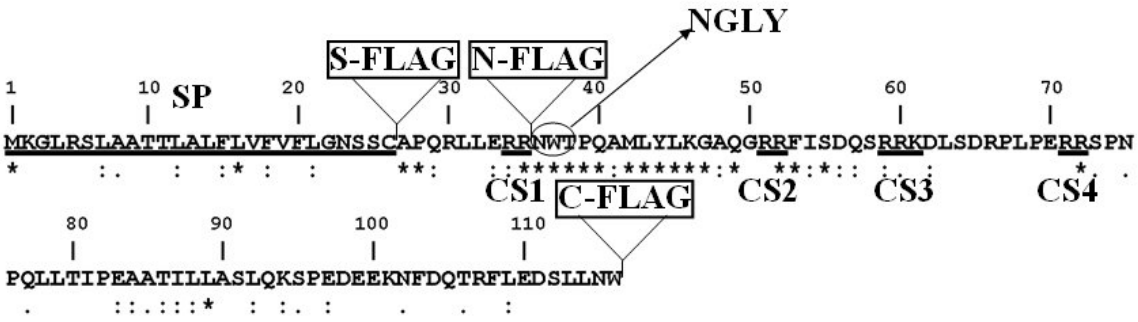

Figure 34: primary structure of human spexin, and description of FLAG constructs.
Conservation among orthologs is show below the spexin primary sequence: (*) identity, (:) high homology; (·) low homology. S-FLAG, N-FLAG, C-FLAG were inserted respectively, after the putative signal peptide, after the first potential cleavage site CS1, and at the c-terminus of the precursor. Predicted features of spexin, including cleavage sites signal peptide (SP), cleavage sites (CS1-4) were underlined, while the putative n-glycosylation site (NGLY) was circled.

Since the exact location of spexin processing site(s) remained elusive, I also made a series of mutants in an effort to locate it/them (see appendix for description of mutants).

3.3.2 Immunoblotting of FLAG-spexin peptides

In a first experiment, β-TC3 cells were transfected with the TKN-FLAG, S-FLAG and N-FLAG constructs, and an immunoprecipitation was performed with M2 anti-FLAG antibodies (called M2 antibodies) on conditioned supernatants, and precipitates were subjected to the Western blot analysis using again M2 antibodies (Figure 35A). The TKN-FLAG western blot showed a pattern described before (Conlon et al. 1988), comforting us in our choice of method to study processing of peptide precursors.

Western blots of S- and N-FLAG samples showed similar band patterns: S-FLAG-spexin transfected cells secrete a protein fragment of about 14-kDa (named P1), which nearly corresponds to the MW of spexin after removal of its signal peptide, and a smaller protein of apparent MW of about 6-kDa (named P4), that may be a processed form of N-FLAG-spexin (Figure 35A, lane 3). N-FLAG-spexin transfected cells were found to secrete a 14-kDa and 12/13-kDa protein that were difficult to resolve on Tris/tricine gels (Figure 35, panel A and C), and one smaller protein of 6-kDa (Figure 35A, lane 4). I reasoned that the 14-kDa band corresponded to the 14-kDa band seen in S-FLAG samples, and that the 12/13-kDa band (P2) corresponded to an N-FLAG-spexin where the region N-terminal of CS1 was removed, thus explaining its absence in the S-FLAG samples. The M1 anti-FLAG antibody was then used that specifically recognizes N-terminal FLAG to confirm that in the S-FLAG fusion protein

cleavage at the predicted signal peptide cleavage site (CS_{SP}) just N-terminal of the FLAG produced the P1 protein (Figure 35D, lane 1). In the same experiment, I witnessed in N-FLAG samples the presence of a 12/13 kDa band, likely to correspond to the P2 protein recognized by the M2 anti-FLAG antibodies, hence showing that cleavage at the predicted dibasic site CS1 does occur in β-TC3 cells (Figure 35D, lane 2). However, if we assume that the high mobility bands seen in both S- and N-FLAG transfections samples correspond to the same protein P4, then it must include the portion between the signal peptide cleavage site (SPCS) and the putative cleavage site CS1. This would mean that CS1 was not used to produce the P4 band, contrasting with my previous observations showing that CS1 was used to produce the P2 protein. The significance of this "OR"-type of processing will be discussed in the next section (conclusions on the secretion and processing of spexin).

To test if the presence of the FLAG peptide influenced processing patterns of spexin, I engineered a DNA molecule encoding a modified spexin protein called C-FLAG-spexin, where a FLAG tag was added at the c-terminus (Figure 34). I transfected β-TC3 cells with this C-FLAG-spexin construct and collected the conditioned supernatant. Two bands were observed in C-FLAG-spexin samples: A low mobility band, corresponding to the 12/13 kDa P2 protein seen in N-FLAG-spexin samples, and a higher mobility band corresponding to a protein fragment which I named P3, and which ran at a MW of about 8-kDa. The apparent 8-kDa band corresponds to a c-terminal product of C-FLAG-spexin after cleavage at CS3 (of theoretical weight 7.5-kDa), while the protein fragment P4 observed in S- and N-FLAG-spexin samples suggested cleavage near the dibasic site CS4 (6.2 kDa). Importantly, the presence of the P2 and P3 proteins shows that the FLAG tag did not interfere qualitatively with secretion and processing. More precisely, it confirmed selective endogenous cleavage of spexin precursor at a position near the cleavage sites CS3 and CS4 (which are only 10 aa apart). The proximity of those two predicted sites is coherent with the existence of a single cleavage site CS3', yielding the two FLAG-fragments P3 and P4, for the two FLAG constructs. The small discrepancy (1-kDa) in the prediction of the cleavage site by the P3 and P4 bands could be explained by the difficulties in precisely assigning apparent MW to peptides in Tris/tricine gel systems.

I then set out to identify the cleavage site responsible for the presence of the multiple band patterns observed in all three S-, N-, and C-FLAG-spexin samples. For this, I modified the

FLAG-spexin constructs by mutating key residues in putative cleavage sites CS1-4 and N-glycosylation site NGLY (cf. Figure 34, and Appendix).

The presence of the protein P4 in mutant samples show that disruptions of critical arginines in the four cleavage sites CS1-4 did not impair processing of N-FLAG-spexin (Figure 35, lanes 1, 3-6). This is in sharp contrast with my previous observations that cleavage at CS1 occurs in the spexin precursor. Before making the mutants, I had hypothesized that the P4 band could be a fully processed spexin peptide with aa ranging from CS1 to CS2, and where n-glycosylation of N36 was responsible for the higher MW than expected (6-kDa vs. 3-kDa). This does not seem to be the case since the presence of a product P5 of clearly lower molecular weight than P4 (Figure 35E, lanes 1-2) in the N-FLAG-Δ51 transfected cells supernatants is good evidence that the CS2 site has not been used to yield the P4 protein.

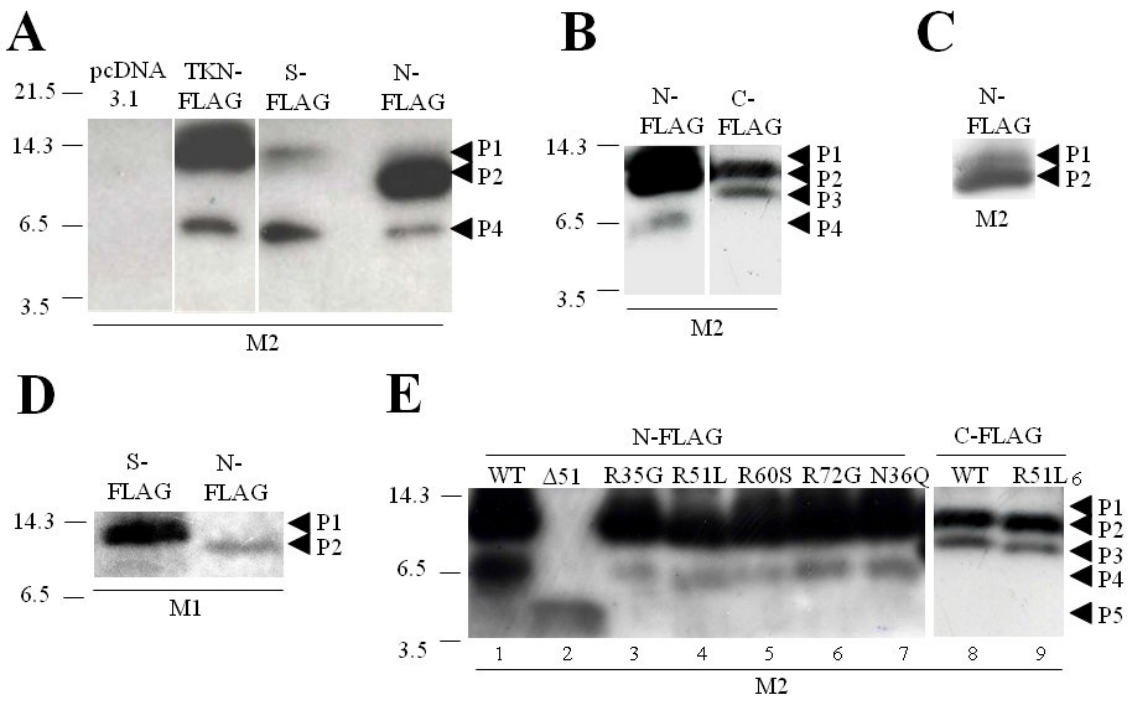

Figure 35: Processing patterns of secreted spexin

Vector control and Flag-tagged NPK, N-FLAG and S-FLAG spexin constructs were transfected into rat pancreatic cells in culture (β-TC3 cell line), and cell supernatants were harvested and submitted to immunoblotting with M1 and M2 anti-FLAG antibodies.

A) M2 anti-FLAG immunoblotting of supernatants from positive control Flag-NPK, S-FLAG-spexin and N-FLAG-spexin transfected cells. The positive control TKN-FLAG is secreted and processed by the cells, as expected. Comparable amounts of S-FLAG-spexin and N-FLAG-spexin were secreted and both showed processing patterns characteristic of PH precursors. N-FLAG transfection samples contained a 14-kDa (P1) and a 6-kDa (P4) protein fragments. Similarly N-FLAG-spexin transfection samples contained a 12/13-kDa (P2) and a 6-kDa (P4) protein fragment. This data reflected processing of FLAG-tagged products from both of these constructs.

B) M2 anti-FLAG immunoblotting of supernatants from N-FLAG-spexin and C-FLAG-spexin transfected cells. C-FLAG-spexin transfected cells secrete a protein which has the same MW as the P2 protein in N-FLAG transfection experiments. A second product is secreted which has an apparent MW of 8-kDa, possibly the complementary c-terminal fragment released with P4 after cleavage by proteases from the PC/Furin family.

C) Low exposure western blot demonstrating the presence of bands P1 and P2 in supernatants of N-FLAG-spexin transfected cells.

D) M1 anti-FLAG immunoblotting of S- and N-FLAG transfected cells supernatants showing that some cleavage occurs just after the predicted signal peptide cleavage site in the S-FLAG-spexin recombinant protein, and just after CS1 in the N-FLAG-spexin.

E) M2 anti-FLAG immunoblotting of supernatants from transfections of N-FLAG, C-FLAG, and mutants derived from those FLAG constructs. Supernatant from N-Flag-□51spexin transfected cells (lane 2) contained a 4-kDa band, suggesting that the 6-kDa product seen for N-FLAG-spexin was the result of cleavage significantly C-terminal to the potential cleavage site CS1. Processing patterns do not seem to be affected from mutations in critical residues of putative cleavage sites, and N-Glycosylation sites, suggesting that none of the cleavage sites proposed where responsible for the processing observed in N-FLAG and C-FLAG precursors, and that N-Glycosylation at residue N36 was not occurring.

Moreover, the pattern observed for the N-FLAG samples was not modified after mutation of the asparagine at position 36 (Figure 35E, lanes 1, 7), arguing against an n-glycosylation of N36. Lastly, when CS2 was mutated in the context of a c-terminal FLAG, the bands pattern remained the same (Figure 35, lanes 8-9); further challenging the hypothesis that CS2 is significantly used in those cells.

3.3.3 Conclusions on the *in vitro* secretion and processing of spexin

Spexin is secreted and processed by pancreatic β-TC3 endocrine cells, as evidenced by the presence in S- and N-FLAG-spexin, and C-FLAG-spexin transfected cells conditioned supernatants of low molecular weight protein fragments P4 and P3. Before starting the experiments I had hypothesized that cleavage at CS2 would be the main cleavage occurring, yielding an amidated peptide that was designated spexin peptide in our publication (Mirabeau et al. 2007). However, the in-vitro processing studies of spexin precursor seem to disprove that hypothesis.

There is strong evidence for the existence of a cleavage site CS3' (Figure 36) situated in the vicinity of CS3 and CS4, but which is unlikely to be any of the two. Furthermore, mutating the asparagine at position 36 into a glutamine did not change the band pattern observed, barring the N36 as a site for n-glycosylation, at least in this cell expression system.

Perhaps the most difficult task I faced was to interpret the data concerning the putative cleavage site CS1. The size of P2 was in good agreement with a FLAG-spexin (36-116) fragment, and the fact that the M2 antibody does recognize a band in N-FLAG transfection experiments suggested that the CS1 was being used by the cells. However, in N-FLAG-spexin experiments, in sharp contrast, I only saw a single 6-kDa band P4, instead of the double band pattern expected if CS1 was used, and mutating the arginine at position 36 of the site CS1 did not change the bands pattern, suggesting that CS1 is not used. There is room for speculation on the reasons for these contradictory observations: the mutation of the arginine might not have been sufficient to disrupt endogenous cleavage of the dibasic site, and this could explain the negative results of the mutant experiments. An explanation for the absence of a double band could be that in order to make the 5-kDa protein, the cell needs to process two different cleavage sites (CS1 and CS3'), making the quantity of the fully cleaved product lower than the one of the partially cleaved ones. Furthermore, transfer of small proteins to membrane is

not trivial. It is known that binding efficacy of small proteins to the membrane is subject to large variations when the experimental transfer conditions are even slightly modified (e.g. methanol percentage in transfer buffer, voltage and duration of transfer). Furthermore, different proteins of same size do not equally interact with the membrane, for identical experimental transfer conditions.

Despite those unsolved questions, I have strong evidence arguing that spexin is secreted and processed, at least in the cell lines I have used. More powerful biochemical techniques, including mass spectrometry and MALDI-TOF analysis of the protein products of spexin processing should confirm the exact cleavage site(s) of the spexin precursor *in vitro*. I have tried such approaches, in collaboration with the EMBL Protein Core Facility. Probably due to insufficient amounts of spexin produced by the cells, I was not able to identify spexin products in cell lysates or supernatants, by MALDI-TOF techniques. Only the *in vivo* isolation and biochemical characterization of the endogenous spexin peptide will provide the final answer to the riddles that I have tried to answer, namely where is spexin precursor precisely cleaved to yield the bioactive product?

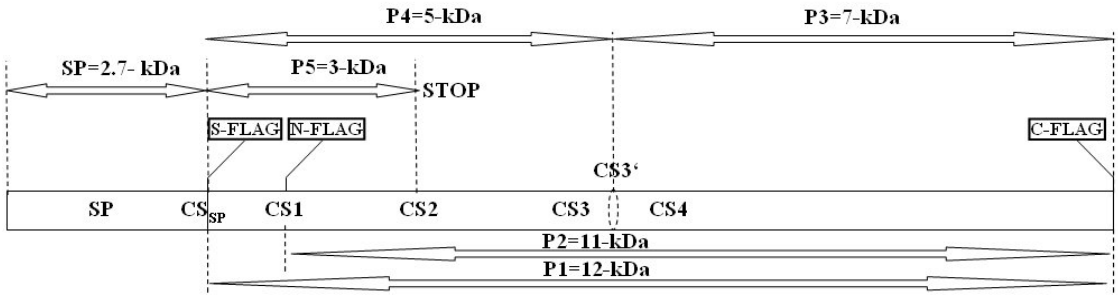

Figure 36: Model of spexin *in vitro* processing

This diagram shows the primary structure of the spexin protein, and proposes a cleavage model that explains the bands patterns that I observed in the transfection/western blot experiments. The positions of the signal peptide cleavage site (CS_{SP}), and the dibasic putative cleavage sites CS1-4 are indicated.

CS3' corresponds to a PC/Furin cleavage site inferred from the western blots experiments. The band P5 was obtained from transfection of a truncated spexin protein construct, N-FLAG-□51.

Sizes shown in the diagram correspond to the protein fragments without their FLAG tags.

3.4 *In vitro* subcellular localization of spexin

In endocrine cells, PH are packaged into secretory granules called dense core granules for their high protein density. PH secretion is regulated: exocytosis follows in response to specific hormonal or neuronal signals and dense core granules release their PH contents into the extracellular space (Burgess and Kelly 1987; Seidah et al. 1996).

Immunocytochemistry with FLAG antibodies following transfection of FLAG-NPK and N-FLAG-spexin into a rat pancreatic cell line (RINm5f, or RIN) demonstrated colocalization of FLAG antigen with endogenous insulin in punctate intracellular bodies (Figure 37) which resembled dense core granules. This colocalization suggested that spexin, like neuropeptide K, underwent trafficking into dense core granules of the secretory pathway, a hallmark of PH. Although the degree of colocalization with insulin was not quantified, anti-FLAG staining of RIN cells after transfection of the N-FLAG-spexin construct showed a striking similarity with anti-insulin staining, suggesting that spexin is a PH.

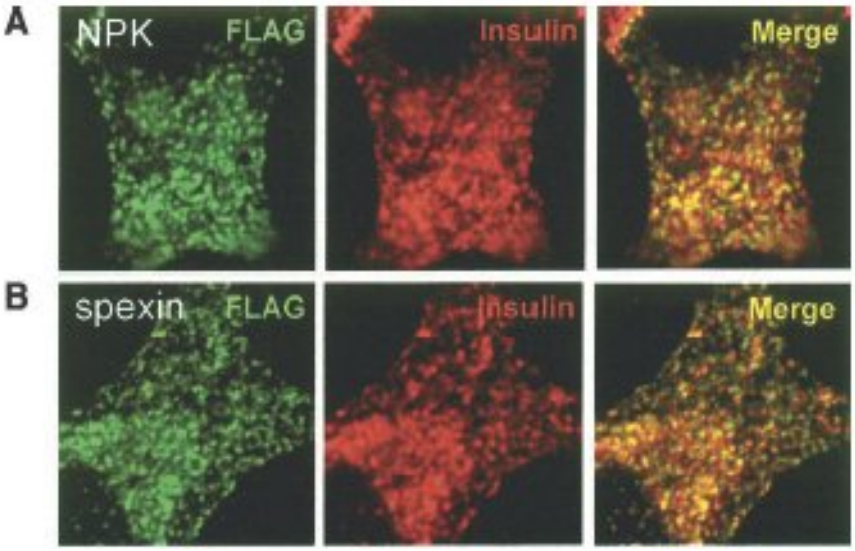

Figure 37: Colocalization of spexin with insulin in endocrine cells.
Flag-tagged NPK and spexin were transfected into rat pancreatic cells and fixed cells were subjected to double immunofluorescence with FLAG and insulin antibodies. Confocal microscopy (Leica Microsystem) analysis revealed that Neuropeptide K (A) and spexin (B) colocalize with insulin in small, cytoplasmic punctate structures with the same characteristics as dense core granules.

3.5 Expression of spexin in mouse tissues

In this section, I present data on spexin mRNA localization in mouse tissues. I first characterize the spexin transcripts that were cloned from mouse tissues (3.5.1), then I present profiles of spexin mRNA expression by RT-PCR in different mouse tissues (3.5.2). In subsections 3.5.3 and 3.5.4, I present data on the *in situ* localization of spexin mRNA in the stomach, oesophagus and brain.

3.5.1 Cloning and characterization of spexin transcript(s)

Spexin transcript was found in a small number of tissues, and not in great quantity, suggesting that it is only expressed in a few cell types. Spexin mRNA was difficult to amplify, and I probed different sets of primers before observing specific bands on the agarose gel. This is likely to be due to two factors. First and foremost, there are reasons to believe that spexin has a low expression in most tissues making the amplification of a low copy number technically more difficult. Second, the region close to the spexin starting methionine codon, where I tried to amplify the transcript, happens to be GC- rich, a factor known to make amplification of DNA more difficult, requiring modifications of standard PCR protocols such as DMSO and betaine inclusion (Henke et al. 1997; Sun et al. 1993). After several attempts, I was able to amplify spexin transcript cDNA, using the following pair of primers:

m56F: 5′-ACAGGGTCGAACATGAAGTAGGG-3′

m56R: 5′-AAGAGTCTGTCTTCCAAGAGTTCGC-3′

After running the PCR product on an agarose gel by electrophoresis I obtained two bands (338 bp and 612 bp, cf. Figure 39) which were cloned and sequenced. I could verify that the lower band (338 bp) corresponded to the expected spexin transcript, as described in the Ensembl database (Figure 38).

Sequence of m56_transcript (lower band):

5′CCCAGCGTCCTGGCAGTGACAGCCGTGGTCCTTCTCCTGGTGCTGTCTGCGCTGGAAAACTCCAGC
GGTGCTCCACAGGCACTGCAGCGACTCTCTGAGAAGAGGAACTGGACTCCCCAAGCTATGCTCTATCT
GAAGGGTGCACAGGGCCGCCGCTTCCTCTCCGACCAGAGCCGTAGGAAGGAGCTTGCAGACCGGCCGC
CTCCAGAAAGACGAAACCCAGATCTTGAACTGCTGACTCTCCCAGAGGCTGCAGCCCTGTTTCTGGCT

TCCTTGGAAAAATCACAAAAAGATGAAGGAGGGAATTTTGATAAAAGCGAACTCTTGGAAGACTCTT-3′

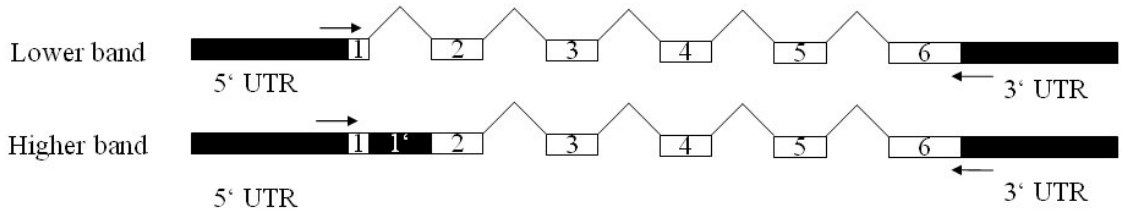

Figure 38: sequences of the two spexin bands amplified, and their relationship with the spexin gene structure

The upper band was sequenced and, unexpectedly, corresponded to a spexin transcript where the first intron (1' in Figure 38) was transcribed. The protein corresponding to this hypothetical transcript would not be functional due to a frameshift created by this insertion. It is thus likely that the piece of DNA corresponding to the higher band is an artefact of the PCR. Furthermore, the unfortunate choice of forward primer that spans exon 1 and 2 contributes to my hunch that the band does not have a biological significance. It is reasonable however to assume that the DNA corresponding to the upper band was created by the PCR from spexin transcripts.

3.5.2 Expression of spexin mRNA by RT-PCR

Several tissues were probed for spexin RNA expression, including heart, testis, uterus, thymus, stomach, gut, adrenal glands, hypothalamus, pituitary, and pancreas (Figure 39). Spexin mRNA (lower band) was found at low levels in the thymus, adrenal glands, cerebral cortex, and brain stem. The upper band (an artefact attesting to the presence of spexin transcripts) was found in the cerebellum. The amplification of spexin transcripts with the primers m56F and m56R proved to be difficult, and success depended highly on the PCR experimental conditions. I was nonetheless able to confirm those results at least once. To further investigate on the *in vivo* expression of spexin, I performed Northern blots experiments on mouse tissues; using the 338-bp probe that was amplified by RT-PCR. I could not see bands for the tissues probed, including heart, lung, liver, testis, uterus, kidney, spleen, thymus, stomach, gut, skeletal muscle, adrenal glands, olfactory bulb, cerebral cortex, cerebellum, brain stem, hypothalamus, pituitary, hippocampus, and pancreas. Precious high-throughput Affymetrix-based expression data from the SymAtlas public database (Walker and

Wiltshire 2006) and mouse *in situ* hybridization data made publicly available in the form of the Allen Brain Atlas (Lein et al. 2007) further supported the idea that spexin expression is very low in most tissues. I have also tried to see expression of spexin *in situ*, using the same probe as for the Northern blots experiments, but I could not detect any expression in sections of the brain, adrenal glands, pituitary and testis.

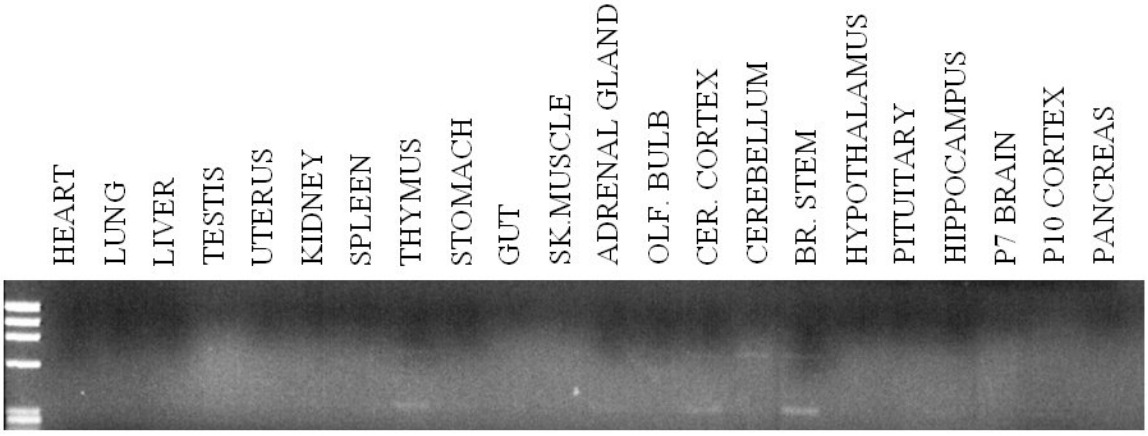

Figure 39: spexin expression by RT-PCR
Total RNA was extracted from heart, lung, liver, testis, uterus, kidney, spleen, thymus, stomach, gut, skeletal muscle, adrenal glands, olfactory bulb, cerebral cortex, cerebellum, brain stem, hypothalamus, pituitary, hippocampus, pancreas, P7 whole brain and P10 cerebral cortex tissues. Spexin expressed sequence was then amplified by PCR using the primers m56F and m56R and 2 distinct DNA bands of 338 and 612 bp were isolated and sequenced. The lower band corresponded to the expected spexin transcript. Spexin mRNA was found at low levels in the thymus, adrenal glands, cerebral cortex, and brain stem.

3.5.3 Expression of spexin mRNA in the gastro-oesophageal system by *in situ* hybridization

Next, the expression of spexin was inspected in a set of tissues where PH are typically expressed, such as the pituitary, adrenal glands and gastrointestinal tract, by standard *in situ* hybridization techniques and using the spexin probe "m56_transcript (lower band)" described above. We could clearly detect a restricted, yet specific, signal in discrete cells of the mucosal layer of the stomach fundus (Figure 40A), oesophagus (Figure 40AB), and lower oesophageal sphincter (LES), (Figure 40C).

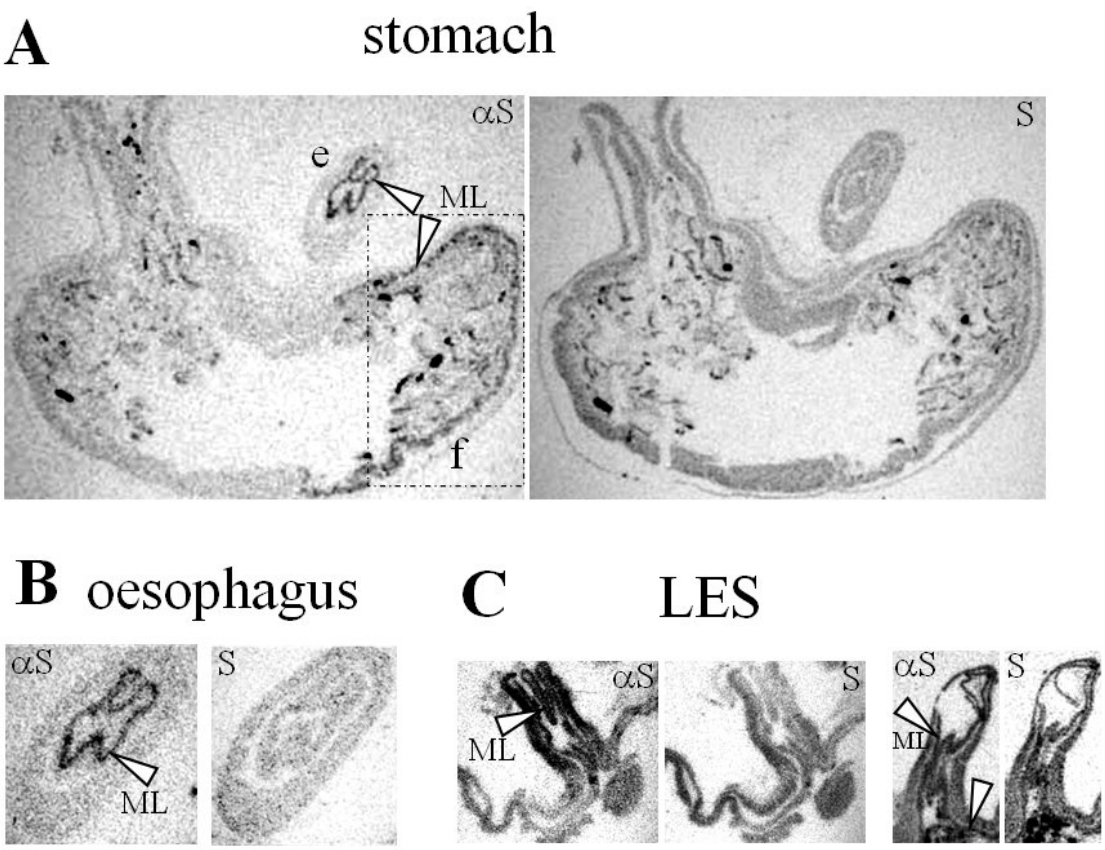

Figure 40: spexin mRNA expression in the stomach, oesophagus, and lower oesophageal sphincter (LES)

In situ hybridizations were performed with sense (S) and anti-sense probes (αS) on different sections of the mouse oesophagus and stomach. We could detect spexin in mucosal layers of the gastric fundus, oesophagus, and LES. Note the rugae (foldings) formed by the fundus spexin-expressing layer of cells, characteristic of the gastric mucosa and submucosa. (S, sense probe; αS, anti-sense probe; e, oesophagus; f, fundus; ML, mucosal layer; LES, lower oesophageal sphincter)

The LES spexin-expressing layer seemed to be continuous with the fundus spexin-expressing layer. From its shape, width, and depth, I could infer that both spexin-expressing layers in the stomach and the oesophagus were part of the mucosa (ML) (Figure 41). In contrast, we did not detect expression of spexin in the intestine, pituitary and adrenal glands. More work is needed to tackle the following question: which types of cells express spexin, and what are their anatomical relationships with smooth muscle cells, and mucosal/submucosal endocrine cells? A more thorough characterization of spexin expression at the cellular level is warranted, that should already give us better insights into the possible roles of spexin in the physiology of the gastro-oesophageal tract.

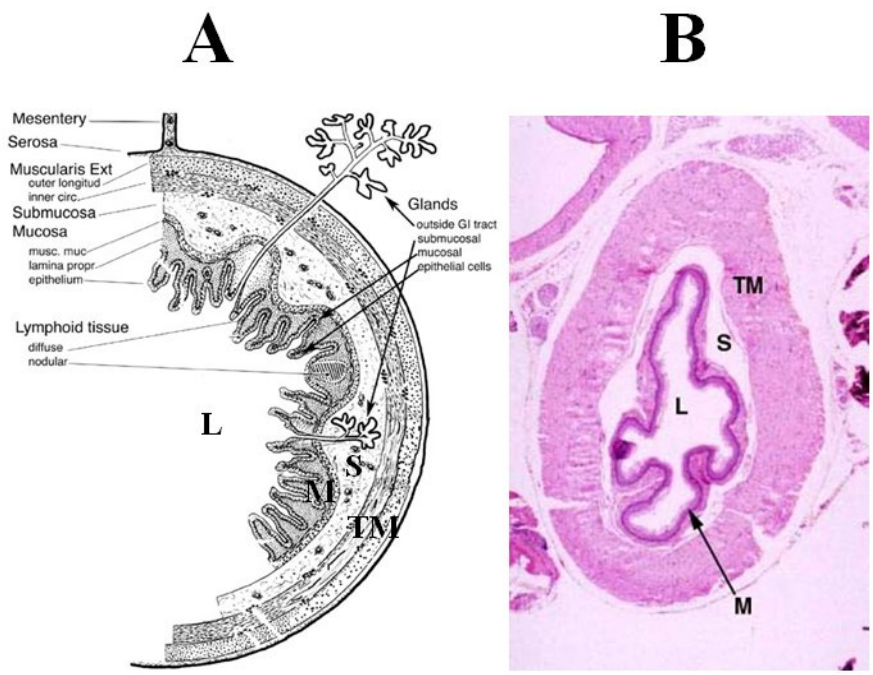

Figure 41: Histology of the stomach.
(A) representation of a stomach section, with annotations. [Source : histology textbook of (Gartner and Hiatt 2001) (B) representative staining of a section of the oesophagus, with annotations.[Source : **http://education.vetmed.vt.edu**]
L, Lumen; M, Mucosa; S, Submucosa; TM, Tunica Muscularis

3.5.4 Expression of spexin in the brain by *in situ* hybridization

I used the extraordinary source of data recently provided by the Allen Brain Atlas project (Lein et al. 2007) to enquire about spexin mRNA expression in the brain. According to this source of data, spexin mRNA is specifically expressed in cells (probably neurons) of the pedunculopontine tegmental nucleus (PPTN), laterodorsal tegmental nucleus (LTD) (Figure 42), and possibly, the pontine central gray (data not shown). A clear signal is also observed in the medial and lateral habenula (Figure 42).

Confirming the expression of spexin in these regions of the brain will be important as I believe this localized pattern of expression in the brain gives important cues for testing spexin function(s). Whether or not the spexin-expressing cells are neurons needs to be determined, although their scattered localization in those regions of the brain strongly suggests that they indeed are.

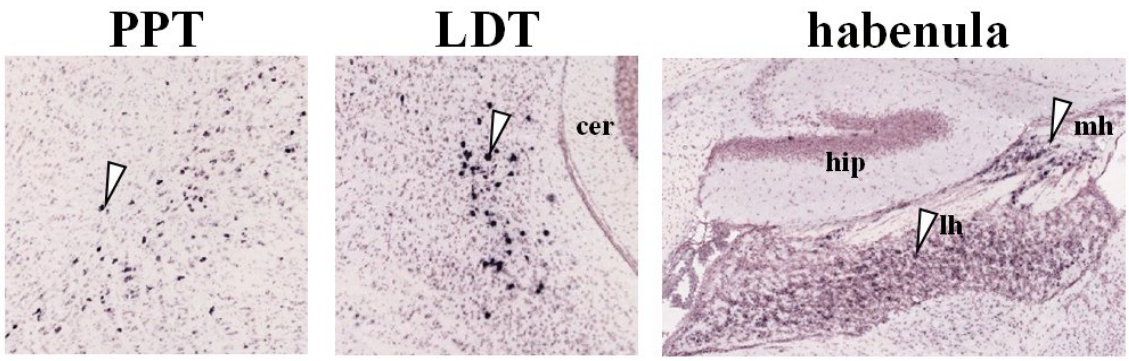

Figure 42: spexin expression in the brain [Allen Brain Atlas, http://www.brainatlas.org]
Spexin is expressed in discrete sets of cells, in at least three brain areas, that included the pedunculopontine nucleus (PPT), laterodorsal (LTD) nucleus, lateral and medial habenulae. Arrows indicate spexin-expressing cells. PPT, pedunculopontine nucleus; LDT, laterodorsal tegmental nucleus; cer, cerebellum; hip, hippocampus; mh, medial habenula; lh, lateral habenula.

3.6 Localization in the gastro-oesophageal system by immunohistochemistry

To study mature spexin protein localisation in the mouse, we had rabbit polyclonal antibodies raised against the conserved core antigenic sequence of spexin precursor protein (Primm, Milan, Italy), Asn-Trp-Thr-Pro-Gln-Ala-Met-Leu-Tyr-Leu-Lys-Gly-Ala-Gln-NH_2. Immunohistochemistry by the DAB staining method (cf. Materials and Methods) done on mouse stomach tissues using that antibody (called α-spexin) revealed a pattern of expression which corroborated our *in situ* hybridization data on spexin mRNA localisation in the LES. Figure 43 shows strong immunoreactivity of spexin in the mucosal layer of the mouse LES suggesting that spexin is a hormone involved in the homeostasis of the gastro-oesophageal system.

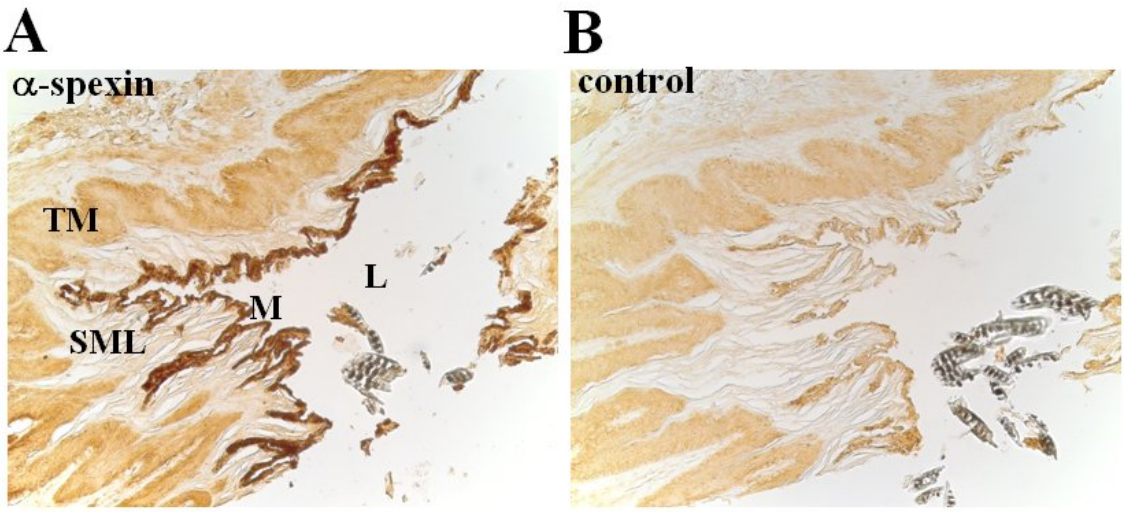

Figure 43: anti-spexin immunohistochemistry of mouse LES
There is strong anti-spexin immunoreactivity in the mucosa (M) of the mouse LES (panel A). Spexin IR seems to be restricted to this layer. We used rabbit polyclonal antibodies. The control corresponds to a procedure where the primary anti-spexin antibody was left out.
L, Lumen; M, Mucosa; S, Submucosa; TM, Tunica Muscularis

3.7 Biological activity of spexin peptide

Many PH have been shown to contract smooth muscle in fundal regions of the stomach, which express a great variety of G protein-coupled receptors. Testing contractility of those muscles ex vivo has emerged as a standard technique to screen and test the activity of peptides (Bolle et al. 2000; Davies et al. 2007). To test bioactivity of the predicted spexin peptide and to test the hypothesis that spexin acts on smooth muscle cells to contract/relax them *ex vivo* smooth muscle contractility experiments were performed in collaboration with Roberta Possenti and Cinzia Severini of the University of Tor Vergata in Rome. Muscle strips were taken from the fundal region of the stomach of live rats, and incubated with the chemically synthesised 15 aa peptide Asn-Trp-Thr-Pro-Gln-Ala-Met-Leu-Tyr-Leu-Lys-Gly-Ala-Gln-NH_2 (Primm SRL, Milan, Italy) that was predicted to be the active peptide in our publication (Mirabeau et al. 2007) and that we named spexin peptide. An apparatus was used that measured contraction amplitudes of the smooth muscle strips. Spexin contracted the muscle in a dose-dependent manner (Figure 44) suggesting that it is a bioactive PH. Furthermore, since it is expressed in the same fundal region where muscle was extracted to test for its contractility, it is reasonable to predict that spexin plays a role in the normal physiology of the fundus. It is noteworthy that spexin contraction of muscle was not changed

in the presence of atropine, a muscarinic cholinergic antagonist. These findings suggest that spexin does not act upstream of a cholinergic transmission mechanism (data not shown).

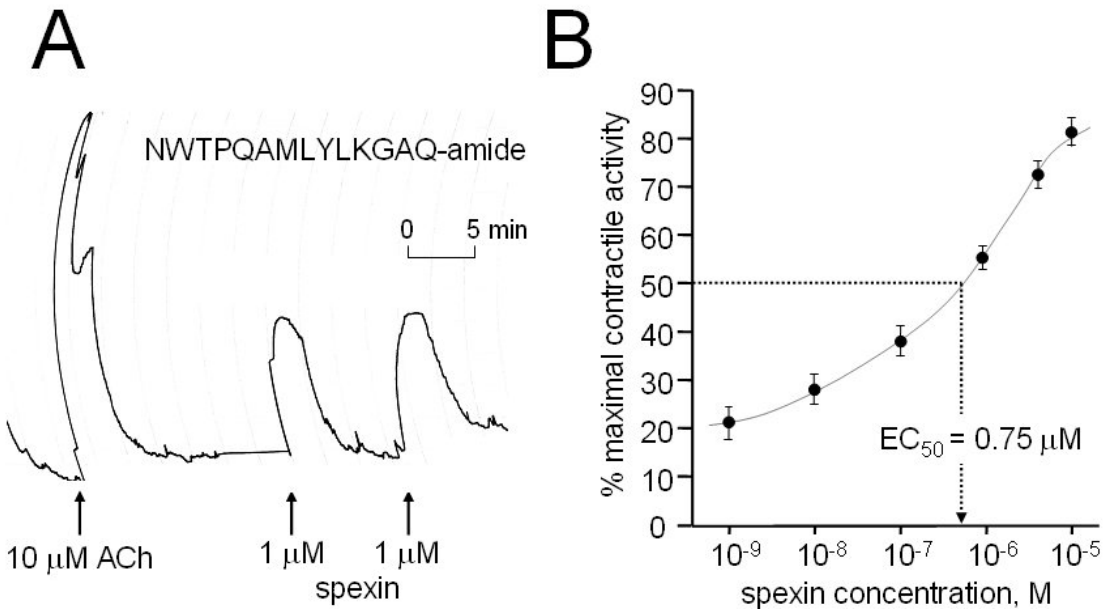

Figure 44: Spexin is a biologically active PH.
(A) Representative muscle contractile response to 10 μM acetylcholine (ACh) and 1 μM spexin peptide (NWTPQAMLYLKGAQ-amide) in a rat stomach explant assay. Repeated administration of spexin peptide produced similar contractile responses. (B) Cumulative dose-response curve for contractile activity of spexin peptide on rat stomach explants (EC50 = 0.75 μM, N = 6). Error bars indicate standard error.

3.8 Possible physiological function(s) of spexin

In this section I speculate on the possible functions of spexin, based on its restricted expression in the gastro-oesophageal system and the brain.

3.8.1 Possible role of spexin in the gastro-oesophageal homeostasis

The main function of the oesophagus is to drive the ingested food (bolus) down to the gastrointestinal tract for its digestion. For this, the oesophagus is traversed by mechanical waves which accompany the bolus all the way down to the stomach, and prevent regurgitation of the ingested food. This process, called peristalsis, necessitates coordination of smooth muscle contractions to form a unidirectional wave (Creamer and Schlegel 1957). The stomach fundus is designed to mash the bolus to prepare for its digestion, and orchestrate storage and emptying of gastric contents. The lower oesophageal sphincter (LES) is a transitory region

between the stomach and the oesophagus which serves as a one-way valve that lets the ingested food inside the stomach and keeps the gastric juice from entering the oesophagus and damaging the oesophageal mucosa. In both of those tissues, smooth muscle cells surrounding the spexin-expressing layer contract in response to specific signals in a coordinated manner to allow those regions to perform their important tasks. The oesophagus and LES, are known to act in a concerted manner to orchestrate oesophageal peristalsis (Spechler and Castell 2001; Zerbib et al. 1998). The fundus contracts differently in presence of solid or liquid food, and responds to signals to empty the gastric contents when digestion is complete. It is not surprising to find that those finely tuned processes are regulated by a number of peptidergic systems. Peptides hormones involved in the physiology of the LES include gastrin which have been shown to contract human LES (Cohen and Lipshutz 1971), vasoactive intestinal peptide (VIP) which has a relaxation effect on human, cat, and baboon LES (Biancani et al. 1984; Siegel et al. 1979) and cholecystokinin which increases transient LES relaxations in humans (Resin et al. 1973). Furthermore, motilin stimulates contractions of the LES in the opossum, while secretin has an inhibitory action on LES contraction in the opossum (Gutierrez et al. 1977), baboon (Siegel et al. 1979) and humans (Cohen and Lipshutz 1971). VIP, secretin and glucagons were all shown to relax LES smooth muscles in baboon (Siegel et al. 1979), and gastrin-releasing peptide (GRP), substance P, somatostatin, and neurotensin are secreted by the endocrine mucosal cells of the stomach in the axolotl (Maake et al. 1999). Spexin localization in the stomach fundus, oesophagus, and LES suggests that it is involved in normal gastric emptying and gastrooesophageal peristalsis.

3.8.2 Spexin, a novel neuropeptide modulating reward mechanisms?

Data from the Allen Brain Atlas (cf. Figure 42) suggest that spexin is expressed in the pedunculopontine tegmental nucleus (PPT) and laterodorsal tegmental (LTD) nuclei. Adjacent PPT and LTD nuclei are involved in addiction/reward mechanisms (Bardo 1998; Mena-Segovia et al. 2004; Wise 1996), and are densely connected to the basal ganglia and dopamine system (Mena-Segovia et al. 2004). It has been shown that PPTN neurons project to the substantia nigra pars compacta (SNc) and provide excitatory cholinergic and glutamatergic inputs to SNc dopamine neurons (Futami et al. 1995; Scarnati et al. 1984).

Spexin is also conspicuously present in the medial and lateral habenulae, which have also been associated with the reward system. Interestingly, afferent synapses in the habenula have

also been shown to provide extensive input to the SNc (Herkenham and Nauta 1979; Sutherland 1982). Very recently, neurons in the lateral habenula of monkeys have been shown to fire in situations associated with negative reward (Matsumoto and Hikosaka 2007). In this paper the authors argue that lateral habenula neurons provide inhibitory input to nigro-striatal dopamine neurons. Strengthening those claims, another recent study argues that the firing of lateral habenula neurons indirectly inhibits the activity of nigro-striatal dopamine neurons by activating inhibitory GABAergic neurons of the ventral midbrain (Ji and Shepard 2007). It is noteworthy that putative spexin-expressing neurons in the PPT/LDT provide excitatory input to ventral dopamine neurons (presumably Ach/Glu), whereas neurons from the lateral habenula provide inhibitory input (presumably GABA). Given that its expression in the brain is remarkably restricted to the lateral habenula and PPT/LTD nuclei, chances are that spexin modulates the reward-related activity of SNc/VTA dopamine neurons.

3.9 Possible links between spexin and human disease

In this section I discuss a number of diseases in which spexin could be involved. Spexin's possible involvement in common diseases is of particular interest, since it is likely to be secreted *in vivo*, and as such, spexin analogs constitute potential drugs. However, one must be aware that only the additional knowledge of its cognate receptor would in practice enable the screening for spexin analogs from chemical compounds banks.

3.9.1 Gastro-oesophageal reflux disease (GERD)

Gastro-oesophageal reflux (GER) occurs naturally in healthy patients (e.g. belching). In non-pathological GER, fundus, LES and oesophageal smooth muscle systems coordinate to regulate entrance of food into the stomach, and prevent gastric acid from exiting the stomach and damaging oesophageal tissues. Normal GER are generally caused by abnormal transient relaxation events in the LES, which cause an influx of gastric acid into the oesophagus, which is then immediately cleared by peristaltic events in healthy subjects (Dent et al. 1980).

Deregulation of the LES system is the major cause of GERD (Ogorek and Cohen 1989). The main symptoms of GERD include frequent heartburns, also called acid indigestion, a burning-type pain occurring in the lower part of the mid-chest. It is estimated that from 10 to 40 % of the European population suffers from heartburns (Mahmood and McNamara 2003). GERD can have several causes including oesophageal motility (Spechler and Castell 2001) and aberrant LES relaxation. Localization of spexin in the LES, together with the stomach smooth

muscle contractility data, is good evidence that spexin is likely to be involved in the contraction/relaxation of the LES. Furthermore, episodes of transient LES relaxations in healthy subjects (those which are not swallow-induced) are triggered by stimulation of the vagus nerve in which afferent input from receptors in the gastric fundus and pharynx is integrated in the central nervous system (Zerbib et al. 1998). Other PH have been involved in the normal physiology of the LES and GERD, including cholecystokinin (Clave et al. 1998) and it is possible that spexin is also involved in normal and pathological GER.

3.9.2 Parkinson disease

The PPT is one of the regions of the brain which is highly associated with Parkinson's disease (Pahapill and Lozano 2000). That region is with the substantia nigra, the only region in which substantial degeneration occurs (Hirsch et al. 1987; Pahapill and Lozano 2000). Parkinson disease symptoms, including defects in swallowing, oesophageal peristalsis, and pathological gastro-oesophageal reflux, cannot be explained entirely by nigro-striatal dopamine deficiency (Hunter et al. 1997). In two papers, (Alfonsi et al. 2007; Hunter et al. 1997), the authors argue that swallowing difficulties encountered by Parkinson patients are due to disturbances in PPTN regulation of the central pattern regulator in the basal ganglia, to which it is highly connected (Mena-Segovia et al. 2004). We have yet to investigate whether spexin is expressed in the upper gastro-oesophageal sphincter (pharynx), which controls swallowing. Nevertheless, given its presence in both the PPT and gastro-oesophageal mucosa, and its possible involvement in the regulation of nigro-striatal dopamine neurons (cf. 3.8.2), spexin stands out as a putative effector molecule in the regulation of processes that are involved in Parkinson disease, including PPT regulation of the basal ganglia and gastro-oesophageal symptoms. Thus, drugs mimicking or antagonizing spexin may have therapeutic value in the treatment of Parkinson disease.

4 Characterization of augurin

In this chapter I present data that has been obtained on augurin, the second candidate PH, and I discuss its possible function(s).

The structure of this chapter follows closely that of the previous one. In the first two sections (4.1) and (4.2) I present what is known about augurin gene structure and the conserved elements of the protein primary structure that the augurin gene encodes. In sections (4.3) and (4.4), I discuss the data I obtained during my thesis on the secretion, processing and subcellular localization of FLAG-tagged augurin proteins in a cell expression system. In section (4.5) the results obtained on augurin mRNA expression in mouse tissues are presented. In section (4.6) I present results on the augurin protein localization in mouse tissues, by immunohistochemistry. The purpose of the last three sections (4.7), (4.8) and (4.9) was to propose various hypotheses for augurin's function in the body, speculate on disease it may be involved in and on possible connections to other peptidergic systems.

4.1 Gene structure of augurin

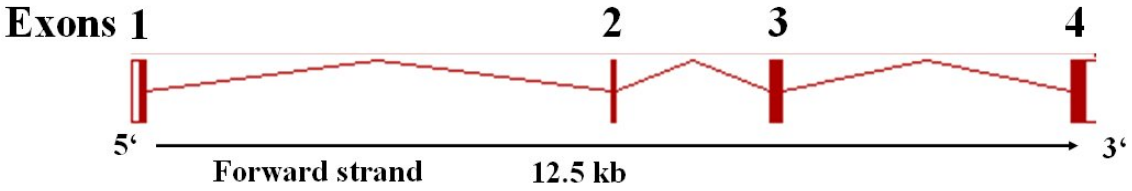

Figure 45: Structure of the augurin gene structure [source:Ensembl]
Human augurin has four exons all of which are coding. Filled red rectangles represent coding-regions of exons, while empty red rectangles define boundaries of non-coding regions of exons. Human augurin is found in chromosome 2 in spans 12.5 kb.

The augurin gene has 4 exons, all coding. No alternative splicing has been reported yet in mammals. However, there are two paralogous augurin genes in the zebrafish, located on two different chromosomes (zebrafish-1 in chromosome 7 and zebrafish-2 in chromosome 8). I noted a strong divergence between the two zebrafish augurin proteins in the putative pro-peptide region.

4.2 Primary sequence features of augurin protein

In this section, I present an analysis of primary sequence features (signal peptide, cleavage site, and putative peptides) and their conservation using augurin orthologous sequences that were made available in the Ensembl database, including (using Linnæus taxonomy) *Macaca*

mulatta (macaque), *Homo sapiens* (human), *Canis familiaris* (dog) *Spermophilus tridecemlineatus* (squirrel), *Equus caballus* (horse), *Myotis lucifugus* (bat), *Rattus norvegicus* (rat), *Mus musculus* (mouse), *Otolemur garnettii*, (lemur), *Bos taurus* (cow), *Cavia porcellus* (guinea pig), *Echinops telfairi* (lesser hedgehog), *Monodelphis domestica* (opossum), lizard, *Ornithorhynchus anatinus* (platypus), *Xenopus tropicalis* (frog), *Gallus gallus* (chicken), *Danio Rerio* (zebrafish), *Tetraodon nigroviridis* (tetraodon), *Gasterosteus aculeatus* (stickleback fish), *Takifugu Rubripes* (fugu).

4.2.1 Signal peptide

N-terminal augurin sequence clearly contains all the characteristic features of a signal peptide: one basic residue close to the starting methionine, a core hydrophobic part (visualised in blue) of about 15 aa and a poorly conserved region c-terminal of the hydrophobic region, which contains many small uncharged residues (isoleucines, valines, glycines and alanines). This alignment illustrates the known fact that the signal peptide length is a variable under low selection pressure, as is the exact signal peptide sequence. For example, the frog SP length (about 15 aa), is twice as long as the opossum SP length (about 32 aa), (cf. Figure 46) and their sequences show now strict homology.

4.2.2 PC/Furin cleavage sites

All three dibasic putative cleavage sites for human and mouse sequences are also predicted with high probability (>0.88) to be PC/Furin CS by the Neuropred software, a program dedicated to prohormone convertase cleavage sites. According to the Neuropred algorithm, R70 is the most likely to be a CS, at a probability of 0.9975. After redoing alignments with sequences of recently available genomes from other organisms (guinea pig, lizard, several fishes), it appeared that the CS1 and CS2 were not conserved, while CS3 retained its prefect conservation across all organisms. I will show in the next sections that this site is indeed used by β-TC3 endocrine cells, while it is unclear whether the others are. This result illustrates the power of comparative orthology studies when done across multiple organisms.

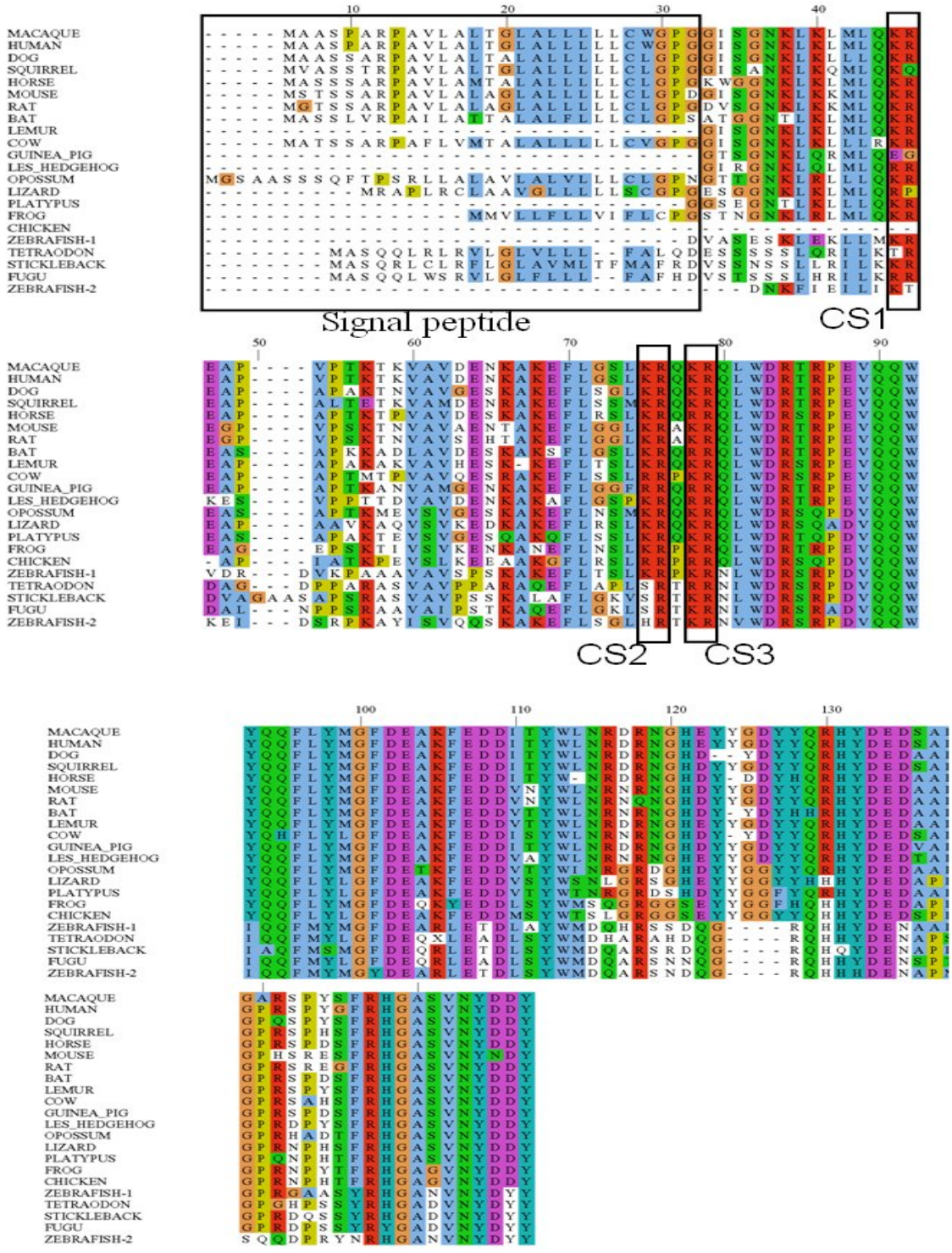

Figure 46: alignment of augurin orthologs.

Augurin orthologs were collected from the Ensembl database. Additional sequences were retrieved with the UCSC genome browser, including those from lizard, horse, opossum and platypus genomes. Proteins were aligned with ClustalW and visualised with the Jalview software. The putative signal peptide is indicated and the 3 putative PC/Furin cleavage sites are denoted CS1-3.

4.3 *In vitro* secretion and processing of augurin

In this section I present the results of experiments I conducted to test if augurin is secreted and processed by cells in culture.

4.3.1 FLAG-tagging strategy

The same strategy as for spexin was used to study intracellular trafficking of augurin (Figure 47).

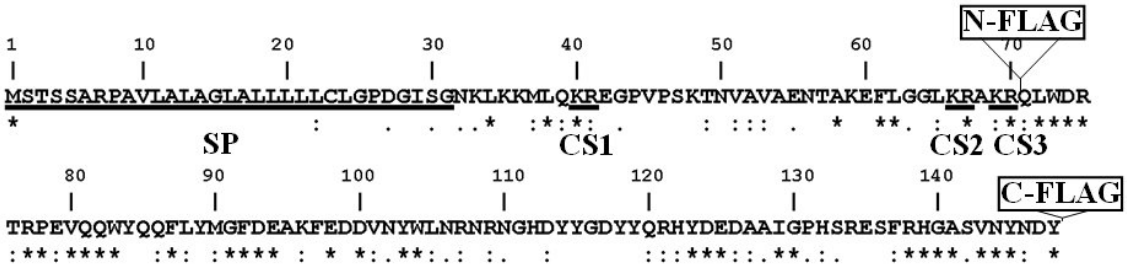

Figure 47: primary structure of mouse augurin, and description of FLAG constructs
SP denote the predicted signal peptide (underlined). CS1-3 indicate the positions of putative dibasic (underlined) cleavage PC/Furin cleavage sites (see also Figure 34). N-FLAG, C-FLAG shows the positions where the FLAG tag was inserted, respectively after CS3 and at the c-terminus of the precursor. Conservation among orthologs is show below: (*) identity, (:) high homology, (·) low homology.

4.3.2 Immunoblotting of FLAG-augurin peptides

Western blotting of supernatant from FLAG-augurin transfected cells revealed a pair of FLAG-immunoreactive bands consistent with secretion of the pro-peptide and a processed variant (11 and 9 kDa), (Figure 48A). Recognition of the N-FLAG-augurin products by a FLAG antibody that binds only N-terminal FLAG antigen (M1) suggests that cleavage occurred at the predicted dibasic cleavage site CS3, just upstream of the FLAG tag. The protein P2 apparent MW of 15-kDa suggests that it corresponds to the c-terminal product of FLAG-augurin cleavage at the signal peptide cleavage site (theoretical MW=15-kDa), (Figure 48Figure 49). Furthermore, the 11-kDa protein P3 has the size of the c-terminal product of FLAG-augurin cleavage at CS3 (10.7-kDa). If we assume that both P3 and P4 are c-terminal products of FLAG-augurin at CS3, then the presence of the 9-kDa fragment P4 would require cleavage at a site CS4' near the augurin c-terminus (Figure 49). As there are no conserved

arginine in this region (Figure 46, Figure 47), it was difficult to make a hypothesis on the exact location of the cleavage site CS4'.

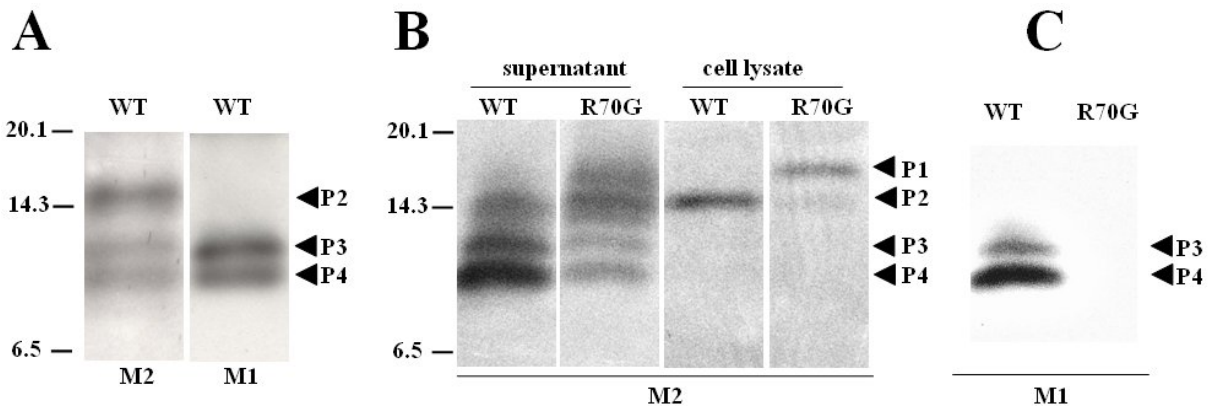

Figure 48: secreted augurin processing patterns

CMV-FLAG-augurin constructs were transfected into pancreatic β-TC3 cells, conditioned supernatants and cell lysates were collected and submitted M1 and M2 anti-FLAG immunoblotting

A) M1 and M2 anti-FLAG immunoblotting of supernatants from N-FLAG-augurin transfected cells. Supernatants were passed through an M2 anti-FLAG affinity column and the retentate was directly re-suspended in tricine buffer, as described in the Materials and Methods section.

Three proteins P2 (15-kDa), P3 (11-kDa) and P4 (9-kDa) are recognized by M2 antibodies, while only P3 and P4 are recognized by M1 antibodies, suggesting that the site CS3, where the FLAG was inserted, is processed to yield two N-terminal-FLAG products (P3 and P4).

B) M2 anti-FLAG western blots of conditioned supernatants and lysates of β-TC3 cells that were transfected with the N-FLAG-augurin and CS3 mutant N-FLAG-R70G-augurin constructs. Supernatants were passed through an M2 anti-FLAG affinity column and the retentate was directly re-suspended in sample buffer, while cell lysates were collected, a M2 anti-FLAG IP was performed, and precipitates were re-suspended in sample buffer. A difference in processing patterns can clearly be seen, both in supernatants and cell lysates. For N-FLAG-augurin transfected cells, the same band pattern (bands P2-4) as in panel A is observed. While in the case of N-FLAG-R70G-augurin transfected cells, I could see the same three bands, and a fourth one P1 running at 18-kDa. Furthermore, there appears to be a shift in the relative intensity of the bands. The lower mobility bands corresponding to P1 and P2 appear more intense in the mutant, in line with my hypothesis that R70 is required for proper processing of augurin precursor. In cell lysates, only larger proteins P1 and P2 are visible, suggesting that, as is commonly the case in endocrine processing studies, only partially processed forms of the protein are detectable inside the cells. As in the supernatants samples the banding pattern is different between N-FLAG-augurin and N-FLAG-R70G-augurin transfected cells, suggesting that the arginine at position 70 is involved in processing of augurin.

C) The same blot as in panel B was probed with anti-FLAG M1 antibodies, which specifically recognize N-terminal FLAG antigens. The two high mobility bands corresponding to P3 and P4 were recognized, only in the case of N-FLAG augurin, but not in the case of the mutant N-FLAG-R70G-augurin, confirming that the augurin predicted cleavage site CS3 is used in β-TC3 cells, yielding two processed products P1 and P2.

To get better insights into the processing of augurin products at its c-terminal end, I generated a construct coding for a protein where a FLAG peptide was placed at the end of the augurin precursor (C-FLAG-augurin), (Figure 47). Unfortunately, I could not detect the presence of

FLAG fusion proteins in supernatants and cell lysates after transfections of C-FLAG-augurin constructs into β-TC3 and RINm5f cells. I speculate that, unlike spexin, augurin needs an intact c-terminus for its stability.

Next, to confirm the existence of the putative cleavage site CS3, a mutant FLAG-augurin was engineered, where the predicted critical arginine residue at position 70 in CS3 was replaced with a glycine (mutant R70G), (Figure 48B). Parallel transfections of N-FLAG-augurin and N-FLAG-R67G-augurin into β-TC3 cells were carried out and both conditioned media and cell lysates were collected. FLAG-containing protein fragment from cell lysates were immunoprecipitated using M2 antibodies (cf. Materials and Methods), and larger volumes of supernatants were passed through a M2 anti-FLAG affinity purification column to isolate secreted FLAG fusion proteins. For both supernatants and cell lysates samples, the processing of the wt augurin construct (FLAG-augurin) differs from the mutant one (N-FLAG-R70G-augurin). A fourth low mobility band P1 (18-kDa), that may correspond to the full-length FLAG-tagged precursor, appears in both supernatants and cell lysate samples, while it is not present in wt samples. Furthermore, there is to be a difference in the relative intensity of the bands in wt and mutant samples: The lower mobility bands corresponding to P1 and P2 appear more intense in the mutant, in line with my hypothesis that R70 is used as a cleavage site. Those last two observations suggest that the arginine at position 70 is required for proper processing of augurin. Moreover, for the same blot as described in panel B, M1 anti-FLAG antibodies were binding with high affinity to the P3 and P4 proteins in the wt lanes but did not reveal any bands in the mutant samples (cf. high intensity band of Figure 48C, lane 1). Clearly, without its arginine at position 70, N-FLAG-augurin is not cleaved at the site CS3 to yield an N-terminal FLAG protein recognized by the M1 antibodies. However, the existence in mutant conditioned medium of two proteins which have the same mobility as the wt P3 and P4 proteins (Figure 48B) prompted me to speculate about the existence of a second cleavage site close to CS3 that could explain the similar processing patterns in the mutant and wt and "rescue" the processing of augurin at its n-terminus. An obvious candidate was CS2, which is situated just 3 residues upstream of CS3 (Figure 47). To test that hypothesis, I generated two mutants of arginine 67 of CS2, using in one case the N-FLAG construct as a template, and in the other case the mutant N-FLAG-R70G-augurin. In both cases I could not observe any expression of the FLAG constructs after probing with M2 antibodies, for at least two different transfections. It is tempting, although possibly premature, to speculate that the conserved arginine at position 67,which mark a strict boundary between the divergent and conserved

regions of augurin (Figure 46, Figure 47) plays a role in the proper processing of the CS3 site (a canonical Arg-X-(Lys/Arg)-Arg site), and thus the stability of augurin.

4.3.3 Conclusions on the secretion and processing of augurin

Augurin is secreted and processed by pancreatic β-TC3 endocrine cells as evidenced by the presence in N-FLAG-spexin transfected cells conditioned supernatants of multiple bands, including the putative processed forms P3 and P4. I have demonstrated, using the M1 anti-FLAG properties (specific binding to NH2-FLAG sequences), that N-FLAG augurin was cleaved exactly at the predicted cleavage site CS3. Furthermore, a mutation in the arginine at position 70 abolished cleavage at CS3, ruling out the possibility that the FLAG sequence had created a spurious cleavage site at CS3. The existence of a second cleavage site CS4' near the c-terminus of the augurin precursor was inferred from the observation of the double band P3/P4 (Figure 48). The experiments did not enable me to conclude on the existence of CS1 and CS2. It is possible that CS2 is used in the case of a mutation of CS3, as discussed in the previous subsection (Figure 48B, lane 2). The presence of a faint extra band near the protein P3 (Figure 48B, lane 2) corresponding to a FLAG-augurin without its signal peptide (Figure 49) could be explained by an alternative cleavage at CS1. Finally, based on near-perfect homology between all vertebrates of the c-terminal region of the augurin precursor (including the fishes) and cleavage patterns of FLAG-tagged constructs, I speculate that the c-terminus contains a second peptide NH_2-His-Gly-Ala-Ser-Val-Asn-Tyr-Asp-Asp-Tyr-COOH, released precisely after cleavage at the conserved arginine 138 (Figure 47) of the postulated site CS4'.

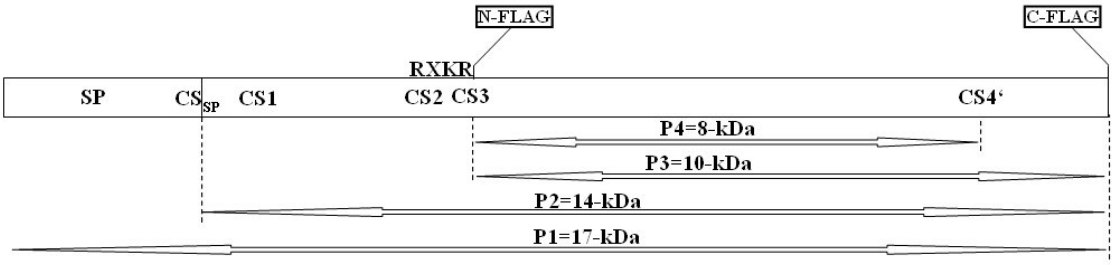

Figure 49: Model of augurin *in vitro* processing
This diagram shows the primary structure of the augurin protein and proposes a cleavage model that explains the banding patterns in the transfection/western blot experiments. The positions of the signal peptide cleavage site (CS_{SP}) and the dibasic putative cleavage sites CS1-3 are indicated. CS4' corresponds to a PC/Furin cleavage site inferred from the western blots experiments. Sizes shown in the diagram correspond to the protein fragments without their FLAG tags.

4.4 *In vitro* subcellular localization of augurin

Just as for spexin, immunocytochemistry with FLAG antibodies following transfection of FLAG- NPK and FLAG-augurin into a rat pancreatic cell line (RINm5f, or RIN) demonstrated colocalization of FLAG antigen with endogenous insulin in punctate intracellular bodies (Figure 37). This colocalization suggested that augurin, like neuropeptide K, underwent trafficking into dense core granules of the secretory pathway, a hallmark of PH. I speculate that augurin, much like spexin, is likely to be localized in dense core granules of secretory cells *in vivo*.

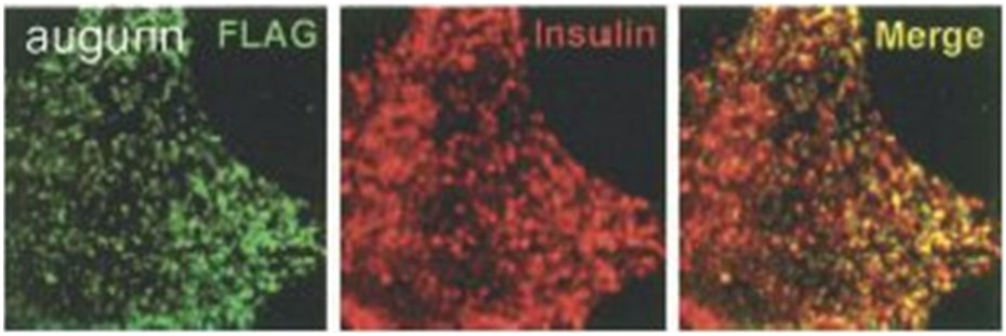

Figure 50:Colocalization of augurin with insulin in endocrine cells.
Flag-tagged augurin was transfected into RIN cells and fixed cells were subjected to double immunofluorescence with FLAG and insulin antibodies. FLAG-tagged augurin show colocalization with insulin in small, cytoplasmic punctate structures (cf. Figure 37: Colocalization of spexin with insulin in endocrine cells.).

4.5 Expression of augurin in mouse tissues

In this section, I present data on augurin mRNA localization in mouse tissues. I first characterize the spexin transcripts that were cloned from mouse tissues (subsection 4.5.1), then I present profiles of augurin mRNA expression by RT-PCR, Northern blot and Affymetrix microarray analyses in different mouse tissues (subsection 4.5.2). Finally, in subsection 4.5.3 I present data on the *in situ* localization of augurin mRNA in adult mouse endocrine organs, the adult mouse brain and late mouse embryo.

4.5.1 Cloning and characterization of the augurin mouse transcript.

Augurin-specific primers m99F1 (5'-caccatgagcacctcgtctgcg-3') and m99R2 (5'-tctgtgggcacctcagga-3') were used to amplify a 472 bp-long DNA band from reverse-transcribed RNA of adrenal glands tissues (Figure 51A). A band corresponding to the amplified fragment was then cloned into a TA vector and sequenced (cf. Materials and Methods). The sequence corresponded to the augurin transcript predicted from the databases (Ensembl and Refseq).

M99_transcript:

CACCCTGAGCACCTCGTCTGCGCGGCCTGCAGTCCTGGCCCTTGCCGGGCTGGCTCTGCTCCTTCTGC
TGTGCCTGGGTCCAGATGGCATAAGTGGAAACAAACTCAAGAAGATGCTCCAGAAACGAGAAGGACCT
GTCCCGTCAAAGACTAATGTAGCTGTAGCCGAGAACACAGCAAAGGAATTCCTAGGTGGCCTGAAGCG
TGCCAAACGACAGCTGTGGGACCGTACGCGGCCTGAGGTACAGCAGTGGTACCAGCAGTTCCTCTACA
TGGGCTTTGATGAGGCTAAATTTGAAGATGATGTCAACTATTGGCTAAACAGAAATCGAAACGGCCAT
GACTACTATGGTGACTACTACCAGCGTCATTATGATGAAGATGCGGCCATTGGTCCCCACAGCCGGGA
AAGCTTCAGGCATGGAGCCAGTGTCAACTATGATGACTATTAAGCTTCCTGAGGTGCCCACAGA

This fragment of DNA containing the entire augurin coding sequence was used as a probe to perform Northern blot experiments.

4.5.2 Expression profiles of augurin mRNA

4.5.2.1 mRNA expression profile of augurin by RT-PCR

An analysis of the mRNA expression of the augurin gene was performed in diverse tissues of the mouse, using the reverse-transcriptase polymerase chain reaction technique (RT-PCR). Total RNA was extracted from tissues using the Trizol reagent (Invitrogen, Carlsbad, CA., USA) and reverse-transcribed into cDNA. A 40 cycles PCR was then performed using the

specific augurin primers m99F1 and m99R2 to amplify the augurin transcript (cf. materials and methods). Augurin mRNA was present in most of the tissues surveyed (Figure 51A) including many classical endocrine tissues (adrenal gland, pituitary gland, testis, uterus, stomach and heart). It was not detected in the brain of young embryos.

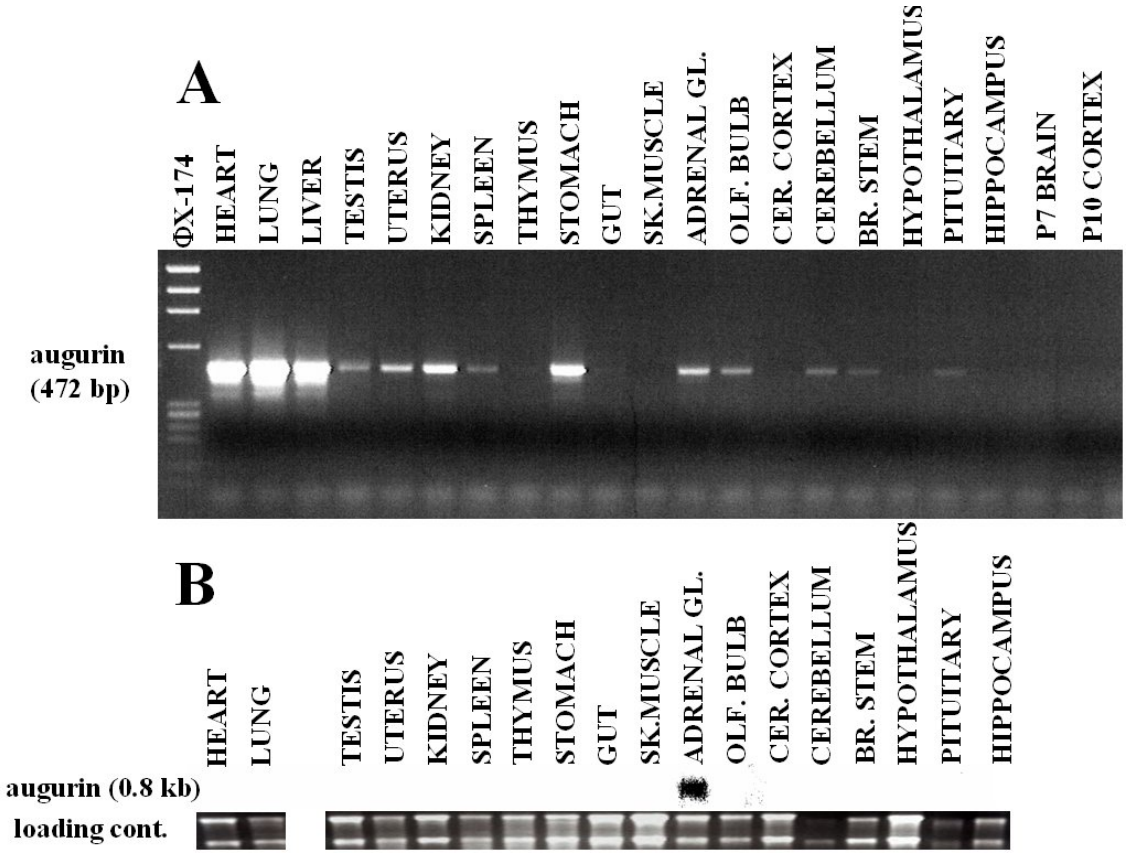

Figure 51: augurin mRNA expression by RT-PCR and Northern blotting.

Northern blot hybridization of augurin probe to a broad range of mouse tissues, including heart, lung, liver, testis, uterus, kidney, spleen, thymus, stomach, gut, skeletal muscle, adrenal glands, olfactory bulb, cerebral cortex, cerebellum, brain stem, hypothalamus, pituitary, hippocampus, P7 whole brain and P10 cerebral cortex.

(A) By RT-PCR, we see augurin expressed across a broad range of tissues in the mouse. Note the presence of augurin in classical endocrine tissues, such as adrenal gland, pituitary gland, testis, uterus, stomach and heart. Its expression could not be detected in the brain of young embryos.

(B) Northern blot of augurin tissues. I could detect a specific 0.8kb band only for adrenal gland tissues. Augurin in the mouse seems to be more highly expressed in the adrenal gland than any other tissues.

4.5.2.2 mRNA expression profile of augurin by Northern Blot

I next investigated augurin mRNA localisation in those same tissues using a more robust, albeit less sensitive method, the Northern blotting technique (cf. Materials and Methods). In the mouse, augurin is expressed in the adrenal glands at a higher level than in any other tissue

(Figure 51B). The presence of a single band in the Northern blot suggests that the augurin transcript is not alternatively spliced. I could estimate that the single endogenous augurin transcript size was around 0.8 kb.

4.5.2.3 Affymetrix microarray analysis

Publicly-available Affymetrix microarray data from the SymAtlas project confirmed, that, in the mouse, augurin is most highly expressed in the adrenal gland. Augurin expression profiles in human and mouse are not as highly correlated as one would expect. The olfactory epithelium/bulb and the trachea are among the tissues where augurin mRNA is found at similarly high levels in both human and mouse tissues. As a whole, we find augurin mRNA in endocrine tissues including adrenal gland, ovary, testis, uterus, pituitary and thyroid glands (Figure 52AC).

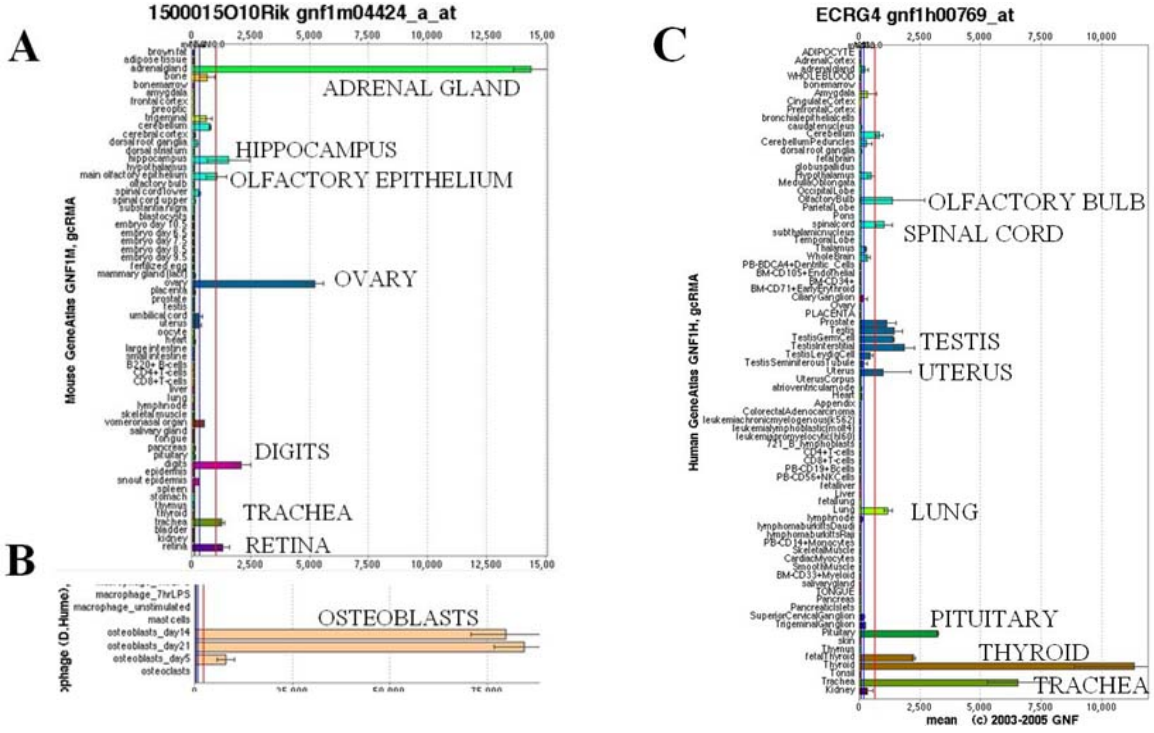

Figure 52:Augurin mRNA expression by Affymetrix microarray data [source: SymAtlas http://symatlas.gnf.org/SymAtlas/]

Augurin expression levels of mouse (AB) and human (C), using the Affymetrix microarray technology. The red line indicate a level of expression 10 times above that of the median level. Augurin is found highly expressed in the adrenal gland, ovary, hippocampus, digits, trachea, retina, and olfactory epithelium.

It is conspicuously expressed in reproduction tissues/organs, including ovary, testis, uterus, and prostate. The presence of augurin in human tissues at high levels in the pituitary and the thyroid do reflect what we see in the mouse, by *in situ* hybridization techniques (Figure 52C).

In addition a particularly high augurin mRNA expression is found in murine osteoblasts (days 5, 14, and 21), (Figure 52B) and human spinal cord (Figure 52C). The significance of augurin high expression in osteoblast cells will be discussed in section (4.7.4).

4.5.3 Expression of augurin mRNA by *in situ* hybridization

4.5.3.1 Augurin expression in adrenal and pituitary glands

I saw strong expression in the adrenal gland by microarray analysis and Northern blotting, and decided to start looking at the mRNA expression of augurin in the adrenal gland of mouse by *in situ* hybridization. Strong expression of augurin was detected in the choroid plexus (CP), which was confirmed by Allen Brain Atlas data, when it was made available. The presence of augurin could be detected in structures from classical endocrine glands of the mouse, including the glomerular layer of the adrenal cortex and the intermediate and anterior pituitary glands. It was also present in heart and CP (Figure 53). Furthermore, the radioactive augurin probe clearly hybridized to the thyroid gland of the mouse late embryo (E17), suggesting that its presence may also persist in the adult thyroid gland (Figure 58).

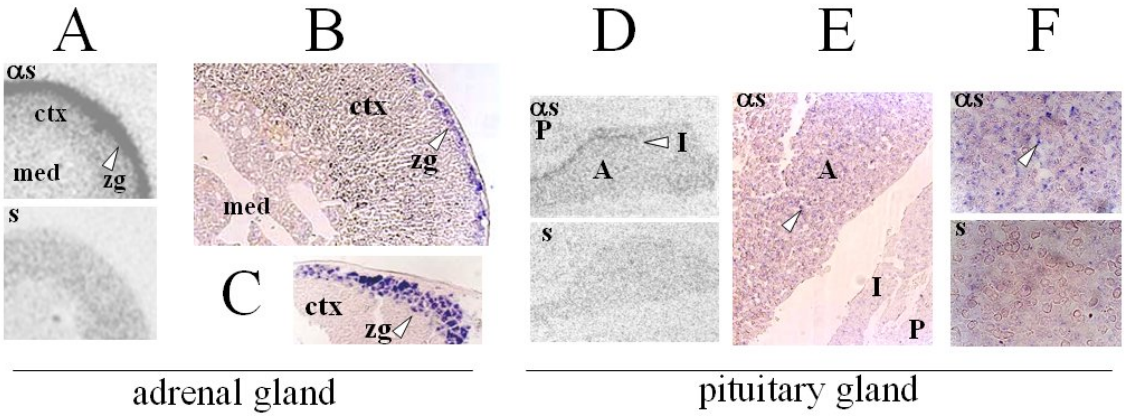

Figure 53: expression of mRNA augurin in the mouse adrenal pituitary glands
In situ hybridization with sense (s) and anti-sense (□s) was performed using ^{35}S (panels A and D) or digoxigenin (panels B, C, E and F) -labelled probes (cf. Materials and Methods). The anti-sense probe could detect the presence of augurin in the glomerular layer of the adrenal gland (A, B, and C), in the intermediate and anterior pituitary, (panels D, E and F). Panel C shows a section of adrenal gland cut deeper into the cortical layer than in panel B. Panel F shows stained anterior pituitary cells, at higher magnification. (Med, medulla; ctx, cortex; zg, zona glomerulosa of the adrenal gland; P, posterior pituitary (or neurohypophysis); I, intermediate pituitary (or pars intemedia); A, anterior pituitary (or adenohypophysis), s, sense, αs, anti-sense.

4.5.3.2 Augurin expression in the heart

Augurin mRNA is found in a restricted region of the heart (Figure 54). Detailed pictures generated from digoxigenin-labelled *in situ* hybridization experiments revealed a pattern of expression proximal to the aortic valve tissues (Figure 54), and suggested that it was likely to be involved in the cell biology or physiology of the heart valves.

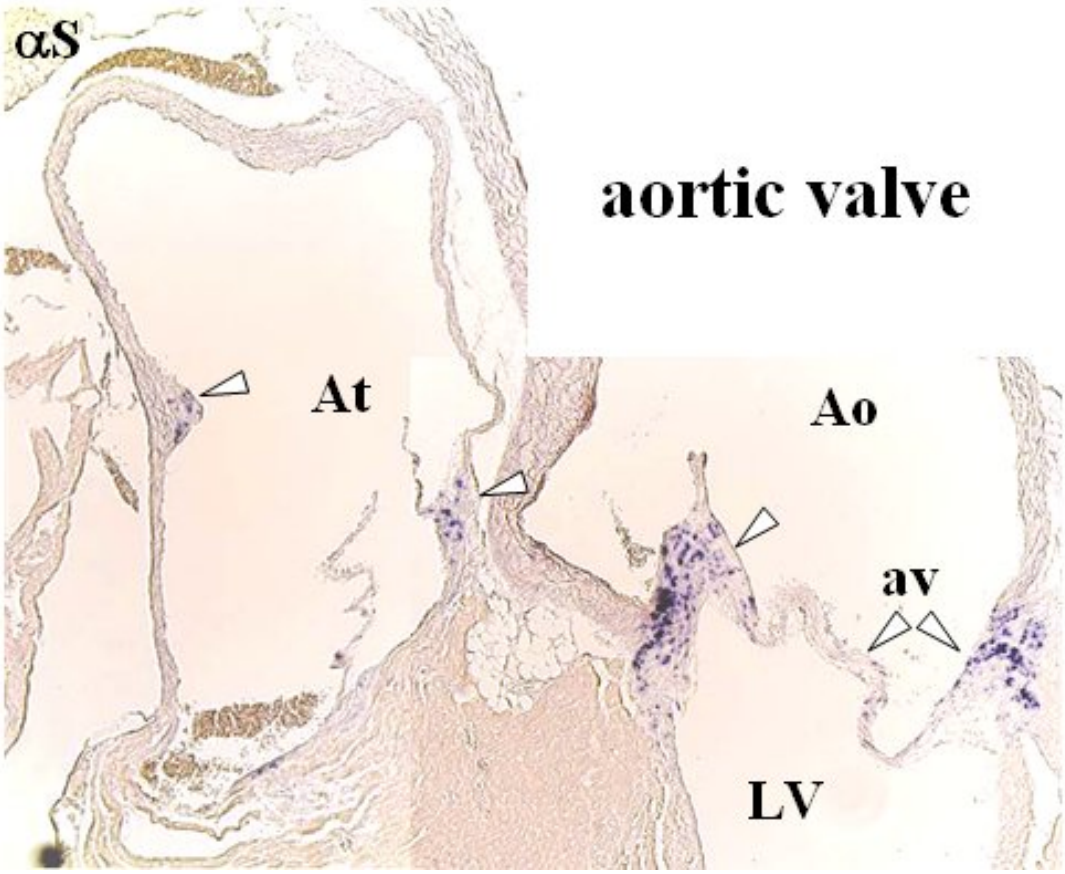

Figure 54:Augurin mRNA expression in the mouse heart

In situ hybridization with anti-sense (αs) probes was performed using digoxigenin-labelled probes (cf. Materials and Methods). The anti-sense probe could detect the presence of augurin (signal in blue) in a region proximal to the aortic valve (noted av). The aortic valve allows passage of blood only in one direction, from the left ventricle (LV) to the aorta (Ao). In some sections, the presence of augurin was also noted in cells at the periphery of the left atrium (At). Arrows indicate expression of augurin mRNA.

4.5.3.3 Augurin mRNA expression in the brain

Augurin was found in circumventricular organs (CVOs) of the brain, including the subcommissural organ (SCO) and the CP (Figure 55). Augurin mRNA was particularly abundant in CP. Note that the high augurin mRNA expression in the hippocampus and

cerebellum found in the microarray data is likely to reflect high augurin expression in choroidal structures (Figure 52A and Figure 55).

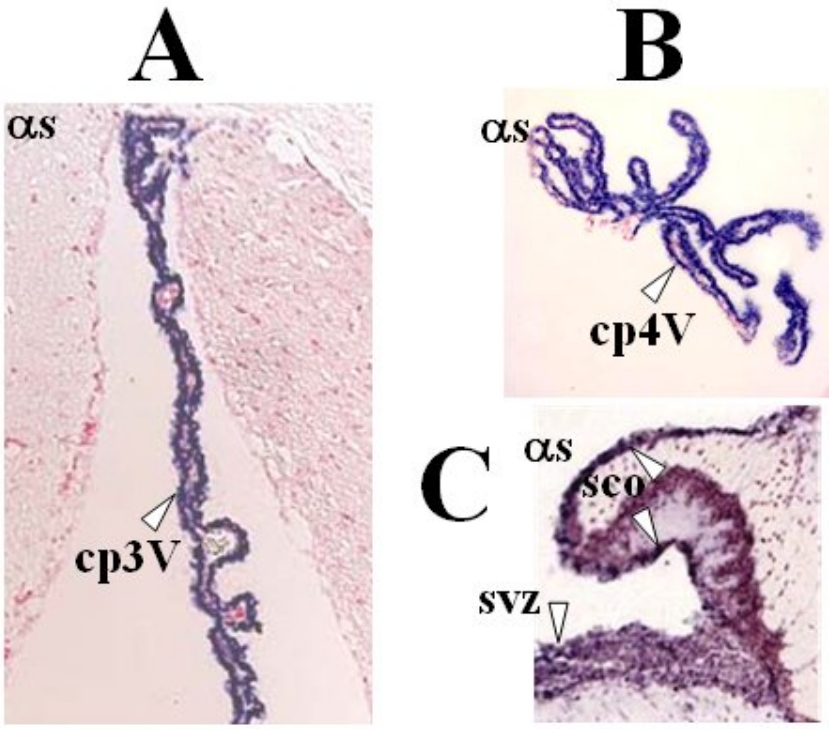

Figure 55: *in situ* hybridization of CP and the subcommissural organ (SCO).
Panels A and B demonstrate the presence of augurin mRNA in the CP of the 3rd (panel A) and 4th ventricle (panel B) of the brain. Panel C shows that augurin is expressed in the SCO, at the level of the RNA. In panels A and B, the augurin probe was labelled with digoxigenin, while in panel C the data was provided by the Allen Brain Atlas project.
(SCO, subcommissural organ; cp3V, CP of the 3rd ventricle; cp4V, CP of the 4th ventricle; svz, subventricular zone).

CVOs are structures found near the ventricles of the brain, and are outside the blood-brain barrier. Augurin was also found in cells lining the brain ventricles, which are likely to be ependymal cells. Ependymal cells are epithelial cells that line the ventricles of the brain and the spinal cord. They form a permeable barrier between the cerebrospinal fluid and the extracellular fluid that surrounds the cells of the CNS, and may serve as a filter between the CSF and brain ECF.

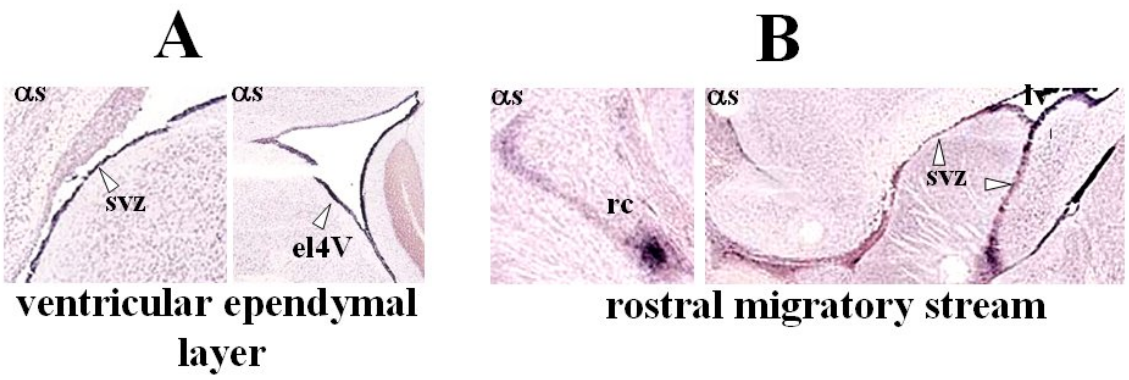

Figure 56: *in situ* hybridization of augurin mRNA [source: Allen Brain Atlas]
Augurin mRNA is present in layers of cells lining the ventricles (panel A). We show its presence in the lateral ventricle, and the 4th ventricle. It is conspicuously present in subventricular cells (SVZ) of the lateral ventricles forming the rostral migratory stream (RMS) (panel B). The RMS ends rostro-laterally in the olfactory bulb, in a region called the rhinocele (rc). All data were generated by the Allen Brain Atlas project. (αs, anti-sense; svz, subventricular zone; el4V, ependymal layer of the 4th ventricle; rc, rhinocele)

Augurin mRNA is strongly expressed in the ependymal layer of the subventricular zone of the lateral ventricles (SVZ), all along the RMS, (Figure 56B). Remarkably, neuronal precursors have been described to originate from the SVZ, and to migrate along a path called the rostral migratory stream (RMS, Figure 57) to repopulate cells of the olfactory bulb (Alvarez-Buylla and Garcia-Verdugo 2002).

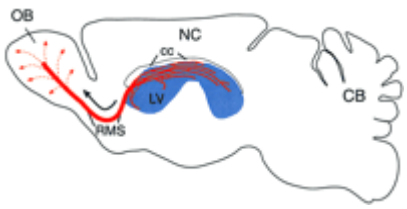

Figure 57:Schematic representation of the rostral migratory stream
RMS, rostral migratory stream; cc, corpus callosum; NC, neocortex; CB, cerebellum; LV, lateral ventricule; OB, olfactory bulb. This figure was extracted from (Alvarez-Buylla and Garcia-Verdugo 2002).

4.5.3.4 Augurin mRNA expression in the late embryo

We then looked for expression of augurin mRNA in the late mouse embryo (E17). In the embryo, augurin is expressed in the brain ventricles and CP, cartilage (all 7 costal cartilage segments), thyroid, adrenal and pituitary glands, heart, primordium of teeth, digits of front- and hind-limbs. Furthermore, it may be expressed in the peripheral system including the lower lumbar dorsal root, celiac, and trigeminal ganglions (data not shown).

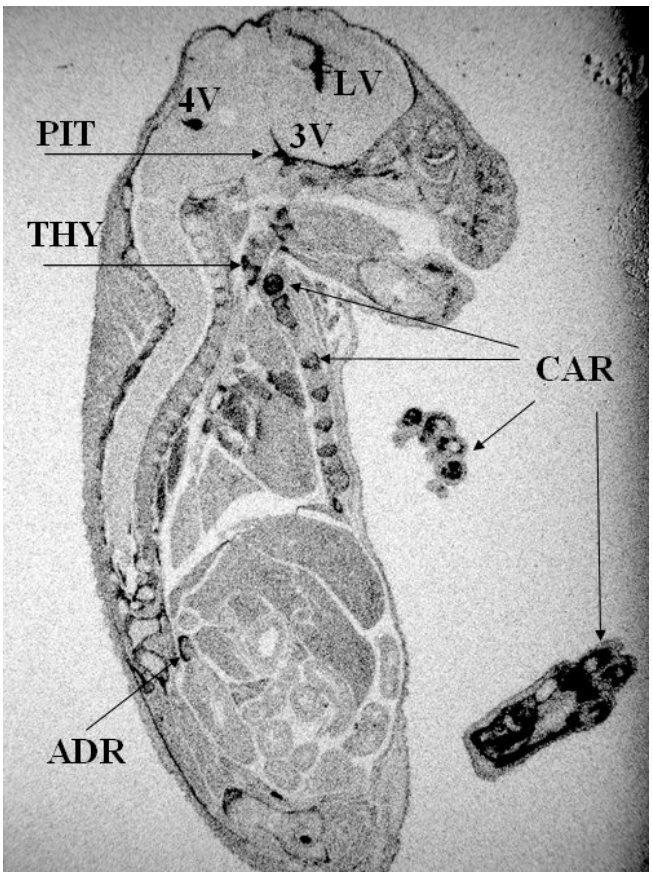

Figure 58: *in situ* hibridization of augurin transcript on a E17 embryo
The pattern of expression in the late embryo (E17) follows that found in the adult. In the embryo, augurin mRNA is clearly found in classic endocrine glands including adrenal (ADR), thyroid (THY), and possibly pituitary (PIT) glands. It is present in all three major ventricles of the brain (3rd, 4th, and lateral), and in numerous bony/cartilagenous structures, including the hyoid, coastal and digital bones of both front- and hind-limbs (CAR). (LV, lateral ventricule; 3V and 4V, 3rd and 4th ventricules; THY, Thyroid gland; PIT, pituitary gland; ADR, adrenal gland; CAR, cartilage primordia).

4.6 Localisation of augurin protein in mouse tissues

To study mature augurin protein localisation in the mouse, we had rabbit polyclonal antibodies raised against the conserved core antigenic sequence of augurin precursor protein (Primm, Milan, Italy), NH_2-Gln-Leu-Trp-Asp-Arg-Thr-Arg-Pro-Glu-Val-Gln-Gln-Trp-Tyr-Gln-Gln-Phe-Leu-Tyr-Met-Gly-Phe-Asp-Glu-Ala-Lys-Phe-Glu-Asp-Asp-COOH.

Immunohistochemistry done on mouse tissue using that antibody (called aug) revealed a pattern of expression which corroborated both our and the Allen Brain Atlas *in situ* hybridization data on augurin mRNA localisation. This suggested that the rabbit antiserum was indeed recognizing the augurin antigenic sequence. Structures where both mRNA and protein appeared to be expressed included the ZG of the adrenals, the adenohypophysis, subcommissural organ, CP, and ependymal linings of the ventricles (Figure 59ABCDFG).

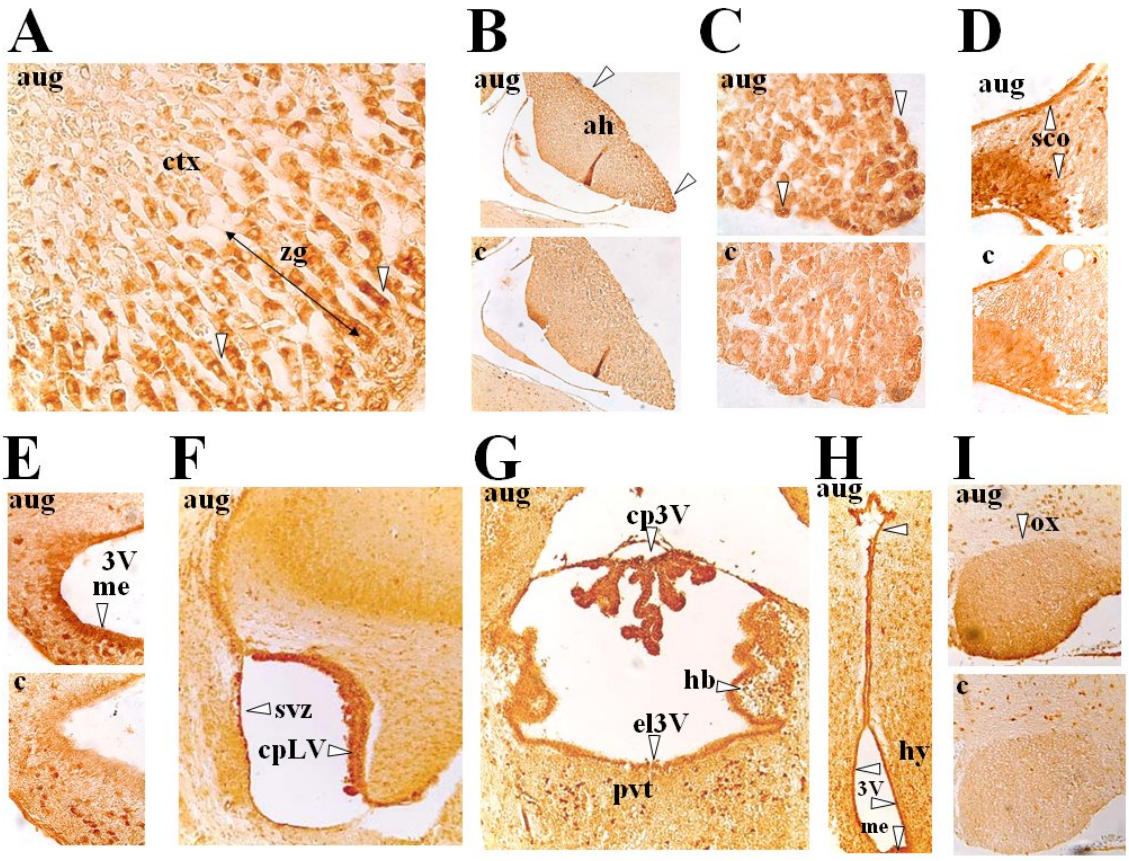

Figure 59: anti-augurin immunohistochemistry of mouse tissues

We show anti-augurin immunoreactivity in adrenal, pituitary glands and brain tissues. Sectioned mouse tissues show immunoreactivity in regions where augurin mRNA was found, including the ZG of the adrenal gland (panel A), the pituitary gland (panel B and C), the circumventricular structures SCO (panel D), CP (panels F and G), and the linings of the 3rd ventricle (panel F, G, and H). Augurin immunoreactivity was also found in the median eminence (E and H), and the optic chiasm (I). Panel B shows a high magnification of ZG adrenal gland staining. A control staining (c), where the augurin antiserum was omitted from the procedure, was performed for panels B, C, D, E and I to demonstrate staining specificity. Panels B, C, D, E and I correspond to brain sagittal sections, while panels F, G and H correspond to brain coronal sections. Arrows point to cells which are augurin-immunoreactive. (ctx, adrenal cortex; zg, zona glomerulosa; ah, adenohypophysis; cpLV, CP of the lateral ventricle; sco, subcommissural organ; 3V, 3rd ventricle; me, median eminence; svz, subventricular zone; el3V, ependymal layer of the 3rd ventricle; hb, habenula, pvt, paraventricular thalamus; hy, hypothalamus; ox, optic chiasm)

Furthermore, augurin immunoreactivity was found in the median eminence (Figure 59EH), and the optic chiasm (Figure 59I). A low expression of augurin mRNA was also noticed in the optic chiasm when I searched specifically for that structure (data not shown) in the Allen Brain Atlas. Additionally, augurin is expressed in the habenular commissure, both at the level of the RNA and protein (data not shown).

4.7 Possible physiological function(s) of augurin

In the following section I discuss the possible function(s) of augurin based largely on our *in vivo* expression data. I assume that augurin undergoes regulated secretion and processing *in vivo*, as suggested by the *in vitro* studies (3.3, 3.4, 4.3 and 4.4), and that it is thus likely to be a secreted signalling molecule. Several studies have showcased the possibility to infer function from anatomical considerations of *in vivo* expression data (Peyron et al. 1998; Vrang et al. 1999). In the first study *in situ* hybridization and immunohistochemistry for the Orexin/Hypocretin gene on brain sections was used to successfully derive predictions for its functions. Remarkably, those predictions all proved to be correct, as demonstrated by subsequent studies, including the hypocretins involvement in feeding behaviour (Sakurai et al. 1998b), blood pressure regulation and cardiovascular function (Kayaba et al. 2003; Smith et al. 2007), neuroendocrine regulation (Horvath et al. 1999; van den Pol et al. 1998), thermoregulation (Szekely et al. 2002; Yoshimichi et al. 2001), and sleep/wake states and the circadian system (Chemelli et al. 1999; Lin et al. 1999). This study demonstrated the power of brain expression studies for the prediction of function and prompted me to propose possible functions for augurin and its involvement in common pathological condition, based on its patterns of expression in human and mouse. In sharp contrast to spexin, augurin expression encompasses a wide range of tissues, including the CP, ependymal layer lining the ventricles of the brain, and the epithelia of several endocrine organs, including the anterior pituitary, adrenal, and thyroid glands. One can form three groups of proteins with hormonal properties based on the tissue-specificity of their expression.

The first group is formed by hormones and PH which are generally highly expressed in a very restricted number of tissues (typically 1 to 3). This group includes secretin (intestine, placenta, dendritic cells in the human), insulin (pancreatic islets), glucagon (pancreatic islets and intestine in the mouse), releasing factors TRH (hypothalamus, pancreas) and CRH (hypothalamus and placenta), tropins ACTH/POMC (pituitary) and somatotropin (pituitary, trigeminal, bone, placenta in human); parathyroid hormone (thyroid and parathyroid glands), oxytocin (hypothalamus), vasopressin (hypothalamus, preoptic area in the mouse), ghrelin (stomach), calcitonin (thyroid, CNS).

The second group is formed by classical neuropeptides which generally have expression restricted to neuronal nuclei of the brain and a few peripheral tissues. Those classical neuropeptides include substance P/tachykinins (CNS/PNS), CART (CNS, retina), orexin

(CNS, mainly hypothalamus, and preoptic area), GRP (CNS, stomach, intestine), cholecystokinin (CNS, prostate, placenta, intestine, mammary gland).

The third group encompasses a large group of growth factors and developmental secreted proteins which are expressed across a much broader range of tissues, as they regulate fundamental aspects of the biology of cells of the same type and can modulate the morphology and physiology of entire organs/tissues (Musaro et al. 2001). Those include growth factors such as IGF1, VEGF, TGF-β1, and BMP-1 that are generally not circulating and act in a paracrine/autocrine fashion.

Spexin clearly relates more to the second group (neuropeptides) while augurin is more likely to belong to the third group (growth factors).

4.7.1 Possible role for augurin in brain barriers maintenance

At the level of the brain, endothelial cells that ensheath capillaries are bound to each other by tight junctions that prevent large blood-borne molecules from seeping through to the brain extracellular fluid (ECF). This impediment, called the "blood-brain barrier" (BBB) protects the brain from circulating viruses, bacteria, inflammatory molecules and other potentially noxious or harmful molecules. However, in a few organs and tissues of the brain bordering the third and fourth ventricles, the endothelial cells surrounding the capillaries form fenestrae, allowing easier exchange of molecules between the plasma and the outside (Groothuis and Levy 1997). For that reason, those circumventricular organs (CVOs) are considered to be "outside" the BBB. Traditionally, CVOs include the subfornical organ (SFO), CP, posterior pituitary, organum vasculosum of the lamina terminalis (OVLT), area postrema (AP), and median eminence (ME). subcommissural organ (SCO), situated below the posterior commissure, at the cerebral aqueduct Sylvian duct, is traditionally considered to be a CVO, for its anatomical (they are periventricular) and histological (they contain secretory ependymal cells) properties. However, the SCO does not have fenestrated capillaries, and is not as permeable as the other CVOs; and for that reason, it is not considered to be outside of the BBB. Because they are outside the BBB, some researchers also include the medial pineal and intermediate pituitary glands among the CVOs, although strictly speaking those are glands and are not part of the brain.

CVOs are generally categorized into sensory and secretory organs. There is a general consensus on classifying the SFO, AP, and OVLT as sensory organs, since they receive

signals from peripheral organs that are integrated in the brain ("the windows of the brain", (Johnson and Gross 1993)). CP, ME and pituitary lobes are categorized as secretory organs (Cottrell and Ferguson 2004) while the SCO is considered to be both sensory and secretory. It is noteworthy that augurin is expressed predominantly in secretory CVOs, including the median eminence, CP, subcommissural organ, and intermediate pituitary. Augurin is also expressed in ependymal cells bordering the ventricles, which are similar to cells of the CVOs in that they constitute a physical barrier separating the ECF (of either the CVOs or the brain) from the CSF (Figure 60). In order to bypass the BBB and reach target cells inside the brain, blood-borne molecules can diffuse through two layers of augurin-expressing cells, either by passing between the cells, or traversing them (Groothuis and Levy 1997).

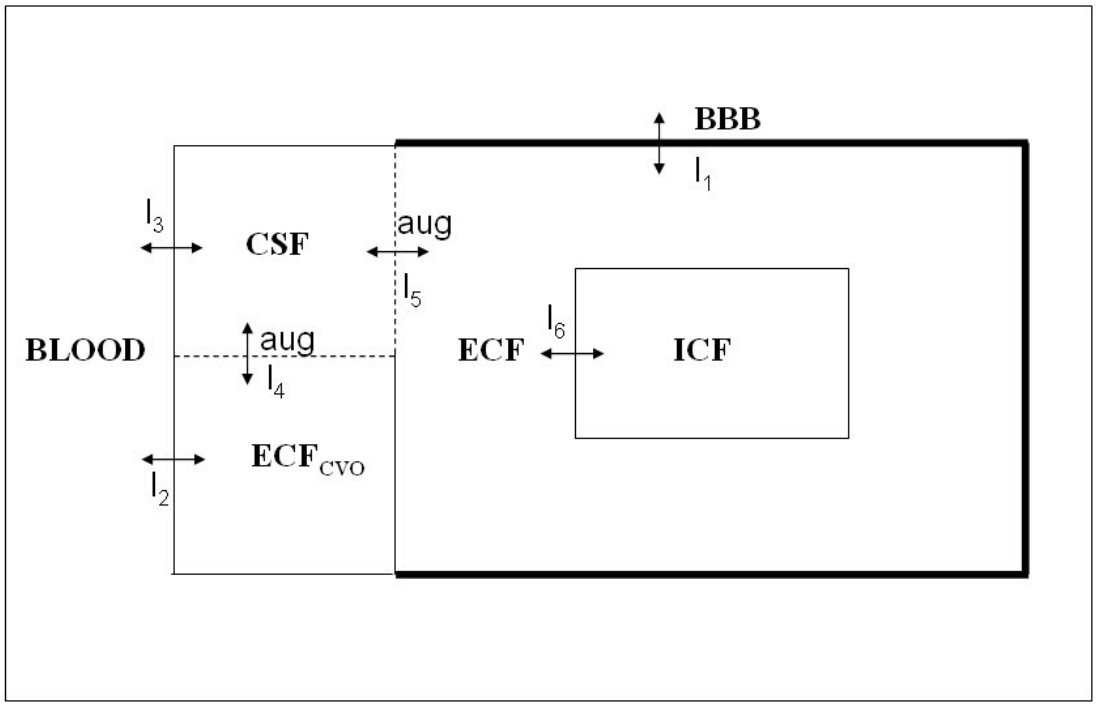

Figure 60: connectivity between the major body fluids in the brain

I_1, I_2, I_3, I_4, I_5, and I_6 denote the interfaces between compartments. The interface between the blood and the extracellular compartment (ECF) defines the blood-brain-barrier (BBB), and the bold line renders the idea that it is relatively impermeable. Both interfaces I_4 and I_5 illustrated by dotted lines, represent augurin-expressing ependymal cells (aug). I_4 is the interface between the extracellular compartment of augurin-expressing circumventricular organs (median eminence, CP, and subcommissural organ) and the cerebrospinal fluid. I_5 is the interface between the cerebrospinal fluid and the extracellular fluid of the brain; and it is materialised by augurin-expressing ependymal cells lining the ventricles. In order to reach the intercellular space of specific cells in the brain, a messenger (hormone, drug, neurotransmitter) must cross the BBB (using specific transporters, for instance), or transit through the cerebrospinal fluid, by crossing boundaries which are much more permissive, including I_4 and I_5. CSF, cerebrospinal fluid; ECF, extracellular fluid (or interstitial fluid); ICF, intracellular fluid; ECF_{CVO}, extracellular fluid of circumventricular organs, including CP.

Given augurin expression in epithelial cells of the CVOs (including the CP) and ependymal cells lining the ventricles, I speculate that augurin could play a role in the specification/maintenance of those two types of "filter cells". I noted that augurin is also expressed in the human testis and the mouse retina (Figure 52); and both organs possess layers of specialised epithelial cells to selectively isolate the tissues from blood-borne molecules (Cumha-Vaz 1978; Holash et al. 1993). In the retina, epithelial cells at the border between the ocular fluids possess ciliae, analogously to ependymal cells from the ventricule linings, SCO, and CP of the brain. Interestingly, in (Cumha-Vaz 1978), the authors suggest

that, similarly to what choroid plexuses do, ciliary processes located between the anterior and posterior part of the eye may play a fundamental role in the regulation of all intraocular fluids (Cumha-Vaz 1978). In the brain ventricular system the beating ciliary processes growing out of ependymal cells from ventricle linings, SCO or CP, facilitate CSF circulation and are required for its homeostasis. In particular, the SCO plays a major role in the maintenance of laminar flow through the cerebral aqueduct, and defects associated to it lead to major health problems, such as stenosis (Pohle et al. 2001). Rfx3 is a gene which has been associated with the regulation of ciliary activity and morphogenesis, and its disruption causes hydrocephaly (Baas et al. 2006). It is perhaps not anecdotal that this gene is expressed in the same structures of the brain than augurin, including the CP, the SCO, linings of the ventricles, and the dorsal part of the median eminence (Allen Brain Atlas). Since both CP and cells lining the ventricles are ciliated cells, it is possible that augurin plays a role in the morphogenesis, or physiology of the ciliae (e.g. the dimensions of the ciliae and their beating frequency).

4.7.2 Possible link between augurin and the renin-angiotensin-aldosterone system

One of the major functions associated to CVOs are the regulation of blood pressure, body fluid regulation, and thirst (Fitzsimons 1998). Especially sensory CVOs (area postrema, subfornical organ, and OVLT) have been associated with the regulation of these autonomic functions (Cottrell and Ferguson 2004). The prominent expression of augurin in medial ventricular structures, including the secretory CVOs (CP, SCO, and median eminence) lead me to hypothesize that it is involved in these homeostatic mechanisms, as a central relay effector molecule conveying its message to the rest of the brain through the CSF (from the SCO, CP, ventricular ependymal cells) or to the periphery (from the pituitary or median eminence). The main regulator of tissue fluid osmolarity and blood pressure in the body is the renin-angiotensin-aldosterone system (RAAS). This circuit involves hormones from the kidney (angiotensins) that are converted by an enzyme (renin) into an active PH, angiotensin I (further cleaved by ACE to yield the more potent angiotensin II) that constrict arteriole smooth muscle and acts on the adrenal cortex to up-regulate the release of the mineralocorticoid aldosterone. The main action of aldosterone is ultimately to increase water retention by cells in the kidney, one of the mechanisms to increase systemic blood pressure. The natriuretic PH (ANP, BNP, and CNP) are PH that antagonise the effects of RAAS activation, and lower blood pressure. ANP acts at different levels to lower blood pressure: it has a direct natriuretic effect (antagonising aldosterone's water retention action), renders the

adrenal ZG cells less responsive to signals triggering secretion of aldosterone (renin), and inhibits secretion of neurohypophysial vasopressin, the major vasoconstrictor/anti-natriuretic PH of the RAAS. ANP binds strongly to sensory CVOs including the AP, and SFO (Gibson et al. 1986). A large body of work has directly implicated CVOs in water and salt homeostasis, particularly the AP (Miselis et al. 1984). It seems clear that sensory CVOs are major regulators of the RAAS.

Several players of the RAAS are expressed in organs and structures in the brain and organs of the body where augurin is expressed. First, adrenal ZG cells synthesize both the essential steroid hormone aldosterone and high quantities of augurin. Angiotensin 2 receptors are found in very few structures of the brain including the SCO (Geerling et al. 2006), where augurin is expressed. ANPR1, an antagonist of the RAAS, is expressed both in the adrenal ZG and CP. It is expressed also conspicuously in the medial habenula, SFO, Purkinje cells of the cerebellum, but not in the SCO. Interestingly, it is also present around the rhinocele, in the mitral cell layer of the olfactory bulb and in periglomerular migrating neurons of the olfactory bulb (Baker et al. 2001; Carlen et al. 2002) (data from the Allen Brain Atlas); and the ANP system could be associated to migration and differentiation of young neurons, just as I predict for augurin. Renin, the enzyme responsible for cleavage of angiotensinogen precursor into angiotensin I, does not seem to be expressed in the brain in substantial amounts. Angiotensinogen is expressed in the brain but does not show a similar expression as augurin, and is not highly expressed in secretory circumventricular organs. ACE, the enzyme which cleaves angiotensin I and yield the more potent angiotensin II, is highly expressed in CP, and pituitary gland (ABA), similarly to augurin. It has been demonstrated that angiotensin II does not cross the BBB (Paul et al. 2006). Presumably, the known hypertensive action of angiotensin II in the brain is partly mediated through sensory CVOs (OVLT, SFO and AP, SCO). It either acts indirectly on those organs, through the action of relay molecules and neurons, or by diffusing into the CSF, where considerable amounts of it has been detected (Paul et al. 2006). Angiotensins in the brain play a prominent role in increasing blood pressure, as transgenic mice inhibiting the action of the RAAS are hypotensive, and higher numbers of angiotensin receptors were found in hypertensive rats (Paul et al. 2006). Angiotensins have been shown to have a prominent role in feeding and drinking behaviour, maintenance of the BBB, and reproduction (Paul et al. 2006). In (Rodriguez et al. 1984b), the authors show a clear spatial and structural relationship between ependymal cells and blood vessels in the SCO, perhaps reminiscent of fractones (Mercier et al. 2002). This evokes the

possibility that augurin is a molecule released by secretory ependymal cells that act directly on endothelial cells of the capillaries to alter blood pressure, or involved in the organisation of the basal laminae and a vascular niche (Nikolova et al. 2007) (this idea is developed further in the next subsections). Lastly, calcification in the aortic valve involves several players of the RAAS (Mohler 2004), again suggesting a link between this system and augurin.

4.7.3 Possible function for augurin in the subcommissural organ

The subcommissural organ (SCO) is a secretory and sensory organ located at the entrance of the Sylvian aqueduct. It is a phylogenetically ancient structure, present in all vertebrates (Sterba et al. 1967), including stickleback fish (Olsson 1959), amphibians, reptiles, birds, and mammals (Rodriguez et al. 1984a), and is one of the first structure of the brain to differentiate (Rodriguez et al. 1998); However to this day, no definite function has been established for this ancient organ. Nevertheless, several roles have been proposed, including regulation of blood pressure, thirst, salt and water balance (Dundore et al. 1984; Gilbert 1963). There is substantial evidence that it modulates the laminar flow of the CSF by secreting a SCO-spondin, a glycoprotein which aggregates to form a structure known as Reissner's fibre (Gobron et al. 1996; Rodriguez et al. 1998; Sterba et al. 1982), and by influencing the activity of ependymal ciliae (Meiniel 2007). Although some studies performed in the 1960s (Foeldvari and Palkovits 1964; Gilbert 1963; Palkovits et al. 1965) and 1980s (Dundore et al. 1987; Dundore et al. 1984; Geerling et al. 2006; Ghiani et al. 1988), have implicated it in the regulation of the renin-angiotensin-aldosterone system (RAAS), its exact role and physiological significance in those processes is still largely unknown (Ganong 2006; Rodriguez et al. 1998). The SCO has the unique feature of being both a sensory and a secretory organ, which means it can be considered as a hub of communication between the brain and peripheral organs. It expresses several receptors, including the leptin receptor (Cecilia et al. 2006), estrogen receptor alpha (Cecilia et al. 2006), and angiotensin II receptor (Ghiani et al. 1988). Recently, (Cecilia et al. 2006) suggested several novel functions for the SCO in reproduction and energy metabolism. In this paper, the authors argue that the ependymal cells from the SCO and neighbouring astrocytes respond to gonadal and nutritional signals from circulating leptin and estrogens by mechanisms which are yet to be discovered. Surprisingly, I found little recent literature on the subcommissural organ. In early physiology-oriented studies of the SCO, that took place in the 1950s and 1960s, Gilbert postulated that the SCO regulates body water volume. In particular, he demonstrated that

subcutaneous or intraperitoneal injection of SCO extracts caused natriuresis in dogs (Gilbert 1963). However, I also found contemporary reports challenging that relationship. In (Upton et al. 1961) the authors wrote: “Our evidence indicates that the SCO plays no role in water metabolism through control of thirst or through control of aldosterone secretion”. However the suspicion that the SCO plays a critical part in body fluid and salt balance still holds today, although little is known about the molecules and mechanisms involved.

There are reports of a cross-talk between the SCO and the adrenal cortex (Dundore et al. 1987; Palkovits et al. 1965). This is of particular interest to us since augurin is expressed in both structures. Dundore and colleagues showed that injection of aldosterone directly in the SCO increased urinary sodium excretion, decreased adrenal medulla size, and elevated adrenal corticosteroids concentration (Dundore et al. 1984). Conversely, (Palkovits et al. 1965) observed a decrease in size and activity of the adrenal ZG after lesions of the SCO, which strongly suggested that the SCO had a trophic effect on the adrenal cortical ZG and medulla. Early on, Gilbert postulated that the SCO secretes a “natriuretic principle” causing natriuresis (Gilbert 1963). I have looked in the databases and in the literature (Allen Brain Atlas and BrainInfo database at http://braininfo.rprc.washington.edu/) for accounts of natriuretic peptides in the brain (ANP, BNP, and CNP) and found no compelling evidence of their presence in the SCO. The fact that there seems to be a cross-talk between the ZG of the adrenal cortex and the SCO through aldosterone, and that molecules produced in the SCO have both a natriuretic action on the body and a trophic effect on the adrenal gland, argues for a possible function of augurin in the central regulation of the RAAS. One may also speculate that elucidation of augurin function in the SCO will aid in understanding the biology of the subcommissural organ.

4.7.4 Possible role of augurin in bone homeostasis

Augurin is expressed at very high levels in osteoblasts, at moderate levels in bones and digits (Figure 52), and in most of the bone/cartilage structures in the late embryo (Figure 58); it is not expressed at all in the osteoclasts. In brain and heart I noted that augurin is expressed in the aortic valve and the linings of the brain ventricles, both structures which are subjected to particular mechanical constraints, and as a result are required to secrete structural proteins such as collagens and elastin to fulfil their function (Vesely 1998). Collagen-1 is found in the lateral ventricle SVZ and is likely to be secreted by fibroblasts or macrophages (Mercier et al. 2002). Collagens forming the basal lamina that are secreted by osteoblast-like cells are

essential for the function of the blood-brain barrier (Godeau and Robert 1979). Basal lamina is a specialized form of ECM that is present in most peripheral organs in close association with epithelial cells. It is a sort of scaffold that supports connectivity between cells and the blood supply, provides mechanical stability and can serve as a barrier between different cell types (Urabe et al. 2002). It can also mediate and promote specific cell-cell and cell-molecule interactions (Nikolova et al. 2007). The basal lamina has been described as an arrangement of collagen and heparin-sulfate glycoproteins (HSPGs) that is thought to provide a solid support for stem cells attachment, sequester growth factors and promote their binding to progenitor cells (Kerever et al. 2007). Osteoblasts and ependymal cells are known to contribute to the permissive environment of stem cells in the CNS and bone marrow by secreting essential growth factors and supporting structures such as basal laminae (Fuchs et al. 2004). Hematopoietic stem cells (HSC) rely on osteoblasts lining the inner cavity of the bone to retain their undifferentiated and slowly proliferating state (Fuchs et al. 2004). Recently, ECM structures termed fractones (for their fractal organization) have been described alongside basal laminae in some of the brain neurogenic zones, including the SVZ (Mercier et al. 2002), and along the walls of the ventral third ventricle of the hypothalamus (Mercier et al. 2003), both regions of high augurin expression. Older papers describe similar structures in the SCO, that connect ependymal secretory cells and the vasculature (Rodriguez et al. 1984b). Interestingly, progenitor cells of the germinal hippocampal SGZ, unlike the SVZ, are in direct contact with endothelial cells of capillaries (Nikolova et al. 2007).

Physiological calcification seems to occur predominantly in regions of the body which require rigidity, and a resistance to mechanical stress, such as bone (Luo et al. 1997) and probably the brain ventricles, heart valves and trachea. However, the process of calcification is believed to be ubiquitous. It can become pathological when in excess; the tissue then interferes with its mechanical properties as is the case with the aortic valve (Boon et al. 1997).

Aortic valve calcification is known to be a direct determinant of aortic valve stenosis, and a major cardiovascular risk factor (Pohle et al. 2001). Aortic valve stenosis is the narrowing, or obstruction of the aortic valve which alters the blood flow between the aorta and the left ventricle and is the main cause of aortic valve failure. Augurin is likely to be a secreted protein expressed at very high levels in osteoblasts, and a number of epithelial tissue calcification diseases have been associated with an osteoblast-like phenotype, including

aortic valve calcification (Rajamannan et al. 2003). In (Rajamannan et al. 2003), the authors draw a parallel between aortic valve calcification in humans and bone formation, and present data supporting that an osteoblast-like phenotype underlies this process. In hearts where calcification is high, typical osteoblasts markers such as osteocalcin, cbfa-1 (Runx2), and osteopontin, are clearly up-regulated. Incidentally, I found that osteopontin had the same mRNA expression pattern in the olfactory bulb as atrial natriuretic peptide receptor 1 (Allen Brain Atlas, mitral cell layer, and scattered neurons in the main olfactory bulb, towards the glomeruli). Large amounts of calcification associated with deposit of collagen whorls have been described in the CP (Alcolado et al. 1986), where large quantities of augurin are produced. Cells in the CP express specific osteoblast markers, including the E11 antigen (podoplanin) (Wetterwald et al. 1996). Interestingly, podoplanin, a type I membrane protein, is expressed in the CNS in the medial habenula, ependymal cells of the CP, subcommissural organ, olfactory bulb, the RMS, the ependymal cells lining the ventricles, and meningeal cells (Allen Brain Atlas and (Schacht et al. 2005)). Podoplanin is specifically expressed in lymphatic endothelial cells (Breiteneder-Geleff et al. 1999) whose functions (clearance of proteins of the interstitial/intracellular fluid, trafficking of immune cells) are critically regulated by the ECM (Pepper and Skobe 2003). Podoplanin has been associated with tumorigenesis of several squamous carcinoma (Schacht et al. 2005) and since no endogenous ligand has been identified, podoplanin could be a candidate receptor for augurin.

Calcification correlates with sclerosis in the heart and cancers of many tissues. Psammona bodies, which refer to nucleated balls of calcium, are found in many common epithelial cancers, such as thyroid, kidney, ovaries, and meninges carcinoma. I have found several reports claiming that a high degree of calcification is correlated with cancer occurrence or malignancy (Alcolado et al. 1986; Khoo et al. 2002; Wienke et al. 2003). While calcification is generally dismissed as side-effect of cancer and merely used as a diagnosis tool, there is growing evidence that calcification might play a role in the development of cancer. I found reports of pathological calcification in carcinomas corresponding to virtually all tissues where augurin is expressed, including the CP (Caputo et al. 1998), pineal and habenular commissures (Macpherson and Matheson 1979), meninges, pineal gland (Alcolado et al. 1986), brain ventricles (Warmuth-Metz et al. 1999), retina (Bullock et al. 1977), anterior pituitary gland (Groisman et al. 1999), thyroid gland (Khoo et al. 2002; Ozaki et al. 1990), and adrenal gland (Ozmen et al. 1992).

Another link between bone morphogenesis, ectopic calcification and maintenance of stem cells/cancer is bone morphogenetic protein 2 (BMP-2). BMP-2 is one of the main factors responsible for setting bone/cartilage homeostasis and together with its antagonist noggin for driving bone and cartilage formation in development and throughout adult life. BMPs have been involved in the maintenance of neuronal SVZ precursor cells (Alvarez-Buylla and Garcia-Verdugo 2002; Temple 2001) and commitment to astroglial lineages (Gross et al. 1996). Noggin, an antagonist of BMPs, participates in both normal neurogenesis and bone/cartilage formation (Lim et al. 2000). In mice, overexpression of noggin impairs the formation of bone, antagonises the differentiation of cells into osteoblasts and causes osteoporosis-like defects in bone (Wu et al. 2003). Finally, high expression of BMP-2 has been described as a hallmark of ectopic calcification (Collett and Canfield 2005). These similarities in expression between augurin and bone morphogenetic proteins lend support to the speculation that augurin plays a role in bone homestasis and calcification.

4.7.5 Possible role of augurin in stem cell niche maintenance

Augurin expression in both the brain SVZ and adrenal ZG suggests its possible implication in processes regulating cell fate. The ZG contains cells which can proliferate and differentiate into fasciculata cells (Connell and Davies 2005; Vinson 2003) in order to replace old cells (Kataoka et al. 1996) and are able to entirely repopulate the adrenal cortex after partial adrenalectomy (Ganong 2006). As for the SVZ, it is along with the hippocampal dentate gyrus the main site of neurogenesis in the mammalian brain (Alvarez-Buylla and Garcia-Verdugo 2002; Kirschenbaum et al. 1999; Yamashita et al. 2006). Neuronal precursors have been identified in the SVZ that are able to migrate along the rostral migratory stream (RMS) and repopulate the olfactory bulb (Kirschenbaum et al. 1999). Recently, (Yamashita et al. 2006) showed that these cells can also repopulate the striatum after a stroke injury. The discovery of neuronal precursors in this region of the brain has generated considerable attention because of its possible therapeutic implications. Many groups have since sought to uncover the regulatory mechanisms underlying neurogenesis from cells of the SVZ. A large group of growth factors have been shown to regulate the proliferation of those cells *in vitro* and *in vivo*. These include epidermal growth factor (EGF) (Reynolds and Weiss 1992), fibroblast growth factor-2 (FGF-2)(Gage et al. 1995), brain-derived neurotrophic factor (BDNF) (Pencea et al. 2001), thyroid hormone T3 (Ben-Hur et al. 1998), transforming growth factor alpha (TGF-α) (Fallon et al. 2000), and bone morphogenetic proteins (BMPs) (Coskun

and Luskin 2002). In their paper, (Alvarez-Buylla and Garcia-Verdugo 2002) speculate that "most of the signalling molecules that allow neurogenesis to occur in the SVZ remain to be discovered."

Augurin is also expressed in ependymal cells lining the ventricles, and forming the CP. Those cells are glial cells, to which proliferative properties have been attributed. However, whether those ependymal cells are postmitotic cells remains controversial, as I have found conflicting reports on that issue. In (Chiasson et al. 1999) the authors claim that both ependymal and subependymal cells are able to proliferate, but only subependymal cells retain stem cells-like attributes. In contrast (Spassky et al. 2005) contend that ependymal cells are post-mitotic, and unable to re-enter the cell cycle. Recent data suggests that, by its action on the CSF flow through the beating of the ciliae, the ependymal cells lining the ventricles play a role in guiding new neurons through the RMS (Sawamoto et al. 2006). In light of those observations, it is tempting to speculate that augurin plays a role in neurogenesis happening in the RMS. Remarkably, even if it has received considerably less attention than the SVZ of the lateral ventricles, neurogenesis also occurs in the CP (Emerich et al. 2005). CP cells express mitogenic markers and once grafted into an injury site of the spinal cord, they were shown to proliferate and differentiate into astrocytes (Li et al. 2002).

In addition, a cell-fate decision function for augurin is supported by the fact that it was originally characterized as an oesophageal cancer-related gene and named ECRG4 by (Yue et al. 2003). In this paper, the authors make the important observation that in human oesophageal tumor tissues and several cancer cell lines, augurin mRNA is down-regulated as a result of its hypermethylation. The authors then conclude that silencing of augurin through hypermethylation may be involved in the development of cancer. Although down-regulation of augurin is not found in all cancer tissues/cell lines these data suggest that augurin may have a protective effect against oesophageal cancer and highlight its potential as a tumor-suppressor gene. Interestingly, another recent study showed that a low level of augurin expression is associated with aggressiveness of clear cell renal cell carcinomas (CCRCC) in a large human cohort (Kosari et al. 2005). These data clearly show that amongst 35 candidate genes tested by RT-PCR, augurin is the single most consistently down-regulated gene in aggressive CCRCCs. Furthermore, in a recent patent (Stache-Crain et al. 2008), augurin is described as a stromal-derived factor regulating survival or proliferation of hematopoietic stem cells (HSCs). In this patent it is stated that augurin/ECRG4 (termed SCR-5) mRNA is

expressed highly in a mouse stromal cell line whose function is to supply an environment to support proliferation or survival of HSCs, and that augurin/ECRG4 gene products (after introduction of the cDNA into the stromal cell line) increased proliferation or survival of HSCs, *in vitro*.

A growing number of organs and specialised tissues are now thought to contain their own pool of adult progenitor cells (Oliver et al. 2004). Recently, several high-profile studies aimed to uncover and characterised those potential organ-specific progenitors. Those questions are still subject to controversy, as the techniques and criteria for identifying and defining those stem cells are hotly debated issues and need perhaps to be standardised. A technique that seems to be growing in popularity looks at isolating a so-called "side population" of cells that retain a number of stem-like characteristics (Hadnagy et al. 2006), such as mitogenic and differentiating capabilities *in vitro* and homing to sites of injury *in vivo* (Oliver et al. 2004; Vankelecom 2007). Endocrine tissues which are thought to harbour stem cells include the anterior pituitary (Subburaju and Aguilera 2007), adrenal gland (Lichtenauer et al. 2007), thyroid gland (Takano 2007), trachea and lungs (Kim et al. 2005), kidney (Oliver et al. 2004), retina/eye (Tropepe et al. 2000), brain SVZ (Alvarez-Buylla and Garcia-Verdugo 2002; Yamashita et al. 2006), CP (Emerich et al. 2005; Li et al. 2002), ependymal and sub-ependymal cells of the ventricles (Li et al. 2002). Remarkably, those are all tissues where augurin is expressed at high levels in mouse or human.

The fact that augurin has not been found in the CSF (Stark et al. 2001) or blood serum argues against an endocrine function for it, and is consistent with this hypothesis of an autocrine/paracrine action of augurin on self or neighbouring tissues. All these observations concur in assigning a role for augurin in the maintenance of resident stem cells fate or environment of stem cells niches. In a recent review on vascular niches (Nikolova et al. 2007), one sentence particularly caught my attention : "haematopoietic stem cells shift from a quiescent osteoblastic niche to a vascular niche that supports their proliferation and further differentiation in the bone marrow". I speculate that augurin plays a role in maintening quiescence in the stem cell niche, and opposes the proliferation/differentiation effects of the vascular niche.

In summary, given its unique expression in neurogenic regions of the brain and stem cells-containing endocrine glands, and its *in vivo* downregulation in invasive renal carcinoma, I

predict that augurin is important in the homeostasis the stem cell niches, either by exerting a direct antimitogenic or antiproliferative effect on stem cells, or by regulating the cell fate of cells which regulate ECM structures around stem cells niches. As more resident stem cells are being characterised in tissues it is becoming plausible that carcinomas can be caused by disruption of signalling pathways not only in stem cells, but also in cells regulating the structure and the environment of stem cells niches ("cancer stem cells hypothesis" (Hadnagy et al. 2006)).

4.8 Possible links between augurin and human disease

In this section I discuss a number of diseases in which augurin could be involved. Augurin's possible involvement in common diseases is of particular interest, since it is likely to be secreted *in vivo*, and as such, augurin analogs constitute potential novel therapeutic drugs. However, one must be aware that only the additional knowledge of its cognate receptor would in practice enable the screening for augurin analogs from among chemical compounds banks.

4.8.1 Aortic valve sclerosis

In the United States calcific aortic stenosis is the third most common cardiovascular disease (Rajamannan et al. 2003) and the most common indication for surgical valve replacement (Rajamannan et al. 2005). Augurin is expressed very highly in osteoblasts (Figure 52) and in cells of the aortic valves. Aortic valve cells have been shown to express many bone matrix proteins and osteoblast markers, including osteopontin, cbfa-1, and alkaline phosphatase (Rajamannan et al. 2002). Cholesterol strengthens this osteoblast-like phenotype, and arthrosclerotic changes correlate with a higher phenotypic similarity of aortic valve cells to osteoblasts and an increase in the expression of osteoblast markers. Conversely, the drug atorvastatin improves the arthrosclerosis-like phenotype in the valves and at the same time reduces the expression of osteoblast markers (Rajamannan et al. 2002). Interestingly, this increase in bone matrix protein production induced by cholesterol is paired with an increase in cellular proliferation, another element hinting at a relationship between cancer and ectopic calcification. Although ectopic calcification in aortic valves has been described for more than a century, little is known concerning the synthesis of bone matrix molecules in the valve or the genes controlling this process (Rajamannan et al. 2005). The study of augurin putative signalling and gene expression control might shed new light onto this pathological state.

4.8.2 Hydrocephalus

Hydrocephalus is a condition in which the abnormal accumulation of CSF in the brain ventricles causes increased intracranial pressure and progressive enlargement of the head, convulsion, and mental retardation. The prevalence of hydrocephalus, while showing a decreasing trend in Europe at 0.82 in every 1000 births still remains one of most common birth defect (Persson et al. 2005). It accounts for approximately 40% of birth defect CNS abnormalities and its causes remain obscure (Lang et al. 2006). There are three general pathogenic mechanisms for hydrocephalus formation: excessive CSF production, physical blockade of CSF flow and insufficient reabsorption of CSF. Stenosis of the aqueduct of Sylvius between the third and fourth ventricles is considered as the main cause of impaired CSF flow in hydrocephalus (Meiniel 2007). The SCO is thought to play a major role in congenital hydrocephalus caused by a physical blockade of CSF flow (non-communicating hydrocephalus) (Meiniel 2007), and is one of the region of the brain where augurin protein is expressed at high levels. One of the known functions of the SCO is to secrete glycoprotein, that form Reissner's fibre (RF) which promote adequate laminar flow through the cerebral aqueduct. Defects in RF formation can lead to stenosis and hydrocephalus (Vio et al. 2000). In two foetuses afflicted by congenital hydrocephalus the size of the SCO was markedly reduced (Castaneyra-Perdomo et al. 1994) and the absence of SCO in mice was postulated to explain spontaneously occurring congenital hydrocephalus in wt mice (Takeuchi et al. 1987) and hydrocephalus in Msx1-deficient mice (Fernandez-Llebrez et al. 2004). Disruption of several developmental genes expressed in the dorsal midline of the diencephalon have been showed to cause hydrocephalus in animal models, including Wnt1, Pax6 and Msx1 (Meiniel 2007). KOs of other genes involved in the regulation of ciliary beating cause hydrocephaly, including RFX3 (Baas et al. 2006) and PAC1R (Lang et al. 2006). In this review (Meiniel 2007), the author claims that a common feature in animal models of hydrocephalus is the lack of SCO secretory function. Another disease causing hydrocephalus is tuberous sclerosis. In this state, tumors grow out from the ventricles and cause a disruption of the ependymal cells lining the ventricles (Castaneyra-Perdomo et al. 1994; Fernandez-Llebrez et al. 2004; Takeuchi et al. 1987). Since augurin is conspicuously expressed in ventricular midline structures of the brain, which play decisive roles in hydrocephalus pathogenesis (SCO), I hypothesize that augurin could be involved in this disease.

4.8.3 Cancer

The augurin/ECRG4 gene has already been implicated in tumor in at least two independent studies. It has been found to be distinctly down-regulated in human cell lines and oesophageal squamous carcinoma tissues (Yue et al. 2003). It has also been found significantly down-regulated in nearly all aggressive clear cells renal cell carcinomas (CCRCCs), compared to non-aggressive CCRCCs (Kosari et al. 2005). In addition, I have found a patent on a possible relationship between augurin/ECRG4 and haematopoietic stem cells regulation and cancer (cf. Resources section). The additional fact that augurin is found in many epithelial/ependymal cells-containing tissues of the body known to harbour stem cells (most notably the SVZ of the brain) prompted me to hypothesize a direct involvement of augurin signalling in carcinomas invasiveness (cf. 4.7.5).

4.8.4 Other diseases

CP, osteoblast cells, and endocrine glands seem to be three of the major sites of augurin expression, suggesting it is essential for proper functioning of those tissues/organs. Defects of CP function, such as those caused by excessive CP calcification, have been involved in several pathologies of the brain including schizophrenia and Alzheimer disease (Caputo et al. 1998; Emerich et al. 2005; Johanson et al. 2004). In this review, the authors argue that the increase of β-amyloid plaques in the brain of Alzheimer disease patients could be partly due to a defect of their clearance from the CNS, through the action of CP on CSF turnover. Furthermore, augurin's suggested relationship with the RAAS implies a possible involvement in pathologies related to the disruption of blood pressure and water/sodium homeostasis (Castaneyra-Perdomo et al. 1998; Connell and Davies 2005). Disruption of augurin signalling in osteoblasts could have consequences in bone metabolism defects, including osteoporosis, while disruption of a putative function in the CSF-blood barrier maintenance has been associated to common neurological diseases, including HIV-induced dementia, multiple sclerosis, and Alzheimer disease (Ballabh et al. 2004).

4.9 Possible connections with other peptidergic systems

Just like augurin, leptin is expressed in many endocrine tissues of the body where it modulates a myriad of physiological processes. It is known that one of the main functions of leptin is to regulate feeding behaviour through its binding to receptors of hypothalamic arcuate nucleus neurons. Leptin also plays an important role in reproduction. Circulating levels of leptin correlate with stages of sexual maturation in humans (Kiess et al. 1998), and

administration of leptin accelerates the onset of puberty in leptin-deficient (*ob/ob*) mice (Ahima et al. 1997; Chehab et al. 1997). However, to my knowledge, leptin physiological action on the reproductive system remains poorly understood and little seems to be known about the physiological significance of specific leptin binding to mouse tissues where augurin is highly expressed, including the CP (Devos et al. 1996; Tartaglia et al. 1995) and the adrenal gland (Glasow et al. 1998). Interestingly, the mouse leptin receptor was cloned from CP tissue, after it was shown that leptin binds to that tissue (Tartaglia et al. 1995). Leptin has been shown to regulate thyroid gland function probably through its regulation of pituitary TSH (Seoane et al. 2000), but the precise mechanisms underlying this regulation remain unclear (Flier et al. 2000). I also noted reports indicating that leptin induces vascular calcification, probably by promoting an osteoblast-like phenotype on vascular cells (Parhami et al. 2001). Those observations raise the possibility that augurin and leptin interact, possibly by a regulation of leptin on augurin.

When I looked for correlated expression patterns between known peptides and augurin, I found that the pattern of expression of IGF-II in the brain was similar to that of augurin (it is highly expressed in CP, but not in the ventricular walls). Moreover, the peptides and monoamines ANP, IGF1, serotonin and vasopressin are all present in the CSF, their corresponding receptors are expressed in the CP, and are known to modulate its function (Nilsson et al. 1992). These systems are thus likely candidates for the modulation of augurin expression in the CP and other organs. Generally, receptors such as ANPR, and transcription factors such as RFX3 which are expressed highly in CP and/or the linings of the brain ventricles are considered ideal candidates as partners of augurin in implementing its function in the brain.

4.10 Final remarks and conclusion

The cancer stem cells/stem cell niche hypothesis (subsection 4.7.5) is surely the most compelling functional hypothesis for augurin. Not only does augurin's expression suggest a relationship between it and resident stem cells, but those anatomical observations are further backed up by the reports from groups that it is involved in epithelial cancer and disruption of augurin signalling is likely to contribute to cancer aggressiveness. The additional fact that many tight junctions proteins, such as claudins, are expressed in those same midline

structures where augurin is expressed, further strengthens the hypothesis that augurin is a growth factor-like protein involved in stem cells -and possibly CSF barriers- maintenance.

Based on our *in vitro* secretion and processing studies, and augurin mRNA and protein expression pattern, I believe that augurin will provide another example of those pleiotropic signalling molecules, and that its study will advance our understanding of human physiology.

5 Preliminary characterization of four extra candidates

In this section, I present partial data on four other candidate human PH which were ranked among the top 300 scoring proteins of my initial search.

The names I use to refer to these candidates are constructed as follows: They begin with the letters “CPH” (for “Candidate PH”) and end with two digits corresponding to their last two Ensembl peptide ID digits (human protein).

The first candidate CPH36 is identified in Ensembl as ENSP00000342336, the second CPH44 corresponds to the Ensembl peptide identifier ENSP00000326044, the third CPH51 corresponds to ENSP00000339251 and the last CPH10 is identified in Ensembl as ENSP00000342110. All four putative genes corresponding to these peptides were annotated in Ensembl as possessing more than 1 exon, and were present in multiple species, thus making it likely that they are real genes and not pseudogenes, as it occurs frequently in the case of single-exon annotated genes. CPH36, CPH44 and CPH51 genes are present in multiple vertebrate species, including the zebrafish and tetraodon while CPH10 seems to be only present in mammals. I decided to study those four candidates after the analysis of the conservation across organisms of a number of their features, including putative signal peptide and PC/Furin cleavage sites. This was made possible by gathering orthologs sequences from the Ensembl database and aligning them using the ClustalW program (cf. Resources). I discarded sequences which contained missing exons so as to have a better representation of conservation in the alignment. Both primary sequence features and partial data on mRNA expression of those genes are discussed in the following section:

5.1 Primary sequence features and mRNA expression of candidates

5.1.1 Primary sequence features of CPH36

CPH36 is a precursor protein with a clear signal peptide and a perfectly conserved motif (Lys/Arg)-X-(Arg/Lys)-Arg highly suggestive of a PC/Furin processing site. The CPH36 gene is present in all vertebrates available through Ensembl. Its N-terminal sequence retains in all species the characteristics of a signal peptide (N-terminal region positively charged, core hydrophobic region and small uncharged residues in a region c-terminal to the hydrophobic part) (Nielsen et al. 1997), and human and mouse N-terminal sequences have been predicted by both my HMM and the SignalP-NN/HMM (http://www.cbs.dtu.dk/services/SignalP/) programs to encode a signal peptide. It is

noteworthy that two CPH36 paralogous genes were found in the stickleback and medaka fish. The fact that those fish conserved two copies of the gene argues for an important role for CPH36.

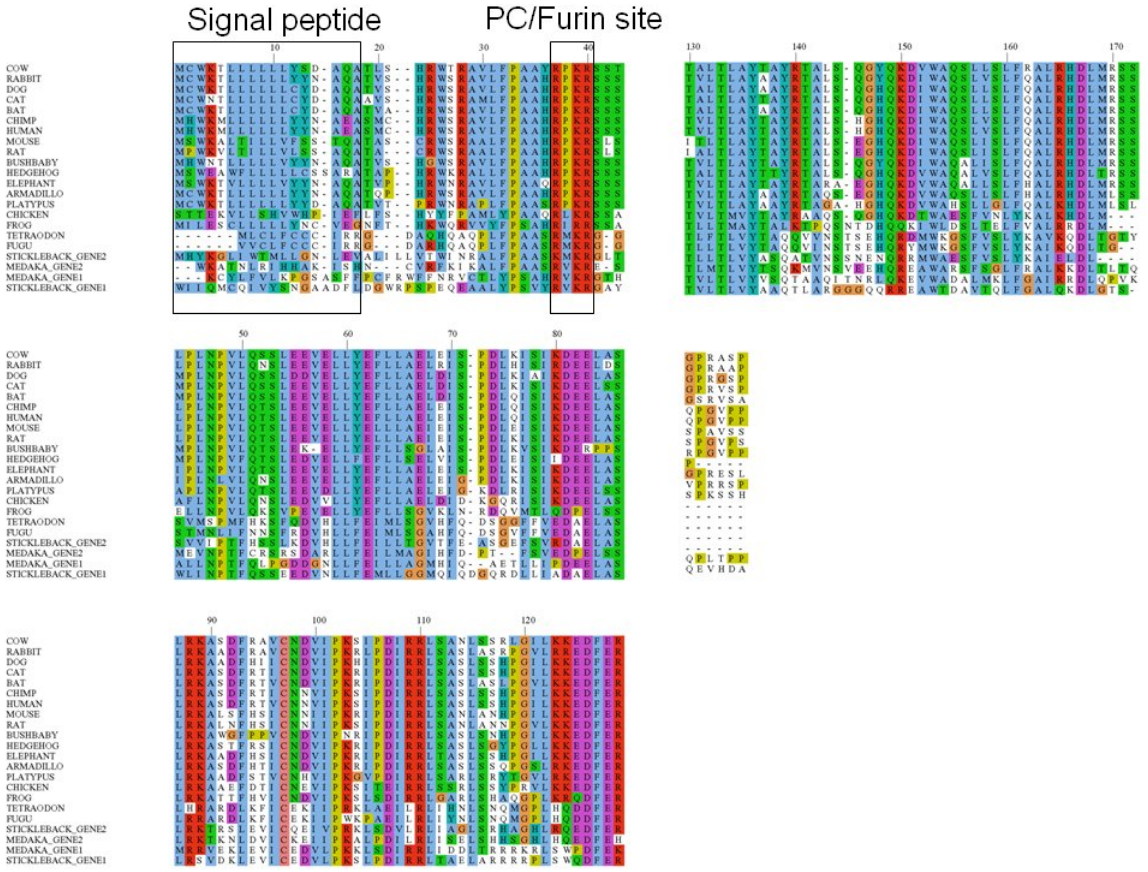

Figure 61: multiple alignment of protein CPH36 orthologs
Notice the characteristic signal peptide, with its core hydrophobic region (residues in blue). The presence of a perfectly conserved Arg-Xaa-Lys-Arg motif (boxed) is characteristic of the presence of a *bona fide* PC/Furin cleavage site

5.1.2 Primary sequence features and mRNA expression of CPH44

5.1.2.1 Primary sequence features of 44

CPH44 encodes a short protein with multiple dibasic cleavage sites (cf. Figure 62). At the time of my search I reasoned that the likelihood of CPH44 harbouring at least one functional PC/Furin cleavage site would be high. At the time of the initial search only mouse and rat orthologs were used to infer cleavage sites and peptides from that protein. As DNA sequences and Ensembl orthologs identification from other vertebrates were made available it became

evident that the pairs of basic residues 2, 3 and 4 (cf. Figure 62) noted in both human and mouse sequences were not present in all vertebrates (for instance in frog and stickleback fish). Only the first dibasic site seems to be conserved across vertebrates, and is likely to be a Furin/PC cleavage site.

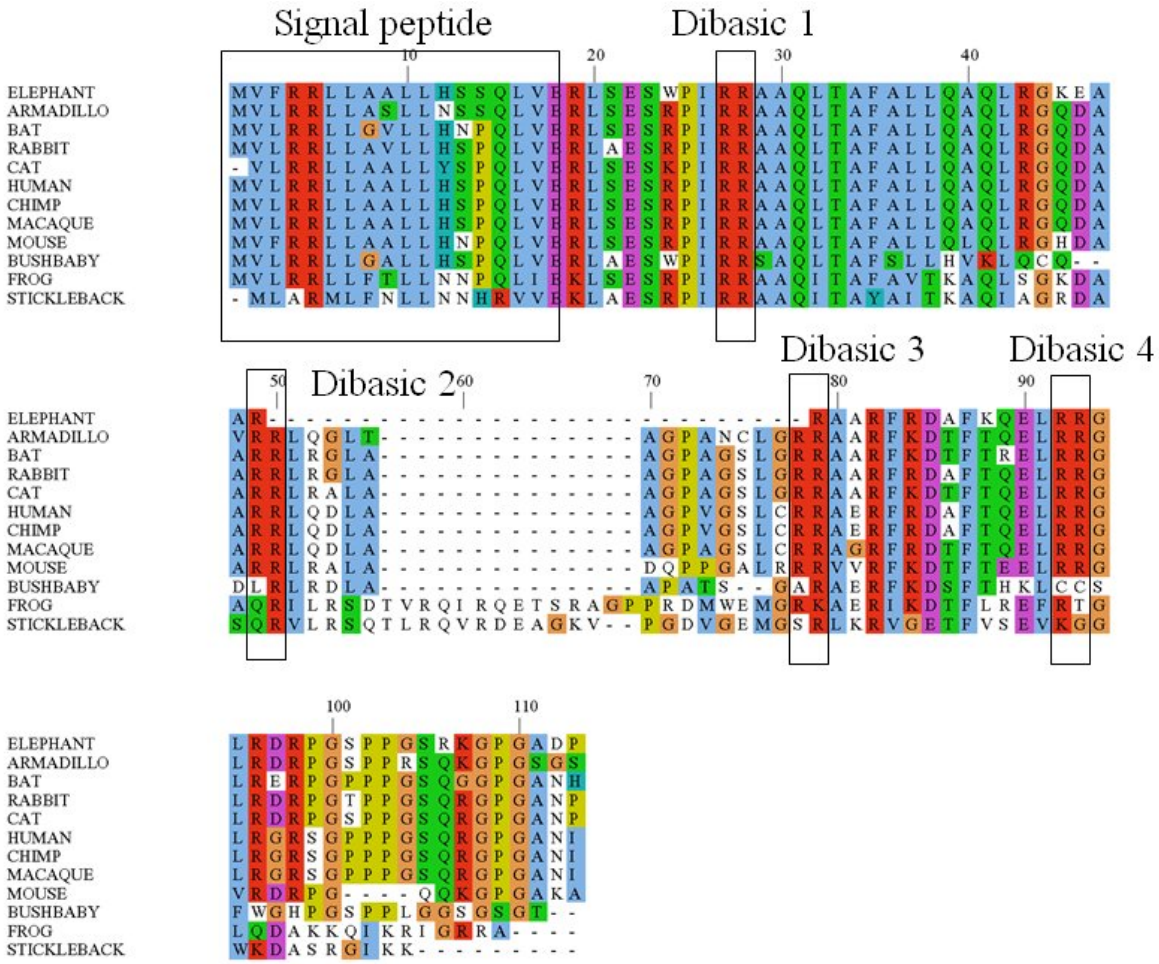

Figure 62: multiple alignment of protein CPH44 orthologs

The putative signal peptide has a short hydrophobic region. The protein and mouse sequences contain four putative dibasic cleavage sites. However only dibasic site 1 is fully conserved across all vertebrates. Note the highly hydrophobic stretch of aa 29-42 which could be a transmembrane domain.

Worryingly, its N-terminal sequence contains a very short stretch of hydrophobic aa, raising doubts about the possibility that it encodes a real signal peptide. I also noted a highly hydrophobic stretch of aa (positions 29-42 in the alignment) that I believe could be a transmembrane domain, contradicting the predictions of the original HMM (not the one I presented in this thesis, cf. Mirabeau et al. 2007). The fact that the HMM did not predict that region to be a transmembrane domain could be explained by the fact that I chose deliberately to tune the HMM so that it did only label transmembrane regions which were clearly long

hydrophobic regions (longer than 18). This choice was made to avoid discarding valid candidates from the screen, at the cost of retrieving a few membrane proteins among the top candidates (3% in the top 500, cf. Mirabeau et al. 2007). However, I have *in vitro* data suggesting that CPH44 is secreted, albeit not efficiently (cf. section 5.2)

5.1.2.2 CPH44 mRNA Expression

CPH44 expressed sequence was cloned from mouse brain tissue by RT-PCR using the following primers:

forward primer m44F1: `5′-GAA GGC GAA GGT AGC GAG CGT C-3′`
reverse primer m44R2: `5′-GGA ATT CAG CTT CAG TCC TGC TG-3′`

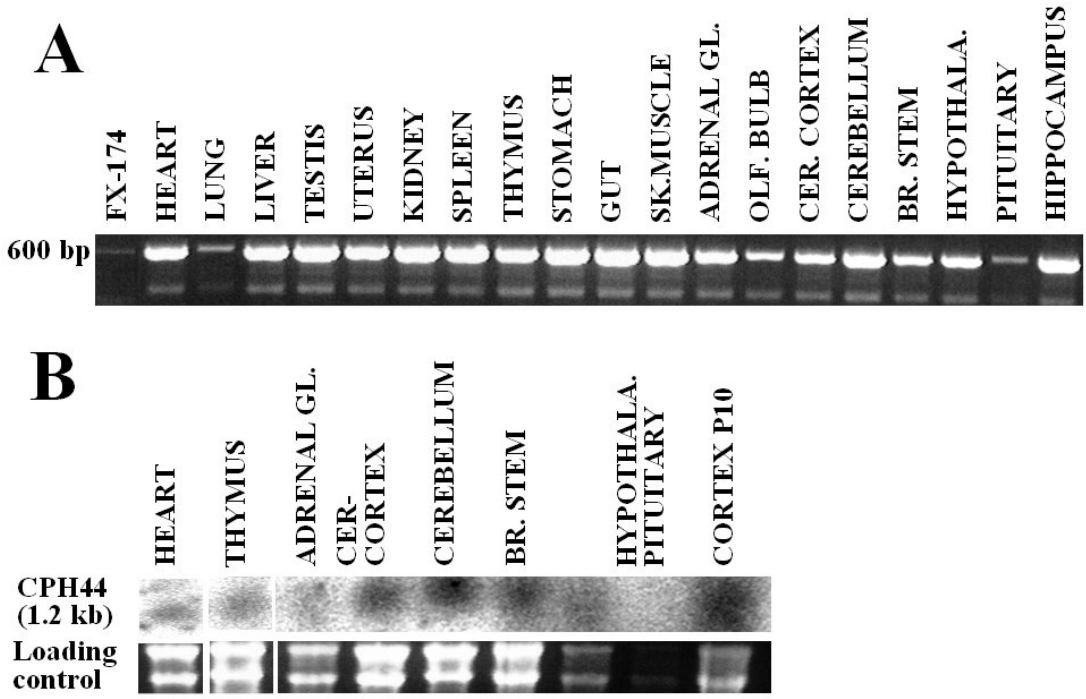

Figure 63: CPH44 mRNA expression in mouse tissues
(A) RT-PCR of heart, lung, liver , testis, uterus, kidney, spleen, thymus, stomach, gut, skeletal muscle, adrenal gland, olfactory bulb, cerebral cortex, cerebellum, brain stem, hypothalamus, pituitary, and hippocampus mouse tissues using primers m44F1 and m44R2.
(B) Northern blot of heart, thymus, adrenal gland, cerebral cortex, cerebellum, brain stem, hypothalamus, pituitary, and cortex from P10 mouse tissues, using m44Probe.
CPH44 mRNA is detected by both techniques in all mouse tissues surveyed.

The 621 bp band that was amplified was cloned into a pCR-II TOPO vector (Invitrogen, Carlsbad, CA., USA), and was used as a probe for Northern blotting:

m44Probe=gaaggcgaaggtagcgagcgtcgggttccgtggcgcggggagagccatggtcttccggcg
gctgctggcggccctgctgcaaacccgcagctggtggagc…
…agcacgactgagagtcatggacaatcagaaagcaagcctgccgtctgccagcaggactgaagctgaa
ttc.

Primers m44F1 and m44R2 were used for the RT-PCR expression studies. RT-PCR data show ubiquitous expression of CPH44 in mouse tissues. Northern blot data show high expression in the brain and some expression in heart and adrenal gland (cf. Figure 63). SymAtlas (cf. Resources) data show homogeneous expression of CPH44 in all tissues at a low amount (data not shown).

5.1.3 Primary sequence features and mRNA expression of CPH10

5.1.3.1 Primary sequence features of CPH10

CPH10 is a short protein (78 aa with the putative signal peptide) well conserved in mammals with four perfectly conserved cysteines spaced in a way characteristic of C-X-C chemokines, forming pairs of disulfide bonds. The putative signal peptide of CPH10 is found to be highly conserved across mammalian sequences collected. Furthermore, the two perfectly conserved cysteines at position 11 and 16 strongly suggest disulfide bridging with the equally conserved pair of cysteines at position 57 and 60 (cf. Figure 64). High conservation of such meaningful functional motifs as pairs of cysteines is unusual in signal peptides.

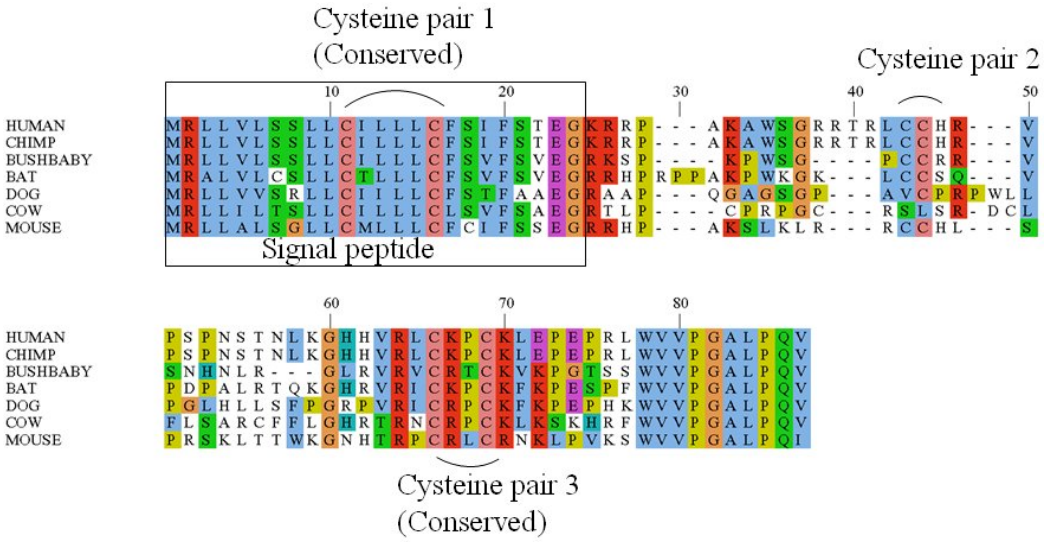

Figure 64: multiple alignment of CPH10 orthologs
Note the highly conserved n-terminal signal-peptide-like region and the two perfectly conserved pairs of cysteines, close to the n- and c- termini.

This strongly advocates for an additional role of CPH10 N-terminal region other than the standard ER targeting function. However, this does not rule out the possibility that CPH10 possesses a signal peptide as some signal peptides fulfil multiple functions in the cell (Martoglio and Dobberstein 1998).

Both N-terminal and C-terminal ends are well conserved, while the central part of the protein is poorly conserved. I noted the presence of perfectly conserved arginines at positions 25 and 64 (numbering according to alignment of Figure 64). Those positions correspond to the border between regions of high and low homology in the alignment, reminiscent of the pre-pro-

insulin primary structure where the central region of the precursor is cleaved off and the insulin A and B chains at both ends remain bound to each other by cysteine residues.

5.1.3.2 CPH10 mRNA Expression

The primers I used to amplify CPH10 cDNA from brain, after total RNA reverse-transcription were:

m110F=5′-ACT CCT TCA CTA TGA GAC TTC TAG-3′

m110R=5′-TCT GAA GTG aaG CAT TCC AGT GTC A-3′

I cloned the 365bp-long product of the PCR in a pCR-II TOPO vector (Invitrogen, Carlsbad, CA., USA), and used it as a probe for Northern blot experiments.

The sequence of the probe was:

m110Probe=gctccttcactatgagacttctagcccttccggtctgctctgcatgctgctcctctgt
ttctgcattttctcctcagaagggagaagacatcctgccaagtccttgaaactcaggcgctgctgtca
cctatctcctagatccaagctgacaacctggaaaggaaaccacacaaggccctgcagactctgcagaa
acaagctaccagtcaagtcatgggtggtgcctggggctctcccacagatatagggcctcctgaagcgt
tgatgcccagatgtggagacaccagaagcatacacactatgttgccttgccccttgccaatgagctgt
gacactggaatgcttcacttcaga

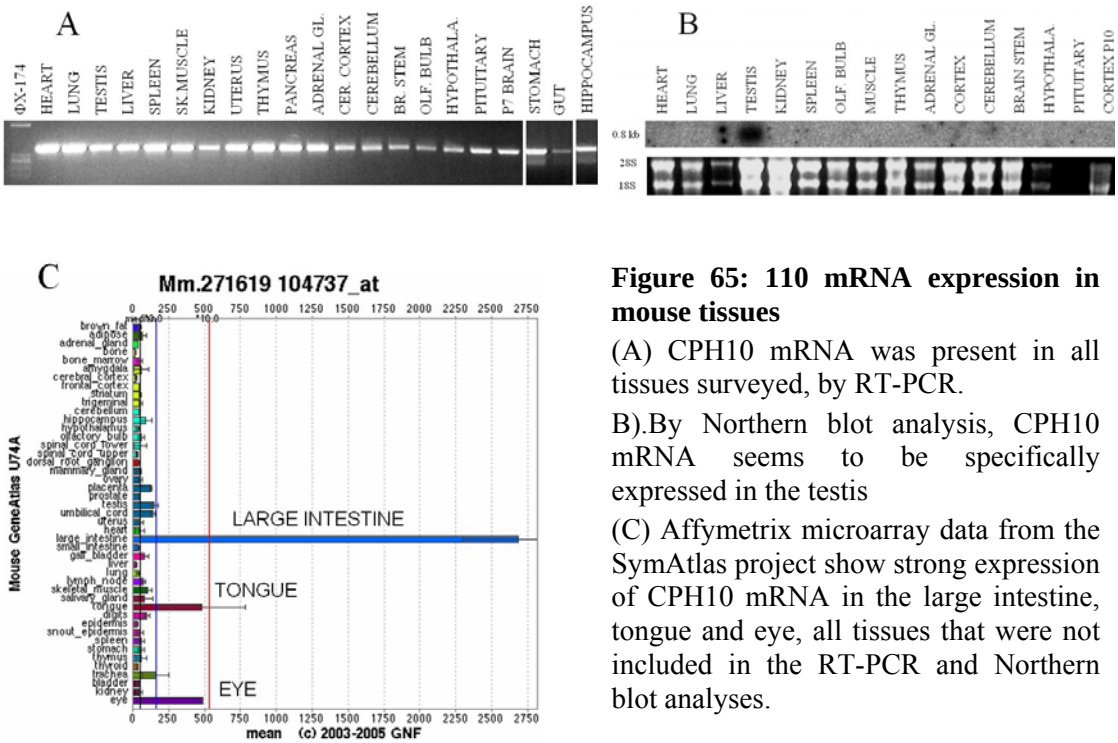

Figure 65: 110 mRNA expression in mouse tissues

(A) CPH10 mRNA was present in all tissues surveyed, by RT-PCR.

B).By Northern blot analysis, CPH10 mRNA seems to be specifically expressed in the testis

(C) Affymetrix microarray data from the SymAtlas project show strong expression of CPH10 mRNA in the large intestine, tongue and eye, all tissues that were not included in the RT-PCR and Northern blot analyses.

Using primers m110F and m110R, I was able to amplify CPH10 transcripts by RT-PCR in all tissue samples surveyed, including the heart, lung, liver testis, spleen, thymus, kidney, uterus,

pancreas, adrenal glands, pituitary gland, stomach, gut, and regions of the brain that include the hypothalamus and hippocampus. However, by Northern blot analysis I could only observe a band in the testis sample (cf. Figure 65), suggesting that the expression of the CPH10 mRNA is stronger in the testis. In contrast, the SymAtlas public database shows strong expression of mouse CPH10 mRNA in the large intestine, tongue and eye, all tissues which were not included in my Northern blots.

Examples of peptides highly expressed in the large intestine include peptide tyrosine-tyrosine (PYY), endothelin, insulin-like 5 hormone, and the chemokine CCL6.

5.1.4 Primary sequence features and mRNA expression of CPH51

CPH51 (recently named U467) is a highly conserved protein only present in mammals, that seems to have an mRNA expression restricted to epithelial tissues (cf. Figure 67) and has recently been described as a keratinocyte-differenciation protein (KDAP).

5.1.4.1 Primary sequence features of CPH51:

It is clear by the alignment of Figure 66 that CPH51 is very likely to contain a signal peptide. Signal peptide prediction programs (SignalP-HMM and SignalP-NN) confirmed that intuition and predicted for human CPH51 cleavage at ^{22}Gly with high probability.

Furthermore, a careful examination of the alignment revealed a conserved repeated motif Phe-(Leu/Iso)-Asn-Trp, reminiscent of short repeated peptides found in the sequence of many insect and molluscs PH precursors (the see hare buccalins, the beetle Allostatins) and human pro-enkephalin and thyrotropin-releasing hormone (TRF) precursors. However, in those cases the conserved motif is typically flanked by conserved dibasic residues, which is not the case for CPH51.

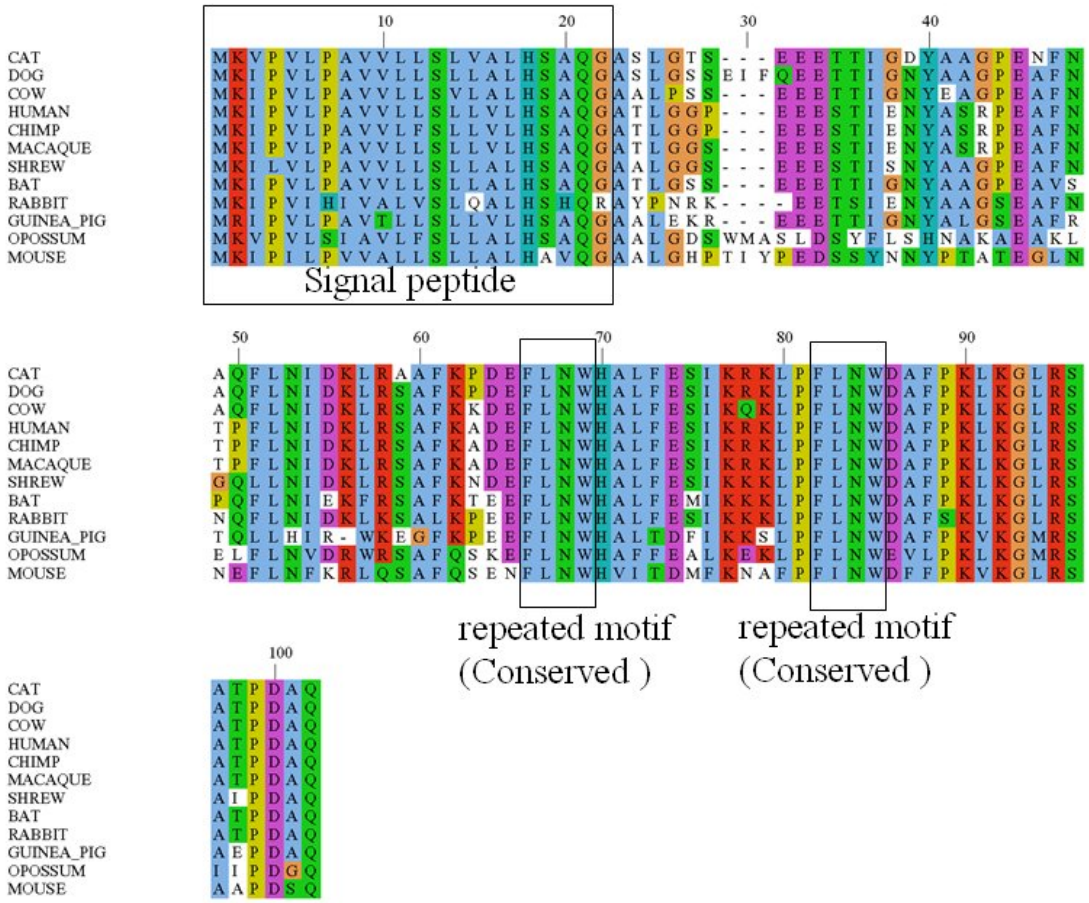

Figure 66: multiple alignment of protein CPH51 orthologs

This multiple alignment of cat, cow, dog, chimp, macaque, mouse tree shrew, has been generated using the ClustalW program with default settings. The signal peptide shows strong conservation across those organisms, pointing towards additional functions for it. Note the perfectly conserved repeated motif F(L/I)NW

5.1.4.2 CPH51 mRNA Expression:

Neither RT-PCR nor Northern blot experiments have been performed to characterize the expression of CPH51 mRNA. However, the expression pattern provided by the SymAtlas suggests that its expression in both mouse and human is restricted to a few tissues including tongue, skin, tonsils, digits, epidermis, trachea and ombilical cord (cf. Figure 67).

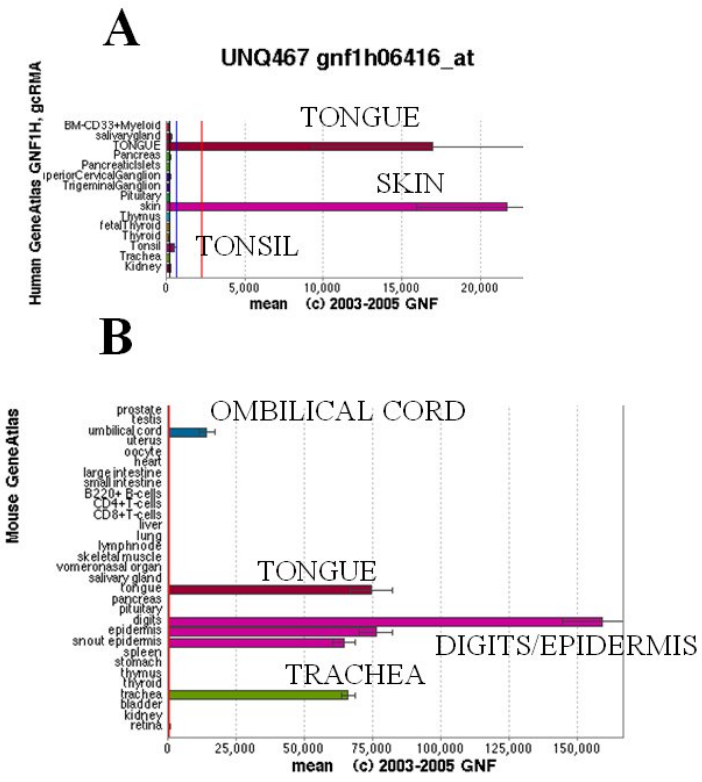

Figure 67: Affymetrix microarrays expression of CPH51 gene [data from SymAtlas]
Levels of CPH51 mRNA expression in human (A) and mouse (B) tissues, assessed with the Affymetrix microarray technology

The expected mouse CPH51 transcript was isolated from skin (epidermis) tissues, using the following primers:

m51F= 5′-AAC CGG ACA CCA TGA AGA TC-3′

m51R=5′-TCT TAA AGA GAA TGG TCA CTG-3′

The fragment amplified was sequenced and shown to correspond to the expected transcript:

Mouse51transcript=5′-
accggacaccatgaagatcccaattcttcccgtcgtggctctcctctctcttctggcattgcatgcgg
tccagggagcagccctggggcatcccacgatatacccggaagatagcagctacaataattaccctacc
gcaacagagggccttaacaatgagttcctgaactttaagaggctacagtctgcctttcagtcagaaaa
cttcctgaactggcacgtcatcactgatatgttcaaaaatgcatttcctttcattaactgggacttct
tccctaaggtgaaaggactgagaagtgccgctcctgattcccagtgaccattctctttaagacctagg
acctaggctgagcccagtagaagaggaggcaagcatggaatctgaagtccatcctacgacaaatcttc
cttgcctcagttcccccaataa-3′

5.2 Secretion and processing of FLAG constructs in cell culture

Next, I studied secretion and processing of these four extra candidates, by inserting in their DNA coding sequence a sequence coding for the FLAG peptide (cf. FLAG-tagging strategy

in subsections 3.3.1 and 4.3.1). I used M2 anti-FLAG antibodies to identify secreted products of the candidates by Western blotting

5.2.1 Description of constructs:

Clones containing coding sequences for mouse orthologs of CPH36 (RZPD: IRAKp961D14138Q2) and CPH44 (IRAUp969B0236D6) genes were ordered from the German Resource Center for Genome Research (RZPD, cf. Resources section). Coding sequences for mouse CPH51 and CPH110 were cloned from mouse tissue RNA, by RTPCR using primers m51F/R and m110F/R (cf. previous section 5.1). Coding sequences were then FLAG-tagged. The FLAG tag was placed at the c-terminal end of candidates CPH36, CPH51, and CPH10 while it was placed after the first putative cleavage site of CPH44 (cf. Figure 68). Constructs were then cloned into the expression vector pcDNA3.1 (Invitrogen, Carlsbad, CA., USA.).

Figure 68 shows the primary structure of candidates with their predicted features including signal peptide and PC/Furin cleavage sites. Conservation of residues across organisms is also given. All putative cleavage sites underlined in Figure 68 were found to be also predicted by either the ProP or the Neuropred program (cf. Resources section). Both softwares are dedicated to the prediction of PC/Furin cleavage sites (Duckert et al. 2004; Southey et al. 2006). All references to residue positions in the following section will be made according to the numbering of Figure 68.

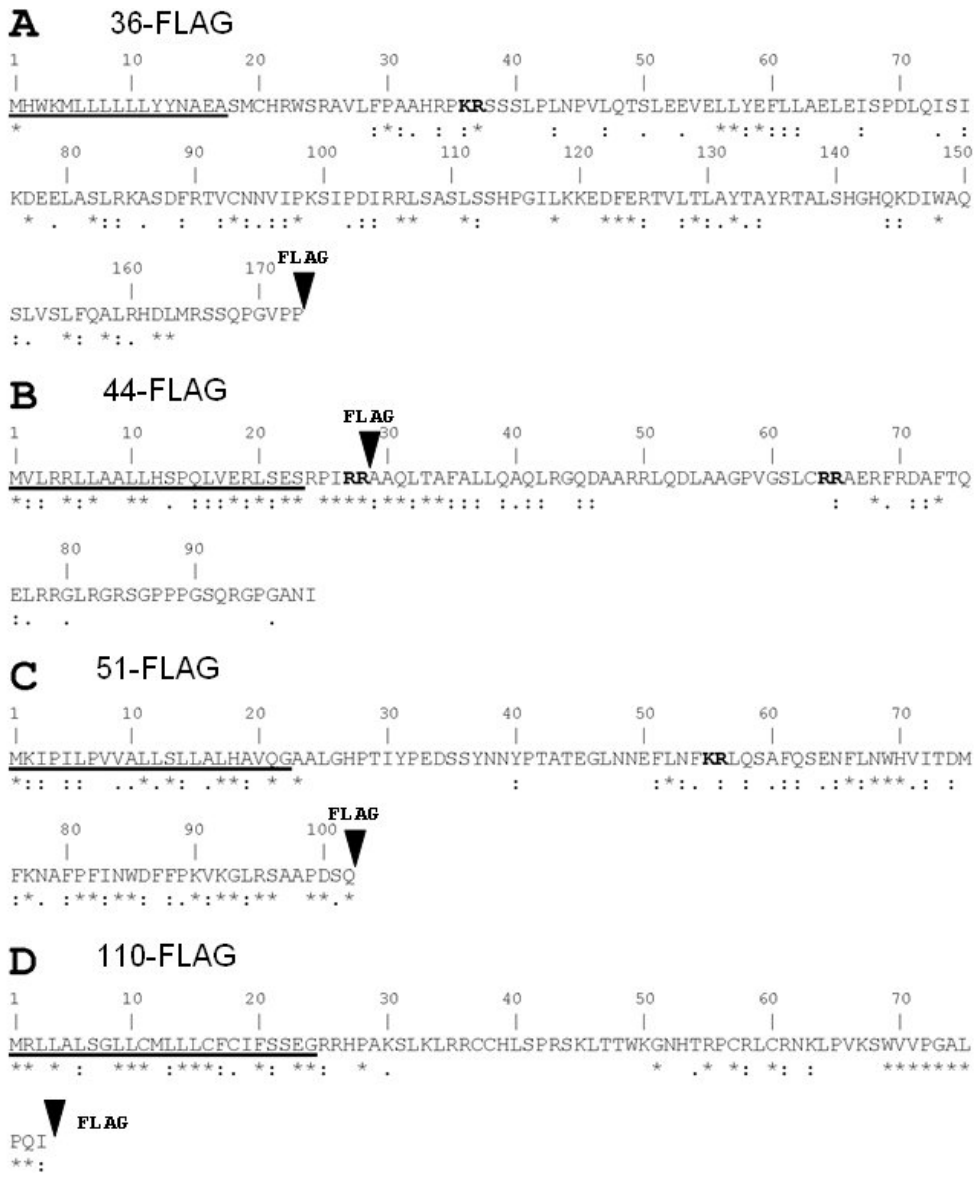

Figure 68: primary structure of FLAG-tagged (A) mouse CPH36, (B) human CPH44, (C) mouse CPH51, and (D) mouse CPH10 proteins.

Signal peptides are underlined, and putative cleavage sites are highlighted in bold. Conservation among orthologs is shown below the sequences. The conservation annotations were generated with the ClustalW program: (*) identity, (:) high homology, lesser homology (.). The positions where FLAG tags were inserted are indicated with an arrow (cf. Materials and methods for the strategy used to insert FLAG antigen sequences).

5.2.2 Secretion and processing of FLAG-tagged constructs, in culture

Pancreatic β-TC3 cells were transfected (cf. Materials and Methods) with expression vectors containing the constructs described in (Figure 68). Cells were left one day to recover in their growing medium. One day after transfection medium was replaced with fresh growing

medium and collection of conditioned-supernatant occurred 24 hours later for analysis. Cells were then washed and cell lysates analysed for the presence of FLAG-containing proteins (cf. Materials and Methods).

At least one FLAG-36 (short for FLAG-CPH36) immunoreactive (IR) protein band was found both in the supernatant and in the cell lysate of transfected β-TC3 cells. Its apparent molecular weight of about 18-19 kDa (cf. Figure 69) matches the expected theoretical size of 18.6 kDa of the 36-FLAG protein without signal peptide (residues 18-173). However, the predicted 36-FLAG peptide (residues 38-173) of theoretical size of 16.2 kDa is unlikely to have been produced by the cells, although a possibility remains that the apparent single band observed accounts for the presence of both protein fragments, since their theoretical size does not differ by much (about 2 kDa).

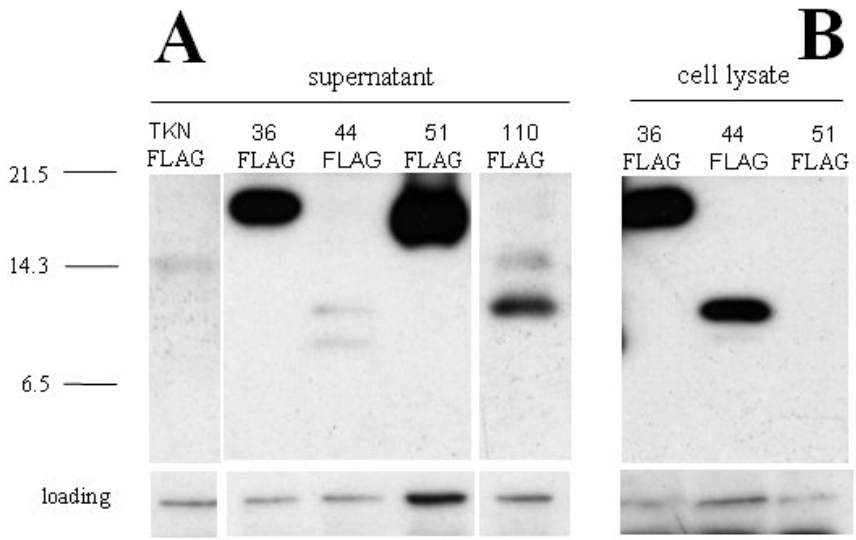

Figure 69: Detection of CPH36, CPH44, CPH51, and CPH110 in cell supernatants (A) and lysates (B).
FLAG-tagged proteins were transfected into rat pancreatic cells in culture, and cell supernatants were harvested and submitted to immunoblotting with an anti-M2 Flag antibody. TKN-FLAG: positive control.

A large amount of 51-FLAG protein was detected in the supernatant. In contrast, the corresponding cell lysates did not contain any trace of 51-FLAG proteins, strongly arguing for CPH51 to be a secreted protein. However, I could only see a single band, corresponding to a rather large protein (15-16 kDa) meaning that, at least in this cell system, CPH51 is not processed as PH generally are. Furthermore, the apparent MW of 51-FLAG, does not correspond to the expected size of the FLAG-protein after cleavage of the signal peptide

(fragment 23-102), which is estimated at 10.2 kDa. Since the full length 51-FLAG precursor should theoretically weigh more than 12 kDa, an absence of processing of the 51-FLAG putative signal peptide cannot account on its own for the size discrepancy. Hence, it is tempting to hypothesize that CPH51 is post-translationally modified (PTM). The sequence of CPH51 was submitted to the Eukaryotic linear motifs (ELM) database site (http://elm.eu.org/help.html, cf. Resources section) to see if it contained possible linear motifs associated with post-translational modifications, which could explain this size gap. Two such sites were found QSAF(58-61) and RSAA(95-98), that could be sites of attachment for glycosaminoglycans.

The presence of at least two bands of MW 8-9 kDa and 10-11 kDa (cf. Figure 69) indicates that the 44-FLAG fusion protein is processed by β-TC3 cells. The theoretical MW of a 44-FLAG protein fragment where the putative signal peptide (1-23) [resp:first putative cleavage site at position 28] has been removed is 9.2 kDa, while this MW becomes 8.6 if the first putative PC/Furin cleavage site at position 28 is used. The sizes observed do not strictly correspond to those expected ones, but that discrepancy could be explained by the difficulty in predicting the exact signal peptide cleavage site in this protein. However, we can say that the lower band (8-9 kDa) is likely to correspond to a processed 44-FLAG where the first CS at position 28 was used (fragment of size 8.6 kDa). No other putative cleavage site seems to be used in β-TC3 cells. Although 44-FLAG seems to be produced in non-negligible quantities inside the cells (cf. Figure 69B) the processed products of 44-FLAG do not seem to be efficiently secreted in this cell system. That could partly be explained by the structure of the putative signal peptide at the n-terminus of the protein, which is atypical, with an unusually short and conserved hydrophobic core, a sign that it might fulfil other functions than just ER targeting. Indeed, when SignalP was run on the CPH44 protein sequence to confirm/infirm the presence of a signal peptide, the answer was unclear. The Neural network output SignalP-NN ruled out that this protein contained a signal peptide, while the HMM-based algorithm strongly indicated that it did, not surprisingly given the probable similarity between my signal-HMM and signalP-HMM (I merely followed the ideas from the authors of signalP-HMM). On the other hand I noted that the intensity ratio between the lower band (presumably the processed from of 44-FLAG) and the higher band (unprocessed form) is larger in supernatants, compared to cell lysates. This observation supports the hypothesis that the multiple bands correspond to secreted Furin/PC-processed products of 44-FLAG and not merely unspecific degradation products of it.

110-FLAG products were seen in both supernatants (cf. Figure 69) and cell lysates (data not shown). As for 44-FLAG, the presence of at least two bands was observed in the supernatant. The lower band, with an apparent MW of 10-11 kDA could correspond to the full-length 110-FLAG protein, with the signal peptide included (theoretical MW of 8.9 kDa). The higher band is likely to represent a modified 110-FLAG protein where carbohydrates radicals, such N-linked oligosaccharides were added. Glycosylation is a frequent post-translational modification occurring in proteins of the secretory pathway (Apweiler et al. 1999). Interestingly, the sequence of CPH10 contains a canonical motif (Mellquist et al. 1998) for N-glycosylation attachment at position 52 (Asn-His-Thr). However, this putative N-glycosylation is not likely to be an important determinant of CPH10 functions, as it is not a conserved feature of that protein (cf. alignment of Figure 64).

5.3 Conclusions

I have presented data arguing that all four candidates are secreted *in vitro*, albeit at varying efficiencies. CPH36 and CPH51 are produced and secreted at high amounts by β-TC3 cells. CPH51 is likely to be PTM. However, they are unlikely to be processed, and as such do not qualify as promising *bona fide* PH precursors.

In contrast with CPH36 and CPH51, lower amounts of CPH44 and CPH10 are secreted by the cells I used. CPH44 is the only candidate likely to be processed; while CPH10 is possibly PTM through N-glycosylation.

One must be aware that the amount of proteins that the cells are forced to produce after transient transfection of the expression vector (CMV promoter) construct could account for a low amount of secreted protein. For that reason, one should be cautious in interpreting the results of the Western blot analysis. In spite of those concerns, I believe that this data strongly suggest that those are likely to be novel secreted proteins, and as such, potentially novel signalling molecules. Some of them, most notably 51 and 110, seem to have an mRNA expression restricted to a few tissues. This should prove to be a source of insights to study their *in vivo* functions. For instance, the high expression of 110 in tissues particularly exposed to germs (large intestine, eye, tongue), together with the fact that it is likely to be a short di-cysteine-containing secreted peptide, suggest that it could be a novel chemokine, antimicrobial peptide, or a peptide involved in innate immunity.

6 Conclusion

6.1 Contribution to PH search algorithms

Several improvements in the HMM algorithms were implemented since the original series of protein sequences screening in 2004. The original A-HMM was only using organisms which constituted most of the well-annotated vertebrate organisms at that time, i.e. mouse, dog, rat and human.

Extension of the A-HMM algorithms to a larger number and greater diversity of vertebrate organisms, including the frog, chicken, opossum and several fish was shown to improve predictions of PH precursors. Incorporation of a conservation score along alignments allowed the prediction of PH features such as PC cleavage sites with greater sensitivity by priming difference of conservation between propeptide and peptide regions of precursors.

In addition, a series of scores were developed empirically, that work well in practice: firstly, the vast majority of known PH are found in the top 300 list. Secondly, very few non-peptide hormone precursors intercalate between classical ones meaning that those which are found high in the list (including spexin, augurin, and a third uncharacterized candidate CPH52) have a high probability of being peptide hormones. It is noteworthy that the quality of alignments (depending chiefly on the quality of annotations and gene models) greatly influences the quality of the predictions for the A-HMM and does so even more for the CSA-HMM. Some alignments score high by chance because, for instance, a cleavage site region was mistaken for a splicing site boundary region. In some cases the signal peptide of orthologous protein sequences are missing and sequences score more poorly than if they were taking as single sequences and a standard HMM ran on them. These cases are uncommon and bound to disappear with the improvement of annotation quality. Approaches that integrate homology information have become commonplace in recent years and are destined to be a main route for biological sequence analysis applications.

6.2 Improvements on PH prediction methods

State of the art gene models rely heavily on expressed sequence tags (EST) mapping on the genome. Although less so everyday, those EST are sometimes incomplete and gene models are missing some exons, often on the 5'-end where signal peptides are encoded.

In order to significantly improve the chances of discovering novel peptide hormones we would probably need to design a hybrid PH HMM coupled with a gene prediction HMM and run it directly on genomic sequences. The computational cost would significantly rise but we would potentially have access to many more candidates: Some PH genes may be missing an exon that encodes the important peptide or the signal peptide part of the protein and this way could be more easily detected and predicted as PH. The strategy I used is highly sensitive to missing signal peptides and incomplete protein sequences and working directly on genomic sequences would be an advantage to reconstitute characteristic features from PH precursor sequences.

The second area for improvement is on the automatic weighing of sequences in the alignment-based HMM algorithms. In the alignment-based HMM algorithms every sequence is weighted equally. This does not take into account how close sequences are from each other or the poor quality of certain sequences (erroneous annotations, long spliced variants, etc.) that may be identified because they align poorly with the other sequences. There are two problems with my simplistic approach. Firstly, organisms which have not diverged much from each other have a bigger influence on the A-HMM predictions. Secondly, in my case the conservation score is highly dependent on the organisms that were chosen to make the alignments. I was concretely confronted with this issue when I decided to run the CSA-HMM I had built (with human alignments) on alignments of sequences from the 5 fish organisms which were available through Ensembl. Conservation score distributions had to be re-learned because they were very different. Implementation of a conservation score and a weighing mechanism which makes this procedure less dependent on the organisms chosen would surely constitute an improvement over what was done in this thesis.

Moreover, PC cleavage sites are best predicted using machine learning methods able to detect non-linear signals. An important limitation in any case is the amount of data available (number of well-characterized cleavage sites) and especially information about sites which are not cleavable. It is possible more sites than we know of are recognized by prohormone convertases, but that the information about bioactivity, stability, and secretion propensity is located elsewhere in the sequence. For that, HMM are well adapted. It is possible that a hybrid method involving a first round of PC cleavage site prediction by specialised non-linear algorithms (SVM, neural networks, etc…) followed by an HMM able to capture information dispersed along the sequence would perform best. The algorithm described in 2.11.2 could be used to constrain the HMM to take into account those former PC cleavage site predictions.

6.3 Have novel peptide hormones been discovered?

The point of my PhD was ultimately to discover novel peptide hormones in the human genome. So can we say at this point that novel peptide hormones were discovered? Some may contend that a peptide only becomes real when it has been purified from of an animal, sequenced and shown to have some pharmacological *in vivo* activity that can be measured (vasodilation, stomach emptying, etc…). This implies that often peptides which have are active *in vitro* are not necessarily relevant for the biology of the organism.

In that restrictive sense neither spexin nor augurin has been proven to be a peptide hormone precursor. However, I believe there are a number of solid pieces of "relevant evidence" that point out towards a peptide hormone function for spexin and augurin.

Firstly, their primary sequence exhibit features which constitute a real signature of peptide hormone precursors including a conserved signal peptide and conservation absolutely restricted to a short charged region flanked by perfectly conserved dibasic sites). It is often tricky to predict the exact active endogenous peptide that is produced by the body from the peptide hormone precursor. However, it is much less of a long shot to claim that there must be an active peptide produced by the spexin and augurin precursor after inspection of alignments of their vertebrate orthologous sequences.

The second line of evidence comes from experiments showing that both spexin and augurin are secreted, processed and localized in dense core granules, after transfection of their coding DNA into endocrine cells. One may challenge the biological relevance of these experiments, since expression of the FLAG constructs were driven by an artificial CMV promoter and because cells in their natural environment behave differently from cells from immortalized cell lines. However, out of a dozen candidates that were cloned and FLAG-tagged, spexin and augurin were the only ones which showed an unambiguous pattern of processing and secretion *in vitro*. In addition, they were also the only candidates which localized in dense core granules of RIN cells. This contrast strengthened our presumption that spexin and augurin are *bona fide* PH.

Thirdly and perhaps most importantly, spexin peptide is a bioactive peptide since it contracts dose-dependently fundal gastric muscle strips *ex vivo* at micromolar concentrations. The concurrent fact that it is expressed at the level of the RNA in the stomach fundus makes it

likely that it is a gastrointestinal peptide hormone. However, caution should be taken about making claims about *in vivo* activity of spexin since known peptides are usually expected to act at nanomolar concentrations in similar assays.

Lastly, both spexin and augurin's expression profiles are compatible with the hypothesis that they are peptide hormones. Spexin is likely to be expressed in neuronal populations of the laterodorsal and pedunculopontine nuclei and the future may reveal that it acts as a neuromodulator in this area of the brain. It is also conspicuously expressed in the lower oesophageal sphincter strongly suggesting a role for it in the gastro-oesophageal system. Augurin's expression is stunningly restricted to endocrine organs of the body and tissues known to harbour stem cells.

Several elements hint on a hormone, growth factor, or neuropeptide-like role for spexin and augurin. However, only further pharmacological (i.e. injection of putative peptides in the circulation or brain ventricles of animals), genetic (disruption of gene with KO techniques or disruption of normal expression levels by insertion of transgene) and biochemical experiments (binding and activation of G protein-coupled receptors of putative peptides) will give us definite answers on whether spexin and augurin are *bona fide* PH.

7 Materials and Methods

7.1 Materials

7.1.1 Chemicals

All chemicals were, if not noted otherwise, purchased from the companies *Fluka*, *Merck*, *Roth* or *Sigma*.

7.1.1.1 Reagents for molecular biology

Description	Source
^{32}P-dGTP (3000 Ci/mM)	*Amersham/GE Healthcare*
BigDye 3.1 sequencing mix	*Applied Biosystems*
Bromphenol blue	*BioRad*
Deoxyribonucleoside triphosphates (dNTPs)	*Promega*
DNA marker	
λ DNA/ *Eco*91I (*Bst*EII)	*MBI Fermentas*
ΦX174, RF DNA (*Hae*III)	*Invitrogen*,
1kb Plus DNA Ladder	*Invitrogen*
EndoFree plasmid maxi preparation kit	*Qiagen*
Plasmids:	
pcDNA3.1 Hygro	Invitrogen
Enzymes:	
Klenow polymerase	*New England Biolabs*
Restriction endonucleases:	*New England Biolabs*
*Bam*HI (20 U/µl)	5´-G'GATCC-3´
EcorV	5´-GAT'ATC-3´
T4-DNA-Ligase (5 U/µl)	*MBI Fermentas*
T4-DNA-Ligase buffer	*MBI Fermentas*
-DNA-Polymerases	
GoTaq Flexi DNA Polymerase	*Promega*
Expland Plus	*Roche*
Oligonucleotide primers:	*Metabion, Sigma*
all oligos used for the design of the constructs are listed in the appendix	
dpN_6 (90 O.D. U/µl)	*Pharmacia*

7.1.1.2 Reagents for biochemical operations

Description	Source
Acrylamide (30%)/Bisacrylamide (0.8%)	*BioRad*,
Ammoniumpersulfate (APS)	*Sigma*
Prestained MW-Marker, "SeeBlue Plus 2"	*Invitrogen*
Coomassie Brillant Blue R250	*BioRad*
ECL	*Amersham/GE Healthcare*
β-Mercaptoethanol	*Sigma*
Milk powder	*Roth*
Protein Markers:	
Rainbow Low molecular weight	*Amersham/GE Healthcare*
PageRuler Prestained Protein Ladder	*Fermentas*
Proteinase Complete Inhibitor, EDTA-free	*Roche*
TritonX-100	*Boehringer Mannheim*
Tween20	*Sigma*

7.1.1.3 Cell culture operations

Description	Source
β-Mercaptoethanol	*Sigma*
Dimethyl sulfoxide (DMSO)	*Sigma*
DMEM cell culture medium	*Gibco/Invitrogen*
Fetal calf serum (FCS)	*Gibco/Invitrogen*
HEPES	*Gibco/Invitrogen*
PBS	*Gibco/Invitrogen*
Penicillin/Streptamycin	*Gibco/Invitrogen*
RPMI 1640 culture medium w/o L-glutamine	*PAA*
Trypsin	*Gibco/Invitrogen*

7.1.2 Cell lines

The following cell lines were used in this work:

Name	Derived from	Description
Beta-TC3	DSMZ (cf. Resources)	rat pancreatic beta cells
RINm5f	gift from David Tosh, University of Bath	rat pancreatic beta cells

7.1.3 Bacteria

Description	Source
E. coli XL-1 blue	*New England Biolabs*
E. coli TOP 10	*Invitrogen*

7.1.4 Mouse and rat strains

Name	Rodent	Source
C57Bl/6	mouse	*Charles River*
Albino Wilstar	rat	*Charles River*

7.1.5 Antibodies, dyes and other high affinity molecules

The table shows the various antibodies, dyes and high affinity molecules that I have used during my doctoral studies and lists their concentration, basic reactivity and companies they were purchased from.

Antigen (Species)	Dilution Stock	Reactivity	Stock	Source
Immunofluorescence/ IHC				
M1 Flag (mouse)	1:100 1 mg/ml	N-terminal FLAG epitope	1 mg/ml	*Sigma*,
M2 Flag (mouse)	1:100 1 mg/ml	FLAG epitope	1 mg/ml	*Sigma*,
Insulin (guinea pig)	1:100	n.s.	1 mg/ml	*Dako Cytomation*
Augurin (rabbit)	1:100	QLWDRTRPEVQQWYQQ FLYMGFDEAKFEDD	n.s.	*Primm*
Spexin (rabbit)	1:100	NWTPQAMLYLKGAQ-amide	n.s.	*Primm*
anti-guinea pig IgG FITC (488 nm)	1:200 2 mg/ml	Mouse IgG (H+L)s	n.s.	*Molecular Probes*
anti-mouse IgG Cy5 (568 nm)	1:200 2 mg/ml	Mouse IgG (H+L)s	n.s.	*Molecular Probes*
anti-rabbit IgG Cy3 (633 nm)	1:100 2 mg/ml	Rabbit IgG (H+L)	n.s.	*Molecular Probes*
Western blotting				
M1 Flag	1:100	N-terminal FLAG epitope	1 mg/ml	*Sigma*
M2 Flag	1:100	FLAG epitope	1 mg/ml	*Sigma*
anti-mouse IgG HRP (goat)	1:2000 0.8 mg/ml	Mouse IgG (H+L)	31430 98052825	*Pierce*
anti-rabbit IgG HRP (goat)	1:2000 0.8 mg/ml	Rabbit IgG (H+L)	31460 98061831	*Pierce*

7.1.6 Culture media, buffer and stock solutions

7.1.6.1 Molecular biology

Description	Composition
Ampicillin stock solution (1000×)	50 mg/ml
DNA loading buffer (6×)	0.25% Bromphenol blue 0.25% Xylencyanol FF 15% Ficoll
dNTP-Mix for PCR	2.5 mM dATP 2.5 mM dCTP 2.5 mM dGTP 2.5 mM dTTP
ethidium bromide stock solution	10 mg/ml ethidium bromide in H_2O
OLB	mix A:B:C, 200:500:300 O: 1.5 M Tris-HCl, pH 8.0 0.15 M $MgCl_2$ A: 0.832 ml solution O 18 µl β-Mercaptoethanol 50 µl dATP (10 mM) 50 µl dCTP (10 mM) 50 µl dTTP (10 mM) B: 2 M HEPES-NaOH, pH 6.6 C: dpN_6, 90 O.D. U/ml
Sodium acetate solution for DNA precipitation	3 M sodium acetate pH 4.6
TAE buffer (50×)	2.0 M Tris-Base pH 8.3 0.6 M Sodium acetate 0.1 M EDTA
TE buffer	10 mM Tris-HCl 0.1 mM EDTA adjust pH to 8.0, autoclave
Culture medium:	
Luria Bertani (LB) medium For selection purposes the appropriate antibiotic (ampicillin, chloramphenicol or kanamycin) was added to the LB broth.	10 g bactotryptone 5 g yeast extract 10 g NaCl ad 1 l H_2O
LB agar	15 g agar ad 1 l LB medi

SOC medium	0.5% Yeast extract 2.0% Bactotryptone 10mM NaCl 2.5mM KCl 10mM $MgCl_2$ 20mM $MgSO_4$ 20mM glucose

7.1.6.2 Biochemistry

Description	Composition
Ammoniumpersulfate (APS) stock solution	10% APS in H_2O
Coomassie Rapid Destain	30% isopropanol 6% acetic acid
Comassie Destain	10% ethanol 6% acetic acid
Coomassie staining solution	40% Coomassie Brilliant Blue R-250 50% methanol 10% acetic acid
IP lysis buffer	150mM NaCl 1mM EGTA 1mM EDTA 1% TritonX-100
PBS (20×)	32 g $Na_2HPO_4 \times 2H_2O$ 5.3 g $NaH_2PO_4 \times 1H_2O$ 164 g NaCl ad 1 l H_2O
Tris/Tricine sample buffer (2×)	100 mM Tris-HCl pH 6.8 30% Glycerol 8% SDS 5% β-Mercaptoethanol Bromphenol blue, ad libidum
Tris/Tricine anode buffer	0.1 M Tris-base, pH 8.9
Tris/Tricine cathode buffer	0.1 M Tris 0.1 M tricine 0.1% SDS
Tyrode's solution	137 mM NaCl 5.4 mM KCl 0.5 mM $MgCl_2$ 1.8 mM $CaCl_2$ 10 mM glucose 11.9 mM $NaHCO_3$ 0.4 mM NaH_2PO_4 at pH 7.4

Western Blot transfer buffer	25 mM Tris-base 192 mM glycine 25% Methanol (v/v)
Western Blot blocking buffer	3% non-fat milk powder in PBS 0.1% Tween20
Western Blot washing buffer	1×PBS with 0.1% Tween20

7.1.6.3 Cell biology

Description	Composition
Fixative	4% PFA/PBS
Gelvatol mounting medium	make 100 ml solution containing: 0.14 M NaCl 0.01M KH_2PO_4/Na_2HPO_4, pH 7.2 To the 100 ml solution, slowly add 25 g polyvinyl alcohol while stirring, stir o.n., adjust pH to 7.2 the next day. add 50 ml Glycerol, stir o Store airtight, at -20°C
PBS (20×)	32 g $Na_2HPO_4 \times 2H_2O$ 5.3 g $NaH_2PO_4 \times 1H_2O$ 164 g NaCl ad 1 l H_2O
Cell culture media:	
beta-TC3 culture medium	82.5 % Dulbecco's MEM 15 % horse serum 2.5 % FBS
beta-TC3 freezing medium	70 % culture medium 20 % FBS 10 % DMSO
RINm5f culturemedium	500 ml RPMI 1640 50 ml FBS

.

7.2 Methods

7.2.1 Methods in molecular biology

All DNA and RNA manipulations, unless otherwise specified, were carried out according to Sambrook (Sambrook et al. 1989).

7.2.1.1 Polymerase Chain Reaction (PCR)

This technique has revolutionised molecular biology by allowing selective replication (or amplification) *in vitro* of any fragment of DNA (Mullis and Faloona 1987; Saiki et al. 1988). This was made possible by harnessing the potential of thermostable enzymes capable of replicating DNA molecules. A PCR reaction consists in three steps: one step of denaturation where all molecules are set at a boiling temperature of 95º C to separate them from each other, one step of selective binding of the primers on DNA molecules (typical temperatures ranging from 58 to 62º C) and one step of amplification where the polymerase replicate DNA molecules starting from the free 3'OH end of primers (the elongation occurs in the 5'→ 3' direction).

For every DNA polymerase the reaction mixture was prepared according to the manufacturer's instruction [ExpandPlus (*Roche*), GoTaq Flexi DNA-polymerase (*Promega*)].

Standard Protocols:

PCR mixture:

1 µl	DNA from tail preparation	
	2 µl	10x Taq buffer
	1.5 µl	dNTPs (2.5 mmol)
	1.5 µl	primer A (5 pmol)
	1.5 µl	primer B (5 pmol, unless stated otherwise)
	0.2 µl	Taq DNA polymerase (Promega)
ad	20 µl	ddH_20

7.2.1.2 Agarose gel electrophoresis

This method can be used to separate, based on their size, fragments of DNA, and to visualise them as single "bands" which fluoresce under ultraviolet (UV) light. The principle is the following: The complex mixture of nucleic acids polymers to resolve is loaded inside a gel of agarose. An electric field is then generated in the gel, generating an electrostatic force that applies to this molecule, and that is proportional to the charge it carries. Since this charge is itself commensurate with the number of nucleic acids it contains (negative charge), the distance travelled in the gel during the application of the electric field is inversely proportional to the size of the molecule. (Maniatis T. 1982; Meyers et al. 1976).

Ethidium bromide added in the gel intercalates between the double strands of the DNA molecule, and the complex becomes fluorescent under ultraviolet light. This allows visualising distinct "bands" of DNA, after they have been separated by size.

An agarose gel preparation is characterized by its percentage in the buffer. A 1% agarose gel contains 1g of agarose per mL of buffer. The buffer used in this procedure is the standard TAE buffer (see the buffer section for the recipe). The greater the percentage of a gel, the better it will be at resolving small bands. Reciprocally, gel with low percentage will be adequate for separating larger DNA molecules. In my case, I had to generate many constructs of small DNA molecules (200-800 base pairs) and most of the time I used high percentage gels (greater than 3%).

7.2.1.3 Gel purification of DNA fragments

To extract DNA fragments of interest from agarose gels the QiaexII kit (*Qiagen*) was used. Briefly, the piece of agarose gel containing DNA was separated and melted at 50°C in the presence of buffered glass milk. The DNA binds to the beads and can then be washed, dried and eluted in a small volume.

7.2.1.4 Ligation of DNA fragments

The coupling of two DNA fragments via a phosphodiester binding is called ligation. The enzymes capable of promoting this reaction under the consumption of ATP are called ligases. The reaction typically took place at 16°C o.n. in a small total volume of 10 µl containing 1-3 units of the T4 DNA ligase, 1 µl of 10x T4 ligation buffer and DNA. The ratio of vector to insert used in the ligation reactions was set between 1:4 and 1:8.

7.2.1.5 Production of chemically competent bacteria

Chemically competent bacteria cells are quickly prepared but do not exhibit the highest transformation capacity (~10^7 transformants/µg DNA). XL1-Blue cells were grown in pure LB medium to an optical density of OD_{600}≈0.5 – 0.6. The pellet was resuspended in 1/10 volume of freshly made TSB and incubated on ice for 10 min. For long-term storage, 200 µl aliquots were kept frozen at -80°C.

7.2.1.6 Transformation of bacteria

Bacteria (100 µl) were thawed on ice and combined with 20 µl 5x KCM, 50 –100 ng DNA, and 100 µl ddH_2O. The transformation preparation was incubated for 20 min on ice followed by 10 min at room temperature. The bacteria were split and plated on two plates containing the appropriate antibiotic (50 + 150 µl).

7.2.1.7 Plasmid preparation from bacteria

Plasmids can be inserted into bacteria and multiplied by the efficient cellular machinery of bacteria. After growth of plasmid-containing bacteria one needs to extract pure DNA from bacterial cells.

First plasmid-containing bacteria were grown in a culture o.n.; 2-5 ml for a Mini prep, 200-300 ml for a Maxi prep. The next morning, a mini prep kit (*Qiagen*) was used to purify DNA for small preparations and the endo-free maxi kit (*Qiagen*) for large amounts of DNA.

7.2.1.8 DNA sequencing

To obtain the exact sequence of a DNA the Didesoxy-method was used (Sanger et al. 1977). Its mechanism is based on the random termination of DNA replication by the insertion of fluorescent

ddNTP analoga, which are detectable in a reader. The sequencing reaction included the following reagents:

		10	ng/100 bp	DNA
		3.2	pmol	primer
		4	µl	BigDye 3.1 (*Applied Biosystems*)
		2	µl	5x Dilution buffer (*Applied Biosystems*)
ad	20	µl		H_20

and a PCR was performed following this program:

Cycles:	95°C	5 sec		1x
	95°C	15 sec	}	25x
	52°C	15 sec		
	60°C	3 min.		

The DNA was then precipitated by addition of 64 µl 100% ethanol and 16 µl H_20 followed by centrifugation (20 min, 14.000 rpm), and washed with 50 µl 70% ethanol. The DNA was then dried and re-dissolved in 10 µl H_20. 2 µl of the sequencing reaction was mixed with 8 µl formamide and heat-denatured before it was sequenced in a capillary sequencer (*Applied Biosystems*).

7.2.1.9 RNA extraction from mouse tissue

Tissue was grinded in glass tube with a pestle and homogenised with 1mL of Trizol per 50-mg of tissue. The homogenisation was done for no more than 5 minutes at room temperature. 200 µL of chloroform was added per 1 mL of Trizol in the sample, to separate proteins from nucleic acids in two distinct phases. After shaking vigorously the sample and 2-3 minutes incubation at room temperature, the sample was centrifuged at 12000 rpm for 15 minutes at 4°C. The aqueous phase was then removed, which contained the nucleic acids. Placed in another tube and 75% of ethanol in DEPC water was added, and the sample mixed by pipetting up and down several times.

Then in order to proceed further in the purification of the RNA from the sample, the instructions from the Qiagen RNeasy Mini Kit protocol were followed: About 700µl of sample was transferred in a mini column, and centrifuged for 15s at 10000 rpm. The flow-through was discarded and 350 µl of proteinase K-containing RW1 buffer (provided in the kit) was added to digest the remaining proteins in the sample, followed by a centrifugation for 15s at 10000 rpm. At this point a DNAse mix (10 µl Dnase I and 70 RDD buffer provided in the Qiagen kit) can be added directly to the silica membrane. 15-30 minutes incubation is necessary to digest the DNA in the sample at room temperature. 350 µl of buffer RW1 was again added to the column, which was spun at 10000 rpm for 15s. The flow-through was then discarded and 500 µl of wash buffer RPE was then added to the sample, which was then centrifuged at 10000 rpm for 15s. This washing step was repeated once, and the pure total RNA was then eluted by adding 30-50 µl of RNAse-free on the column water and centrifuging the sample for 1 minute at 10000 rpm.

The total amount of nucleic acids extracted from the tissues was then quantified using a spectrophotometer, and the sample purity was assessed by calculating the ratio of its absorbances at wavelengths 260nm and 280nm (the ratio is noted A260/A280) (Maniatis T. 1982). Only RNA samples with a ratio greater or equal to 1.8 were kept for analysis.

7.2.1.10 Radioactive labeling of DNA after Vogelstein

20–50 ng of a linear DNA fragment was heat denatured (5 min, 100°C), briefly chilled on ice and added to the labelling solution:

	10 µl	OLB
	2 µl	BSA (10 mg/ml)
	5 µl	^{32}P-dGTP, 3000 Ci/mM
	1 µl	Klenow enzyme (5 U/µl)
add	25 µl	H_20

The mixture was incubated o.n. at RT or 2-3 h at 37°C. The reaction was terminated by the addition of 100 µl stop solution (50 mM Tris-HCl pH 8.0, 20 mM EDTA) and purified using a BioGel column. 1 ml BioGel was given into a 1 ml syringe with a bit of glass wool at the bottom to avoid the flow through of the gel. The column was packed by centrifugation for 2 min at 2000 rpm. The labelled probe was given onto the column and purified by an additional centrifugation (2 min, 2000 rpm, Megafuge) during which the probe travels through the columns while unbound nucleotides remain in the bedding. The probe needs to be denatured (5 min, 100°C) before it can be used in a Southern Blot.

7.2.1.11 Northern blot

For a Northern blot 10-15 µg of RNA were loaded on an agarose gel and slowly separated by size. A picture with a ruler was taken before the RNA was transferred by capillary transfer on a nitrocellulose membrane (GeneScreen Plus). After the transfer the RNA was cross-linked to the membrane. Specific constructs were detected using radioactively labelled probes.

Non specific binding sites were blocked by incubating the blot for 1 hr in hybridization buffer at 62°C, before the labelled probe was given onto the blot in fresh hybridization buffer and incubated o.n. at 62°C. The next day the blot was washed several times with a wash buffer before it was exposed on a phosphoimager screen or an X-ray film.

7.2.2 Spexin and augurin FLAG constructs

7.2.2.1 Strategy for cloning the FLAG constructs

The following strategy was used to insert FLAG tags into DNA coding for the proteins studied:

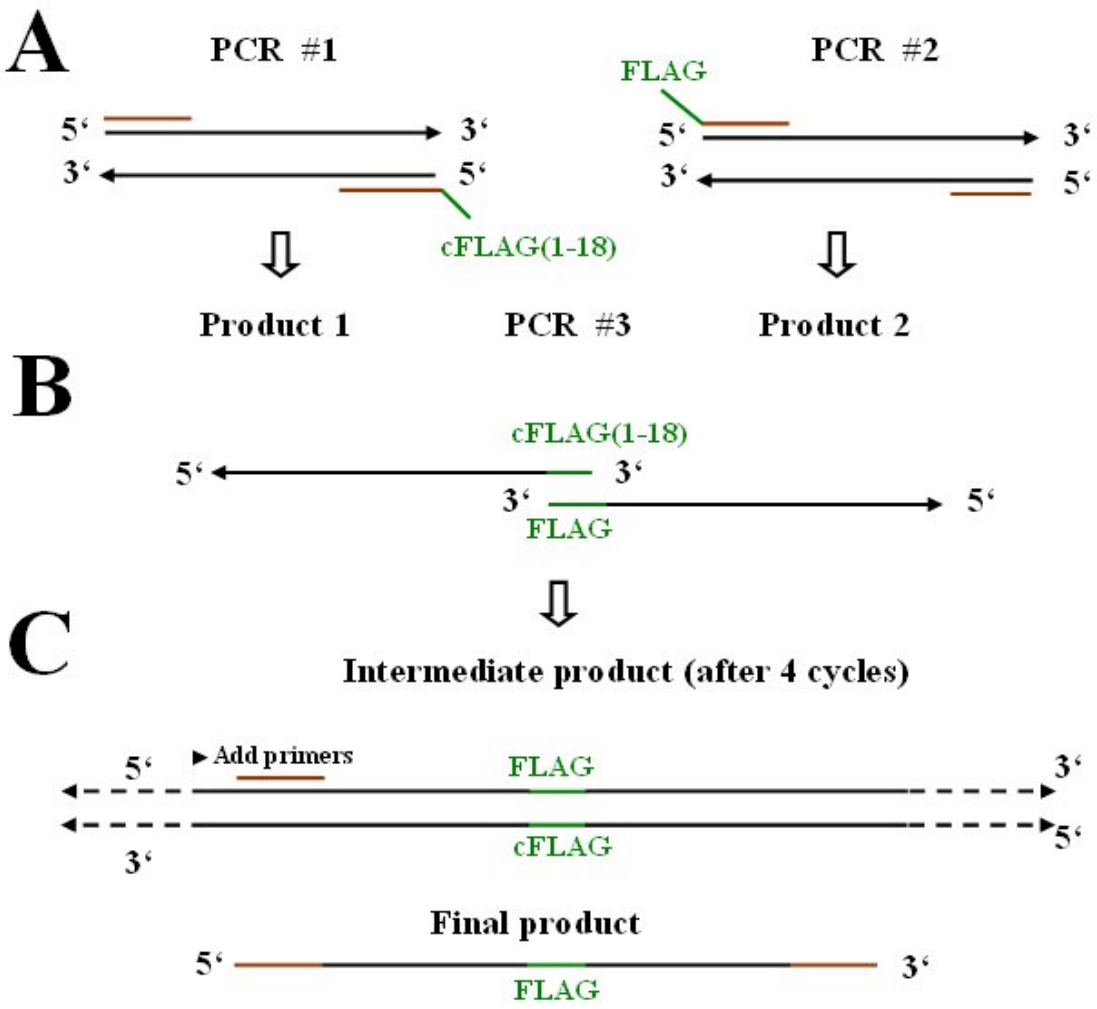

Figure 70: PCR strategy for cloning FLAG constructs

To insert DNA coding for FLAG tags 3 PCRs were performed. (A) The first two can be done in parallel to generate products with either partial complementary (PCR#1) or full (PCR#2) FLAG DNA sequences at their 3' end. (B) In a third PCR, those two products were then mixed for a low number of cycles (4 is the standard) to generate full lengths PCR products containing the FLAG sequence inserted. (C) PCR#3 was then interrupted and primers were added to generate the desired fragment with FLAG DNA inserted.

cFLAG stands for the complementary DNA sequence of the FLAG peptide. cFLAG(1-18) represents the complementary sequence of the first 18 nucleic acids coding for the FLAG protein that was sufficient to bind selectively to the FLAG sequence. The different primer pairs used to generate the different constructs are listed in the appendix.

All constructs were cloned into a PCDNA3.1 vector (Invitrogen), and expression of spexin- and augurin-FLAG fusion proteins were driven by the constitutive CMV (cytomegalovirus) promoter.

7.2.2.2 Description of FLAG-spexin constructs

cf. APPENDIX A for a list of the primers that led to the cloning of these constructs.

S-FLAG: DNA coding for the FLAG antigen sequence DYKDDDDK was inserted just downstream of the potential signal peptide cleavage site coding sequence (cf. Figure 34).

N-FLAG: The FLAG antigen sequence DYKDDDDK was inserted just c-terminal of the first potential PC/Furin cleavage site (cf. Figure 34).

C-FLAG: The FLAG antigen sequence DYKDDDDK was inserted at the c-terminus of the potential spexin precursor (cf. Figure 34).

N-FLAG/Δ51: Construct made from the N-FLAG template where the first arginine of CS2 at position 51 (boxed in Figure 34) was mutated into a stop codon.

N-FLAG/R35G: Construct made from the N-FLAG template, where the triplet coding for the arginine (R) of the first putative cleavage site CS1 at position 35 (underlined in Figure 34) was replaced by a glycine (G) codon.

N-FLAG/R51L: Construct made from the N-FLAG template, where the triplet coding for the arginine (R) of the second putative cleavage site CS2 at position 51 (underlined in Figure 34) was replaced by a leucine (L) codon.

N-FLAG/R60S: Construct made from the N-FLAG template, where the triplet coding for the arginine (R) of the third putative cleavage site CS3 at position 60 (underlined in Figure 34) was replaced by a serine (S) codon.

N-FLAG/R72G: Construct made from the N-FLAG template, where the triplet coding for the arginine (R) of the first fourth cleavage site CS4 at position 72 (underlined in Figure 34) was replaced by a glycine (G) codon.

N-FLAG/N36Q: Construct made from the N-FLAG template, where the triplet coding for the asparagine (N) of the putative n-glycosylation site at position 36 (underlined in Figure 34) was replaced by a codon for glutamine (Q).

C-FLAG/R51G: Construct made from the C-FLAG template, where the triplet coding for the arginine (R) of the first putative cleavage site CS2 at position 51 was replaced by a codon for glycine (G)

7.2.2.3 Description of FLAG-augurin constructs

N-FLAG: The FLAG antigen sequence DYKDDDDK was inserted just c-terminal of the third potential PC/Furin cleavage site CS3 (cf. Figure 47).

C-FLAG: The FLAG antigen sequence DYKDDDDK was inserted at the c-terminus of the potential augurin precursor (cf. Figure 47).

N-FLAG/R70G: Construct made from the N-FLAG template, where the triplet coding for the arginine (R) of the third putative cleavage site CS3 at position 70 (underlined in Figure 47) was replaced by a glycine (G) codon.

N-FLAG /R67G: Construct made from the N-FLAG template, where the triplet coding for the arginine (R) of the third putative cleavage site CS2 at position 67 (underlined in Figure 47) was replaced by a glycine (G) codon.

N-FLAG/R70G/R67G: Construct made from the N-FLAG/ R70G template, where the triplet coding for the arginine (R) of the third putative cleavage site CS2 at position 67 (underlined in Figure 47) was replaced by a glycine (G) codon.

7.2.3 Biochemical methods

7.2.3.1 Polyacrylamide gel electrophoresis (Tricine-PAGE)

Proteins can be separated according to their size using a polyacrylamide gel. I used a variation of the classical SDS-PAGE using a Tris/Tricine buffer system to obtain a good separation of small proteins. Protein gels were prepared and run with a discontinuous buffer system (Laemmli, 1970):

	15 %	9.8 %	stack
30 % Acrylamide (ml)	15	9.8	1.6
3 M Tris-HCl, pH 8.45, 0.3%SDS(ml)	10	10	…
87% glycerol (ml)	3.17	3.17	…
H_20 (ml)	21.1	32.8	7.78
0.5 M Tris-HCl, pH 6.8, 0.4%SDS (ml)	…	…	3.1
10 % APS (ml)	0.50	0.50	0.025
TEMED (ml)	0.01	0.01	0.005
	30	30	12.5

Protein samples were compounded with 2x Tris/Tricine loading buffer, heat denatured (5 min, 95°C) and spun down. On every gel, 6 µl marker (prestained for Western Blots) was loaded to determine the size of the proteins. Protein separation was achieved by applying for approximately two hours an electrical current of 100 V.

7.2.3.2 Western blot

For Western blots, proteins were electrophoretically transferred onto a polyvinylidone difluoride (PVDF) membrane. In order to retain very small proteins PSQ membranes (Millipore) were used that have pore < 0,2 µm. Prior to the transfer, the PVDF membrane was made hydrophilic by a brief incubation in methanol followed by an equilibration step in transfer buffer. The transfer was done using a wet-transfer system (BioRAD) at 4°C for 16h at 30 V.

After the transfer, the membrane was incubated in blocking solution (3 % non-fat milk powder dissolved in PBS/0.05% Tween20) for 1 hr at RT or o.n. at 4 °C. Primary and secondary antibodies were diluted in blocking solution to the appropriate concentrations. The enhanced chemiluminescence system (GE Healthcare) was used, followed by exposure to X-ray film. Between the antibody incubations, the membrane was washed five times in PBS/0.05% Tween20. For detection, horseradish peroxidase (HRP) conjugated secondary antibodies were used.

7.2.3.3 Protein precipitation using acetone

10 ml of conditioned cell culture supernatant were added to 40 ml of ice-cold acetone in a 50 ml polypropylene falcon tube. The solution was shaken vigorously and incubated o.n. at -20°C. At this step samples can be kept for longer periods at -20°C. Precipitates were collected by a centrifugation step (10.000 rpm, 4°C, 20 minutes) and after removal of the supernatant the pellet was left to dry for 30 – 60 minutes at RT. For Western blot analysis the pellet was directly resuspended in sample buffer (for instance, 2x Tricine buffer), boiled for 5 minutes at 95°C and loaded on a protein gel.

7.2.3.4 Immunoprecipitation (using M2 anti-FLAG antibody)

Immunoprecipitations were carried out on cell lysates as well as on cultures supernatants. To prepare cell lysates the cells were incubated for 30 minutes on ice in IP lysis buffer.All following manipulations were carried out at 4°C. 5-10 µg of M2 anti-FLAG Ab (from Sigma or Stratagene) were added to samples (1 ml of cell lysate or supernatant) and incubated for at least 1h. Then 50µl of protein G sepharose beads (GE Healthcare) were added to the sample and put for 1h on a rotating wheel. The precipitated beads were washed with PBS several times before they were directly resuspended in sample buffer (2x tricine buffer).

7.2.3.5 Lysis of tissues and cells

For further processing of lysates, e.g. affinity-chromatography, fresh tissues were shock frozen and routinely lysed in ice-cold HEPES-lysis buffer (HLB). Decomposition was achieved by 30 strokes at 800 rpm in a teflon-pestle Dounce- homogenizer in the cold. After centrifugation at 60.000 rpm at 4°C in a TLA 120.2 rotor (Beckman), the clear supernatant was used for further experiments.

Tissue culture cells were generally lysed in 1xSDS sample buffer in the culture dish after several washes with PBS.

7.2.3.6 Explant assay

Albino Wistar female rats (250–350 g; Charles River) were sacrificed by inspiration of 75% CO2, and stomach fundus muscles strips were isolated, washed in fresh Tyrode's solution, mounted vertically in a 5-mL organ bath in oxygenated (95% O_2, 5%CO_2) Tyrode's solution, and maintained at 37°C. The segments were stretched to a tension of 2.0 g and allowed to equilibrate for 30–60 min, with the superfusion buffer changed every 15–20 min. At the beginning of each experiment, acetylcholine chloride (ACh 10^{-5} M) was applied to achieve a maximal control contraction. The potency of contractions was recorded isometrically by a strain gauge transducer (DY 1; Ugo Basile) and displayed on a recording microdynamometer (Unirecord; Ugo Basile). When reproducible responses to ACh were obtained, increasing concentrations (from 10^{-9} to 10^{-5} M) of synthetic amidated spexin peptide (NWTPQAMLYLKGAQ-amide; Primm) were applied every 2 min to establish a cumulative dose-response curve followed by washing and recovery for minimum 20 min. The EC50 was calculated by interpolation from the cumulative dose-response curve. Consecutively, single doses of spexin 10^{-6} M were applied until reproducible responses were obtained.

7.2.4 Cell biology methods

7.2.4.1 General conditions of cell culture and sterilisation

The following cell culture methods were accomplished in a sterile environment using a Laminar Flow Hood. Pancreatic cells were kept in incubators at 37°C in the presence of 7.5 % CO_2 (pancreatic cells have a high metabolism). Solutions and plastic materials were autoclaved for 20 minutes at 135°C and 2.2 bar pressure prior to use. Glass devices were sterilized for 3 hrs at 180°C.

7.2.4.2 Thawing and freezing of eukaryotic cells

Eukaryotic cells can be stored for a long time in liquid nitrogen (-196°C) and after thawing put into culture. To prevent formation of ice crystals inside the cells 10% DMSO was added to the freezing medium. At least $5x10^6$ cells/ml were transferred into cryo-tubes and cooled to -80°C, before they were stored into liquid nitrogen.

After thawing cryo-conserved cells, they were washed twice in medium to remove the toxic DMSO and then seeded for culture.

7.2.4.3 Growing conditions of pancreatic cells

βTC-3 are adherent cells growing as multilayers. The growing medium used was: 82.5% Dulbecco's MEM (4.5 g/L glucose) + 15% horse serum + 2.5% FBS. Cells were split at confluency 1:5 to 1:10 every 2-5 days by mechanical disruption or a low amount of trypsin. Cells were incubated at 37 °C with 7.5% CO_2. The doubling time of βTC-3 cells is approximately 34 hours. Cells were frozen with 70% medium+20% FBS+10% DMSO.

7.2.4.4 Transfection of eukaryotic cells with Lipofectamine 2000

Transfection of Beta-TC3 and RINm5f were routinely done in 10-cm plate using the Lipofectamine reagent (Invitrogen) following the manufacturer's instructions. Briefly, 24 μg of plasmid DNA were diluted in 1.5 ml serum-free medium. 50 μl of Lipofectamine 2000 were mixed into another 1.5 ml serum-free medium and incubate for 5 minutes at RT. After combination of the two solutions, DNA-Lipofectamine complexes were allowed to be formed for 30 minutes at RT. The mix was added drop by drop to the cells. After 4-6 hours the transfection medium was replaced with new one to ensure a high viability of the cells. For smaller transfections reagents were scaled down proportionally.

7.2.4.5 Immunofluorescence staining

Cultured cells were plated on cover slips and washed twice with PBS prior to fixation in 4% PFA/PBS for 20 min. Cells were then permeabilized for 10 min in 0.2 % TX-100/PBS and washed in 50 mM glycine in PBT (PBS/0.05% Tween20), and incubated in blocking buffer for 1 hr. Primary antibodies were diluted in blocking buffer and incubated for 1 hr in a wet chamber (for M2 and anti-insulin Ab~ 10 μg/ml). After washing in PBT, fluorescence-conjugated secondary antibodies were allowed to bind for 30 min in the dark diluted in blocking buffer. Nuclei were stained by including a dilution of Hoechst 33342 in PBT in one of the final washing steps. After briefly dipping in water the cover slips were mounted on glass slides in gelvatol.

7.2.5 Methods in histology

7.2.5.1 Paraffin embedding of mouse tissues

Tissues were fixed in 4% PFA at 4°C o.n. The next day tissues were washed for at least 1h in PBS and subsequently dehydrated by incubation increasing ethanol solutions at 4°C (i.e., 2 x 30 min in 50% EtOH/dH_2O, 2 x 30 min in 70% EtOH/dH_2O, 2 x 30 min in 96% EtOH/dH_2O, 100% ethanol for 1 hr). Tissues were brought to RT and transferred into xylene, which was exchanged after 30 minutes. The xylene was then replaced by a xylene:paraffin mix (1:1) and incubated at 58-60° C for 30 min, followed by several incubations in pure paraffin at 58-60° C. The next day tissues were transferred into molds with paraffin, orientated in the desired position and left to harden at RT. 6-8 µm sections were with a microtome and mounted on Superfrost Plus slides. Sections were dried in a 42°C oven o.n.

7.2.5.2 Immunohistochemistry

Immunohistochemistry was used to detect spexin and augurin proteins in mouse tissues. Briefly, paraffin sections were dewaxed and rehydrated by incubation in xylene and descending dilutions of ethanol (100% - 50%). After an equilibration in TBS, antigen removal was achieved by boiling the slides for 5 minutes in citrate buffer. After blocking of free protein binding sides, the slides were incubated with the primary antibodys o.n. In order to obtain a more sensitive staining The DAB development system was used, in which a biotinylated secondary antibody builds a second layer on the sample, before detection and color precipitation is done using Avidin-HRP. The slides were counterstained and mounted with xylene containing medium.

7.2.5.3 *In situ* hybridization

Tissues and E18.5 embryos were dissected, fixed overnight in 4% paraformaldehyde, and embedded in paraffin. *In situ* hybridization using digoxigenin-labeled or 35S-CTP-labeled probes on 8-µm paraffin sections was performed according to procedures previously described (Neubuser et al. 1995; Niederreither and Dolle 1998). Briefly, sections were dewaxed, rehydrated, digested with proteinase K, and hybridized with probe at 65°C. Posthybridization washes in 20% formamide, 0.5% SSC were done at 60°C. The spexin probe was a 0.3-kb cDNA fragment cloned from mouse brain RNA:

spexin-5': 5'-ACAGGGTCGGAACATGAAGGG-3'
spexin-3': 5'-AAGAGTCTGTCTTCCAAGAGTTCGC-3'

The augurin probe was a 0.4-kb fragment amplified from mouse adrenal RNA:

augurin-5': 5'-CACCATGAGCACCTCGTCTGCG-3'
augurin-3': 5'-TCTGTGGGCACCTCAGGG-3-

8 Resources

8.1 Protein Databases and genome browsers

8.1.1 Ensembl

Ensembl (http://www.ensembl.org/index.html), (Hubbard et al. 2006) is a project which aims to gather information about Eukaryotic genome sequences and develops tools to automatically annotate those genomes. For any given gene, Ensembl provides information about orthologs genes in other organisms. The number of (especially vertebrate and mammalian) organisms made available and annotated by Ensembl has greatly increased over the last 3-4 years. I have made extensive use of this resource throughout my doctoral studies.

8.1.2 SwissProt

Swissprot/Trembl is the largest public database/portal dedicated to protein annotation. The Sequence Retrieval System (SRS, http://expasy.org/srs5/), (Boeckmann et al. 2003) is a technology which allows indexing and querying of large amount of Swissprot/Trembl data. This was my main source of publicly available data, along with Ensembl and without those two resources I believe carrying out of my project would not have been feasible.

8.1.3 UCSC genome browser

The UCSC genome browser (http://genome.ucsc.edu), (Karolchik et al. 2003) is a genome browser developed at the University of California Santa Cruz and many whole genome alignments that I have used to gather orthologs proteins from my candidate PH

8.2 Gene expression databases

8.2.1 Allen Brain Atlas

This recently completed project (http://www.brain-map.org/welcome.do), (Lein et al. 2007) provides high quality *in situ* hybridization data on coronal and sagittal sections of the mouse brain for a large number of genes. The visualisation of the data in the desired region of the brain is aided by a great interactive navigation. This database has greatly helped me as it provided important clues on where spexin, augurin and other candidate PH are expressed in the mammalian brain.

8.2.2 Symatlas

The SymAtlas project (http://symatlas.gnf.org/SymAtlas/), (Walker and Wiltshire 2006) provides expression data by microarrays of nearly all known transcripts in the genome, for a large collection of tissues. I have made extensive use of this resource during my project.

8.2.3 BrainInfo

Braininfo (http:/braininfo.rprc.washington.edu) is a publicly-available database where one can find structures of the brain where specific genes are expressed or genes expressed in a specific structure of the brain

8.3 Protein features prediction tools

8.3.1 SignalP

SignalP is a publicly-available program which identifies signal peptides in protein sequences (http://www.cbs.dtu.dk/services/SignalP/), (Kall et al. 2004).

8.3.2 ProtParam

ProtParam (http://expasy.org/cgi-bin/protparam), a program part of the Expasy toolkit, was used to compute the theoretical molecular weight of proteins, using their primary sequence as input.

8.3.3 ProP

ProP (http://www.cbs.dtu.dk/services/ProP/), (Duckert et al. 2004) is a publicly-available program which predicts prohormone/furin cleavage sites in a protein sequence.

8.3.4 Neuropred

Neuropred (http://neuroproteomics.scs.uiuc.edu/cgi-bin/neuropred.py), (Southey et al. 2006) is a freely-available program which predicts prohormone convertases cleavage sites such as those found in neuropeptide precursor sequences.

8.3.5 Eukayotic Linear Motif resource (ELM)

The ELM database (http://elm.eu.org/), (Puntervoll et al. 2003) predicts short linear motifs in a protein sequence, such as proteolytic cleavage sites.

8.4 Other softwares

8.4.1 Clustalw

ClustalW is the program I used to align multiple sequences (Chenna et al. 2003).

8.4.2 Jalview

Jalview is a software very useful to visualise multiple alignments (Clamp et al. 2004).

8.5 Biological material resources

8.5.1 German resource centre for biological material (DSMZ)

The β-TC3 cell line was ordered from the German resource centre for biological material (www.dsmz.de).

8.5.2 German resource center for genome research (RZPD)

Several candidate PH-containing plasmids were ordered from the German resource center for genome research.

9 Bibliography

Abney, S. 1996. Part-of-Speech Tagging and Partial Parsing.

Ahima, R.S., J. Dushay, S.N. Flier, D. Prabakaran, and J.S. Flier. 1997. Leptin accelerates the onset of puberty in normal female mice. *J Clin Invest* **99:** 391-395.

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. 1989. *Molecular biology of the cell*. Garland Publishing, New York.

Alcolado, J.C., I.E. Moore, and R.O. Weller. 1986. Calcification in the human choroid plexus, meningiomas and pineal gland. *Neuropathol Appl Neurobiol* **12:** 235-250.

Alfonsi, E., M. Versino, I.M. Merlo, C. Pacchetti, E. Martignoni, G. Bertino, A. Moglia, C. Tassorelli, and G. Nappi. 2007. Electrophysiologic patterns of oral-pharyngeal swallowing in parkinsonian syndromes. *Neurology* **68:** 583-589.

Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* **219:** 555-565.

Alvarez-Buylla, A. and J.M. Garcia-Verdugo. 2002. Neurogenesis in adult subventricular zone. *J Neurosci* **22:** 629-634.

Apweiler, R., H. Hermjakob, and N. Sharon. 1999. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* **1473:** 4-8.

Asai, K., S. Hayamizu, and K. Handa. 1993. Prediction of protein secondary structure by the hidden Markov model. *Comput Appl Biosci* **9:** 141-146.

Baas, D., A. Meiniel, C. Benadiba, E. Bonnafe, O. Meiniel, W. Reith, and B. Durand. 2006. A deficiency in RFX3 causes hydrocephalus associated with abnormal differentiation of ependymal cells. *Eur J Neurosci* **24:** 1020-1030.

Baker, H., N. Liu, H.S. Chun, S. Saino, R. Berlin, B. Volpe, and J.H. Son. 2001. Phenotypic differentiation during migration of dopaminergic progenitor cells to the olfactory bulb. *J Neurosci* **21:** 8505-8513.

Baldi, P. and Y. Chauvin. 1994. Hidden Markov Models of the G-protein-coupled receptor family. *J Comput Biol* **1:** 311-336.

Baldi, P., Y. Chauvin, T. Hunkapiller, and M.A. McClure. 1994. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci U S A* **91:** 1059-1063.

Ballabh, P., A. Braun, and M. Nedergaard. 2004. The blood-brain barrier: an overview: structure, regulation, and clinical implications. *Neurobiol Dis* **16:** 1-13.

Bardo, M.T. 1998. Neuropharmacological mechanisms of drug reward: beyond dopamine in the nucleus accumbens. *Crit Rev Neurobiol* **12:** 37-67.

Bateman, A. and C. Chothia. 1996. Fibronectin type III domains in yeast detected by a hidden Markov model. *Curr Biol* **6:** 1544-1547.

Baum, L. and J. Eagon. 1967. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73:** 360-363.

Baum, L.E. and T. Petrie. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **37:** 1554-1563.

Baum, L.E., T. Petrie, G. Soules, and N. Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions
of Markov chains. *Ann. Math. Stat.*, **41:** 164-171.

Ben-Hur, T., B. Rogister, K. Murray, G. Rougon, and M. Dubois-Dalcq. 1998. Growth and fate of PSA-NCAM+ precursors of the postnatal brain. *J Neurosci* **18:** 5777-5788.

Bendtsen, J.D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340:** 783-795.

Bhatnagar, S., V. Viau, A. Chu, L. Soriano, O.C. Meijer, and M.F. Dallman. 2000. A cholecystokinin-mediated pathway to the paraventricular thalamus is recruited in chronically stressed rats and regulates hypothalamic-pituitary-adrenal function. *J Neurosci* **20:** 5564-5573.

Biancani, P., J.H. Walsh, and J. Behar. 1984. Vasoactive intestinal polypeptide. A neurotransmitter for lower esophageal sphincter relaxation. *J Clin Invest* **73:** 963-967.

Birney, E., A. Bateman, M.E. Clamp, and T.J. Hubbard. 2001. Mining the draft human genome. *Nature* **409:** 827-828.

Birney, E. and R. Durbin. 2000. Using GeneWise in the Drosophila annotation experiment. *Genome Res* **10:** 547-548.

Boeckmann, B., A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31:** 365-370.

Bolle, P., C. Severini, G. Falconieri-Erspamer, P. Tucci, and V. Erspamer. 2000. Effects of natural tachykinins on porcine lower urinary tract smooth muscle. *J Auton Pharmacol* **20:** 157-161.

Boon, A., E. Cheriex, J. Lodder, and F. Kessels. 1997. Cardiac valve calcification: characteristics of patients with calcification of the mitral annulus or aortic valve. *Heart* **78:** 472-474.

Breiteneder-Geleff, S., A. Soleiman, H. Kowalski, R. Horvat, G. Amann, E. Kriehuber, K. Diem, W. Weninger, E. Tschachler, K. Alitalo, and D. Kerjaschki. 1999. Angiosarcomas express mixed endothelial phenotypes of blood and lymphatic capillaries: podoplanin as a specific marker for lymphatic endothelium. *Am J Pathol* **154:** 385-394.

Brownstein, M.J. 1993. A brief history of opiates, opioid peptides, and opioid receptors. *Proc Natl Acad Sci U S A* **90:** 5391-5393.

Bucher, P., K. Karplus, N. Moeri, and K. Hofmann. 1996. A flexible motif search technique based on generalized profiles. *Comput Chem* **20:** 3-23.

Bullock, J.D., R.J. Campbell, and R.R. Waller. 1977. Calcification in retinoblastoma. *Invest Ophthalmol Vis Sci* **16:** 252-255.

Burgess, T.L. and R.B. Kelly. 1987. Constitutive and regulated secretion of proteins. *Annu Rev Cell Biol* **3:** 243-293.

Camproux, A.C., R. Gautier, and P. Tuffery. 2004. A hidden markov model derived structural alphabet for proteins. *J Mol Biol* **339:** 591-605.

Caputo, A., L. Ghiringhelli, M. Dieci, G.M. Giobbio, F. Tenconi, L. Ferrari, E. Gimosti, K. Prato, and A. Vita. 1998. Epithalamus calcifications in schizophrenia. *Eur Arch Psychiatry Clin Neurosci* **248:** 272-276.

Carlen, M., R.M. Cassidy, H. Brismar, G.A. Smith, L.W. Enquist, and J. Frisen. 2002. Functional integration of adult-born neurons. *Curr Biol* **12:** 606-608.

Castaneyra-Perdomo, A., E. Carmona-Calero, G. Meyer, H. Perez-Gonzalez, M.M. Perez-Delgado, N. Marrero-Gordillo, S. Rodriguez, and E.M. Rodriguez. 1998. Changes in the secretory activity of the subcommissural organ of spontaneously hypertensive rats. *Neurosci Lett* **246:** 133-136.

Castaneyra-Perdomo, A., G. Meyer, E. Carmona-Calero, J. Banuelos-Pineda, R. Mendez-Medina, C. Ormazabal-Ramos, and R. Ferres-Torres. 1994. Alterations of the subcommissural organ in the hydrocephalic human fetal brain. *Brain Res Dev Brain Res* **79:** 316-320.

Cecilia, D., C. Piero, P. Luisa, B. Gabriele, and B. Cristiano. 2006. Receptors for leptin and estrogen in the subcommissural organ of rabbits are differentially modulated by fasting. *Brain Res* **1124:** 62-69.

Chartrel, N., C. Dujardin, Y. Anouar, J. Leprince, A. Decker, S. Clerens, J.C. Do-Rego, F. Vandesande, C. Llorens-Cortes, J. Costentin, J.C. Beauvillain, and H. Vaudry. 2003. Identification of 26RFa, a hypothalamic neuropeptide of the RFamide peptide family with orexigenic activity. *Proc Natl Acad Sci U S A* **100:** 15247-15252.

Chehab, F.F., K. Mounzih, R. Lu, and M.E. Lim. 1997. Early onset of reproductive function in normal female mice treated with leptin. *Science* **275:** 88-90.

Chemelli, R.M., J.T. Willie, C.M. Sinton, J.K. Elmquist, T. Scammell, C. Lee, J.A. Richardson, S.C. Williams, Y. Xiong, Y. Kisanuki, T.E. Fitch, M. Nakazato, R.E. Hammer, C.B. Saper, and M. Yanagisawa. 1999. Narcolepsy in orexin knockout mice: molecular genetics of sleep regulation. *Cell* **98:** 437-451.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T.J. Gibson, D.G. Higgins, and J.D. Thompson. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31:** 3497-3500.

Chiasson, B.J., V. Tropepe, C.M. Morshead, and D. van der Kooy. 1999. Adult mammalian forebrain ependymal and subependymal cells demonstrate proliferative potential, but only subependymal cells have neural stem cell characteristics. *J Neurosci* **19:** 4462-4471.

Chu, W., Z. Ghahramani, A. Podtelezhnikov, and D.L. Wild. 2006. Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. *IEEE/ACM Trans Comput Biol Bioinform* **3:** 98-113.

Civelli, O. 1998. Functional genomics: the search for novel neurotransmitters and neuropeptides. *FEBS Lett* **430:** 55-58.

Civelli, O., H.P. Nothacker, and R. Reinscheid. 1998. Reverse physiology: discovery of the novel neuropeptide, orphanin FQ/nociceptin. *Crit Rev Neurobiol* **12:** 163-176.

Clamp, M., J. Cuff, S.M. Searle, and G.J. Barton. 2004. The Jalview Java alignment editor. *Bioinformatics* **20:** 426-427.

Clave, P., A. Gonzalez, A. Moreno, R. Lopez, A. Farre, X. Cusso, M. D'Amato, F. Azpiroz, and F. Lluis. 1998. Endogenous cholecystokinin enhances postprandial gastroesophageal reflux in humans through extrasphincteric receptors. *Gastroenterology* **115:** 597-604.

Cohen, S. and W. Lipshutz. 1971. Hormonal regulation of human lower esophageal sphincter competence: interaction of gastrin and secretin. *J Clin Invest* **50:** 449-454.

Collett, G.D. and A.E. Canfield. 2005. Angiogenesis and pericytes in the initiation of ectopic calcification. *Circ Res* **96:** 930-938.

Conlon, J.M., C.F. Deacon, L. Grimelius, B. Cedermark, R.F. Murphy, L. Thim, and W. Creutzfeldt. 1988. Neuropeptide K-(1-24)-peptide: storage and release by carcinoid tumors. *Peptides* **9:** 859-866.

Connell, J.M. and E. Davies. 2005. The new biology of aldosterone. *J Endocrinol* **186:** 1-20.

Coskun, V. and M.B. Luskin. 2002. Intrinsic and extrinsic regulation of the proliferation and differentiation of cells in the rodent rostral migratory stream. *J Neurosci Res* **69:** 795-802.

Cota, D., K. Proulx, K.A. Smith, S.C. Kozma, G. Thomas, S.C. Woods, and R.J. Seeley. 2006. Hypothalamic mTOR signaling regulates food intake. *Science* **312:** 927-930.

Cottrell, G.T. and A.V. Ferguson. 2004. Sensory circumventricular organs: central roles in integrated autonomic regulation. *Regul Pept* **117:** 11-23.

Crawley, J.N. and R.L. Corwin. 1994. Biological actions of cholecystokinin. *Peptides* **15:** 731-755.

Creamer, B. and J. Schlegel. 1957. Motor responses of the esophagus to distention. *J Appl Physiol* **10:** 498-504.

Cumha-Vaz, J.G. 1978. The blood-ocular barriers. *Invest Ophthalmol Vis Sci* **17:** 1037-1039.

Davies, K.P., M. Tar, C. Rougeot, and A. Melman. 2007. Sialorphin (the mature peptide product of Vcsa1) relaxes corporal smooth muscle tissue and increases erectile function in the ageing rat. *BJU Int* **99:** 431-435.

De Loupy, C., M. El-Beze, and P.F. Marteau. 1998. Word Sense Disambiguation using HMM Tagger. *Actes de The First International Conference on Language Resources & Evaluation***:** 1255-1258.

Dent, J., W.J. Dodds, R.H. Friedman, T. Sekiguchi, W.J. Hogan, R.C. Arndorfer, and D.J. Petrie. 1980. Mechanism of gastroesophageal reflux in recumbent asymptomatic human subjects. *J Clin Invest* **65:** 256-267.

Devos, R., J.G. Richards, L.A. Campfield, L.A. Tartaglia, Y. Guisez, J. van der Heyden, J. Travernier, G. Plaetinck, and P. Burn. 1996. OB protein binds specifically to the choroid plexus of mice and rats. *Proc Natl Acad Sci U S A* **93:** 5668-5673.

Dey, A., X. Xhu, R. Carroll, C.W. Turck, J. Stein, and D.F. Steiner. 2003. Biological processing of the cocaine and amphetamine-regulated transcript precursors by prohormone convertases, PC2 and PC1/3. *J Biol Chem* **278:** 15007-15014.

Duckert, P., S. Brunak, and N. Blom. 2004. Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel* **17:** 107-112.

Duguay, S.J., J. Lai-Zhang, and D.F. Steiner. 1995. Mutational analysis of the insulin-like growth factor I prohormone processing site. *J Biol Chem* **270:** 17566-17574.

Dundore, R.L., J.N. Wurpel, C.D. Balaban, T.S. Harrison, L.C. Keil, J.F. Seaton, and W.B. Severs. 1987. Site-dependent central effects of aldosterone in rats. *Brain Res* **401:** 122-131.

Dundore, R.L., J.N. Wurpel, C.D. Balaban, L.C. Keil, and W.B. Severs. 1984. Central effects of aldosterone infused into the rat subcommissural organ region. *Neurosci Res* **1:** 341-351.

Durbin, R., S.R. Eddy, A. Krogh, and G. Mitchison. 1998. *Biological sequence analysis:probabilistic models of proteins and nucleic acids*, Cambridge, UK.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755-763.

Eipper, B.A., D.A. Stoffers, and R.E. Mains. 1992. The biosynthesis of neuropeptides: peptide alpha-amidation. *Annu Rev Neurosci* **15:** 57-85.

Emanuelsson, O., H. Nielsen, S. Brunak, and G. von Heijne. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300:** 1005-1016.

Emerich, D.F., S.J. Skinner, C.V. Borlongan, A.V. Vasconcellos, and C.G. Thanos. 2005. The choroid plexus in the rise, fall and repair of the brain. *Bioessays* **27:** 262-274.

Fallon, J., S. Reid, R. Kinyamu, I. Opole, R. Opole, J. Baratta, M. Korc, T.L. Endo, A. Duong, G. Nguyen, M. Karkehabadhi, D. Twardzik, S. Patel, and S. Loughlin. 2000. In vivo induction of massive proliferation, directed migration, and differentiation of neural cells in the adult mammalian brain. *Proc Natl Acad Sci U S A* **97:** 14686-14691.

Fariselli, P., P.L. Martelli, and R. Casadio. 2005. A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics* **6 Suppl 4:** S12.

Felsenstein, J. and G.A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol* **13:** 93-104.

Fernandez-Llebrez, P., J.M. Grondona, J. Perez, M.F. Lopez-Aranda, G. Estivill-Torrus, P.F. Llebrez-Zayas, E. Soriano, C. Ramos, Y. Lallemand, A. Bach, and B. Robert. 2004. Msx1-

deficient mice fail to form prosomere 1 derivatives, subcommissural organ, and posterior commissure and develop hydrocephalus. *J Neuropathol Exp Neurol* **63:** 574-586.

Fitzsimons, J.T. 1998. Angiotensin, thirst, and sodium appetite. *Physiol Rev* **78:** 583-686.

Flier, J.S., M. Harris, and A.N. Hollenberg. 2000. Leptin, nutrition, and the thyroid: the why, the wherefore, and the wiring. *J Clin Invest* **105:** 859-861.

Foeldvari, I.P. and M. Palkovits. 1964. Effect Of Sodium And Potassium Restriction On The Functional Morphology Of The Subcommissural Organ. *Nature* **202:** 905-906.

Freitag, D. and McCallum. 1999. Information extraction using hmms and shrinkage. In *Papers from the AAAI-99 Workshop on Machine Learning for Information Extraction*, pp. 31-36. AAAI., Menlo Park, California.

Fuchs, E., T. Tumbar, and G. Guasch. 2004. Socializing with the neighbors: stem cells and their niche. *Cell* **116:** 769-778.

Fuller, R.S., A.J. Brake, and J. Thorner. 1989. Intracellular targeting and structural conservation of a prohormone-processing endoprotease. *Science* **246:** 482-486.

Futami, T., K. Takakusaki, and S.T. Kitai. 1995. Glutamatergic and cholinergic inputs from the pedunculopontine tegmental nucleus to dopamine neurons in the substantia nigra pars compacta. *Neurosci Res* **21:** 331-342.

Gage, F.H., P.W. Coates, T.D. Palmer, H.G. Kuhn, L.J. Fisher, J.O. Suhonen, D.A. Peterson, S.T. Suhr, and J. Ray. 1995. Survival and differentiation of adult neuronal progenitor cells transplanted to the adult brain. *Proc Natl Acad Sci U S A* **92:** 11879-11883.

Ganong, W.F. 2006. *Review of Medical Physiology, 22nd edition*.

Gartner, L.P. and J.L. Hiatt. 2001. *Color textbook of histology*. Elsevier.

Gascuel, O. and A. Danchin. 1986. Protein export in prokaryotes and eukaryotes: indications of a difference in the mechanism of exportation. *J Mol Evol* **24:** 130-142.

Geerling, J.C., M. Kawata, and A.D. Loewy. 2006. Aldosterone-sensitive neurons in the rat central nervous system. *J Comp Neurol* **494:** 515-527.

Ghiani, P., B. Uva, M. Vallarino, A. Mandich, and M.A. Masini. 1988. Angiotensin II specific receptors in subcommissural organ. *Neurosci Lett* **85:** 212-216.

Gibson, T.R., G.M. Wildey, S. Manaker, and C.C. Glembotski. 1986. Autoradiographic localization and characterization of atrial natriuretic peptide binding sites in the rat central nervous system and adrenal gland. *J Neurosci* **6:** 2004-2011.

Gilbert, G.J. 1963. Renal effect of subcommissural extract. *Neurology* **13:** 43-55.

Gillick, L., Y. Ito, L. Manganaro, M. Newman, F. Scattone, S. Wegmann, J. Yamron, and P. Zhan. 1998. Dragon systems' automatic transcription of new TDT corpus. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*.

Glasow, A., A. Haidan, U. Hilbers, M. Breidert, J. Gillespie, W.A. Scherbaum, G.P. Chrousos, and S.R. Bornstein. 1998. Expression of Ob receptor in normal human adrenals: differential regulation of adrenocortical and adrenomedullary function by leptin. *J Clin Endocrinol Metab* **83:** 4459-4466.

Gobron, S., H. Monnerie, R. Meiniel, I. Creveaux, W. Lehmann, D. Lamalle, B. Dastugue, and A. Meiniel. 1996. SCO-spondin: a new member of the thrombospondin family secreted by the subcommissural organ is a candidate in the modulation of neuronal aggregation. *J Cell Sci* **109 (Pt 5):** 1053-1061.

Godeau, G. and A.M. Robert. 1979. Mechanism of action of collagenase on the blood-brain barrier permeability. Increase of endothelial cell pinocytotic activity as shown with horse-radish peroxidase as a tracer. *Cell Biol Int Rep* **3:** 747-751.

Grimsholm, O., S. Rantapaa-Dahlqvist, and S. Forsgren. 2005. Levels of gastrin-releasing peptide and substance P in synovial fluid and serum correlate with levels of cytokines in rheumatoid arthritis. *Arthritis Res Ther* **7:** R416-426.

Groisman, G.M., M. Amar, and S. Polak-Charcon. 1999. Microcalcifications in the anterior pituitary gland of the fetus and the newborn: a histochemical and immunohistochemical study. *Hum Pathol* **30:** 199-202.

Groothuis, D.R. and R.M. Levy. 1997. The entry of antiviral and antiretroviral drugs into the central nervous system. *J Neurovirol* **3:** 387-400.

Gross, R.E., M.F. Mehler, P.C. Mabie, Z. Zang, L. Santschi, and J.A. Kessler. 1996. Bone morphogenetic proteins promote astroglial lineage commitment by mammalian subventricular zone progenitor cells. *Neuron* **17:** 595-606.

Gutierrez, J.G., K.D. Thanik, W.Y. Chey, and H. Yajima. 1977. Effect of motilin on the lower esophageal sphincter of the opossum. *Am J Dig Dis* **22:** 402-405.

Hadnagy, A., L. Gaboury, R. Beaulieu, and D. Balicki. 2006. SP analysis may be used to identify cancer stem cell populations. *Exp Cell Res* **312:** 3701-3710.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Element of Statistical Learning*. Springer, New York.

Hauser, F., G. Cazzamali, M. Williamson, Y. Park, B. Li, Y. Tanaka, R. Predel, S. Neupert, J. Schachtner, P. Verleyen, and C.J. Grimmelikhuijzen. 2008. A genome-wide inventory of neurohormone GPCRs in the red flour beetle Tribolium castaneum. *Front Neuroendocrinol* **29:** 142-165.

Heinrichs, S.C., B.J. Cole, E.M. Pich, F. Menzaghi, G.F. Koob, and R.L. Hauger. 1992. Endogenous corticotropin-releasing factor modulates feeding induced by neuropeptide Y or a tail-pinch stressor. *Peptides* **13:** 879-884.

Henke, W., K. Herdel, K. Jung, D. Schnorr, and S.A. Loening. 1997. Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Res* **25:** 3957-3958.

Herkenham, M. and W.J. Nauta. 1979. Efferent connections of the habenular nuclei in the rat. *J Comp Neurol* **187:** 19-47.

Hewes, R.S. and P.H. Taghert. 2001. Neuropeptides and neuropeptide receptors in the Drosophila melanogaster genome. *Genome Res* **11:** 1126-1142.

Hirsch, E.C., A.M. Graybiel, C. Duyckaerts, and F. Javoy-Agid. 1987. Neuronal loss in the pedunculopontine tegmental nucleus in Parkinson disease and in progressive supranuclear palsy. *Proc Natl Acad Sci U S A* **84:** 5976-5980.

Holash, J.A., S.I. Harik, G. Perry, and P.A. Stewart. 1993. Barrier properties of testis microvessels. *Proc Natl Acad Sci U S A* **90:** 11069-11073.

Horvath, T.L., S. Diano, and A.N. van den Pol. 1999. Synaptic interaction between hypocretin (orexin) and neuropeptide Y cells in the rodent and primate hypothalamus: a novel circuit implicated in metabolic and endocrine regulations. *J Neurosci* **19:** 1072-1087.

Hubbard, T.J.P., B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S.C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X.M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. 2006. Ensembl 2007. *Nucl. Acids Res.***:** gkl996.

Hughey, R. and A. Krogh. 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput Appl Biosci* **12:** 95-107.

Hummon, A.B., T.A. Richmond, P. Verleyen, G. Baggerman, J. Huybrechts, M.A. Ewing, E. Vierstraete, S.L. Rodriguez-Zas, L. Schoofs, G.E. Robinson, and J.V. Sweedler. 2006. From the genome to the proteome: uncovering peptides in the Apis brain. *Science* **314:** 647-649.

Hunter, P.C., J. Crameri, S. Austin, M.C. Woodward, and A.J. Hughes. 1997. Response of parkinsonian swallowing dysfunction to dopaminergic stimulation. *J Neurol Neurosurg Psychiatry* **63:** 579-583.

Ivy, A.C. and E. Oldberg. 1928. A hormone mechanism for gall-bladder contraction and evacuation. *American Journal of Physiology* **86:** 599-613.

Ji, H. and P.D. Shepard. 2007. Lateral habenula stimulation inhibits rat midbrain dopamine neurons through a GABA(A) receptor-mediated mechanism. *J Neurosci* **27:** 6923-6930.

Johanson, C., P. McMillan, R. Tavares, A. Spangenberger, J. Duncan, G. Silverberg, and E. Stopa. 2004. Homeostatic capabilities of the choroid plexus epithelium in Alzheimer's disease. *Cerebrospinal Fluid Res* **1:** 3.

Johnson, A.K. and P.M. Gross. 1993. Sensory circumventricular organs and brain homeostatic pathways. *Faseb J* **7:** 678-686.

Julius, D., A. Brake, L. Blair, R. Kunisawa, and J. Thorner. 1984. Isolation of the putative structural gene for the lysine-arginine-cleaving endopeptidase required for processing of yeast prepro-alpha-factor. *Cell* **37:** 1075-1089.

Kall, L., A. Krogh, and E.L. Sonnhammer. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338:** 1027-1036.

Kall, L., A. Krogh, and E.L. Sonnhammer. 2005. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* **21 Suppl 1:** i251-257.

Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent. 2003. The UCSC Genome Browser Database. *Nucl. Acids Res.* **31:** 51-54.

Kataoka, Y., Y. Ikehara, and T. Hattori. 1996. Cell proliferation and renewal of mouse adrenal cortex. *J Anat* **188 (Pt 2):** 375-381.

Kayaba, Y., A. Nakamura, Y. Kasuya, T. Ohuchi, M. Yanagisawa, I. Komuro, Y. Fukuda, and T. Kuwaki. 2003. Attenuated defense response and low basal blood pressure in orexin knockout mice. *Am J Physiol Regul Integr Comp Physiol* **285:** R581-593.

Kenny, P., M. Lennig, and P. Mermelstein. 1990. A linear predictive HMM for vector-valued observations withapplications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.* **32:** 220-225.

Kerever, A., J. Schnack, D. Vellinga, N. Ichikawa, C. Moon, E. Arikawa-Hirasawa, J.T. Efird, and F. Mercier. 2007. Novel Extracellular Matrix Structures in the Neural Stem Cell Niche Capture the Neurogenic Factor FGF-2 from the Extracellular Milieu. *Stem Cells*.

Khoo, M.L., S.L. Asa, I.J. Witterick, and J.L. Freeman. 2002. Thyroid calcification and its association with thyroid carcinoma. *Head Neck* **24:** 651-655.

Kiess, W., W.F. Blum, and M.L. Aubert. 1998. Leptin, puberty and reproductive function: lessons from animal studies and observations in humans. *Eur J Endocrinol* **138:** 26-29.

Kim, C.F., E.L. Jackson, A.E. Woolfenden, S. Lawrence, I. Babar, S. Vogel, D. Crowley, R.T. Bronson, and T. Jacks. 2005. Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell* **121:** 823-835.

Kirschenbaum, B., F. Doetsch, C. Lois, and A. Alvarez-Buylla. 1999. Adult subventricular zone neuronal precursors continue to proliferate and migrate in the absence of the olfactory bulb. *J Neurosci* **19:** 2171-2180.

Kojima, M., H. Hosoda, Y. Date, M. Nakazato, H. Matsuo, and K. Kangawa. 1999. Ghrelin is a growth-hormonereleasing acylated peptide from stomach. *Nature* **402**.

Kosari, F., A.S. Parker, D.M. Kube, C.M. Lohse, B.C. Leibovich, M.L. Blute, J.C. Cheville, and G. Vasmatzis. 2005. Clear cell renal cell carcinoma: gene expression analyses identify a potential signature for tumor aggressiveness. *Clin Cancer Res* **11:** 5128-5139.

Kosfeld, M., M. Heinrichs, P.J. Zak, E. Fischbacher, and E. Fehr. 2005. Oxytocin increases trust in human. *Nature* **435:** 673-676.

Krogh, A., M. Brown, I.S. Mian, K. Sjolander, and D. Haussler. 1994a. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235:** 1501-1531.

Krogh, A., B. Larsson, G. von Heijne, and E.L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305:** 567-580.

Krogh, A., I.S. Mian, and D. Haussler. 1994b. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res* **22:** 4768-4778.

Kumar, M., R. Verma, and G.P. Raghava. 2006. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *J Biol Chem* **281:** 5357-5363.

Lang, B., B. Song, W. Davidson, A. MacKenzie, N. Smith, C.D. McCaig, A.J. Harmar, and S. Shen. 2006. Expression of the human PAC1 receptor leads to dose-dependent hydrocephalus-related abnormalities in mice. *J Clin Invest* **116:** 1924-1934.

Leder, E.H. and J.T. Silverstein. 2006. The pro-opiomelanocortin genes in rainbow trout (Oncorhynchus mykiss): duplications, splice variants, and differential expression. *J Endocrinol* **188:** 355-363.

Lein, E.S. M.J. Hawrylycz N. Ao M. Ayres A. Bensinger A. Bernard A.F. Boe M.S. Boguski K.S. Brockway E.J. Byrnes L. Chen L. Chen T.M. Chen M.C. Chin J. Chong B.E. Crook A. Czaplinska C.N. Dang S. Datta N.R. Dee A.L. Desaki T. Desta E. Diep T.A. Dolbeare M.J. Donelan H.W. Dong J.G. Dougherty B.J. Duncan A.J. Ebbert G. Eichele L.K. Estin C. Faber B.A. Facer R. Fields S.R. Fischer T.P. Fliss C. Frensley S.N. Gates K.J. Glattfelder K.R. Halverson M.R. Hart J.G. Hohmann M.P. Howell D.P. Jeung R.A. Johnson P.T. Karr R. Kawal J.M. Kidney R.H. Knapik C.L. Kuan J.H. Lake A.R. Laramee K.D. Larsen C. Lau T.A. Lemon A.J. Liang Y. Liu L.T. Luong J. Michaels J.J. Morgan R.J. Morgan M.T. Mortrud N.F. Mosqueda L.L. Ng R. Ng G.J. Orta C.C. Overly T.H. Pak S.E. Parry S.D. Pathak O.C. Pearson R.B. Puchalski Z.L. Riley H.R. Rockett S.A. Rowland J.J. Royall M.J. Ruiz N.R. Sarno K. Schaffnit N.V. Shapovalova T. Sivisay C.R. Slaughterbeck S.C. Smith K.A. Smith B.I. Smith A.J. Sodt N.N. Stewart K.R. Stumpf S.M. Sunkin M. Sutram A. Tam C.D. Teemer C. Thaller C.L. Thompson L.R. Varnam A. Visel R.M. Whitlock P.E. Wohnoutka C.K. Wolkey V.Y. Wong M. Wood M.B. Yaylaoglu R.C. Young B.L. Youngstrom X.F. Yuan B. Zhang T.A. Zwingman and A.R. Jones. 2007. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445:** 168-176.

Li, Y., J. Chen, and M. Chopp. 2002. Cell proliferation and differentiation from ependymal, subependymal and choroid plexus cells in response to stroke in rats. *J Neurol Sci* **193:** 137-146.

Lichtenauer, U., I. Shapiro, K. Geiger, P. AM., J. Drouin, and F. Beuschlein. 2007. Isolation of the side population (SP) from murine adrenal glands renders cells with adrenocortical stem cell properties.

Lim, D.A., A.D. Tramontin, J.M. Trevejo, D.G. Herrera, J.M. Garcia-Verdugo, and A. Alvarez-Buylla. 2000. Noggin antagonizes BMP signaling to create a niche for adult neurogenesis. *Neuron* **28:** 713-726.

Lin, L., J. Faraco, R. Li, H. Kadotani, W. Rogers, X. Lin, X. Qiu, P.J. de Jong, S. Nishino, and E. Mignot. 1999. The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* **98:** 365-376.

Liu, C., J. Chen, S. Sutton, B. Roland, C. Kuei, N. Farmer, R. Sillard, and T.W. Lovenberg. 2003a. Identification of relaxin-3/INSL7 as a ligand for GPCR142. *J Biol Chem* **278:** 50765-50770.

Liu, Q., Y.S. Zhu, B.H. Wang, and Y.X. Li. 2003b. A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Comput Biol Chem* **27:** 69-76.

Lodge, D.J. and A.A. Grace. 2006. The laterodorsal tegmentum is essential for burst firing of ventral tegmental area dopamine neurons. *Proc Natl Acad Sci U S A* **103:** 5167-5172.

Luo, G., P. Ducy, M.D. McKee, G.J. Pinero, E. Loyer, R.R. Behringer, and G. Karsenty. 1997. Spontaneous calcification of arteries and cartilage in mice lacking matrix GLA protein. *Nature* **386:** 78-81.

Maake, C., W. Kloas, M. Szendefi, and M. Reinecke. 1999. Neurohormonal peptides, serotonin, and nitric oxide synthase in the enteric nervous system and endocrine cells of the gastrointestinal tract of neotenic and thyroid hormone-treated axolotls (Ambystoma mexicanum). *Cell Tissue Res* **297:** 91-101.

Macpherson, P. and M.S. Matheson. 1979. Comparison of calcification of pineal, habenular commissure and choroid plexus on plain films and computed tomography. *Neuroradiology* **18:** 67-72.

Mahmood, Z. and D. McNamara. 2003. Gastro-oesophageal reflux disease and ulcer disease. *Aliment Pharmacol Ther* **18 Suppl 3:** 31-37.

Majoros, W.H., M. Pertea, A.L. Delcher, and S.L. Salzberg. 2005. Efficient decoding algorithms for generalized hidden Markov model gene finders. *BMC Bioinformatics* **6:** 16.

Maniatis T., E.F.F., and J. Sambrook. 1982. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Springs Harbor, NY.

Martoglio, B. and B. Dobberstein. 1998. Signal sequences: more than just greasy peptides. *Trends Cell Biol* **8:** 410-415.

Maruyama, M., H. Matsumoto, K. Fujiwara, C. Kitada, S. Hinuma, H. Onda, M. Fujino, and K. Inoue. 1999a. Immunocytochemical localization of prolactin-releasing peptide in the rat brain. *Endocrinology* **140:** 2326-2333.

Maruyama, M., H. Matsumoto, K. Fujiwara, J. Noguchi, C. Kitada, S. Hinuma, H. Onda, O. Nishimura, M. Fujino, T. Higuchi, and K. Inoue. 1999b. Central administration of prolactin-releasing peptide stimulates oxytocin release in rats. *Neurosci Lett* **276:** 193-196.

Matsumoto, M. and O. Hikosaka. 2007. Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* **447:** 1111-1115.

McDonald, T.J., H. Jornvall, G. Nilsson, M. Vagne, M. Ghatei, S.R. Bloom, and V. Mutt. 1979. Characterization of a gastrin releasing peptide from porcine non-antral gastric tissue. *Biochem Biophys Res Commun* **90:** 227-233.

Meiniel, A. 2007. The secretory ependymal cells of the subcommissural organ: which role in hydrocephalus? *Int J Biochem Cell Biol* **39:** 463-468.

Mellquist, J.L., L. Kasturi, S.L. Spitalnik, and S.H. Shakin-Eshleman. 1998. The amino acid following an asn-X-Ser/Thr sequon is an important determinant of N-linked core glycosylation efficiency. *Biochemistry* **37:** 6833-6837.

Mena-Segovia, J., J.P. Bolam, and P.J. Magill. 2004. Pedunculopontine nucleus and basal ganglia: distant relatives or part of the same family? *Trends Neurosci* **27:** 585-588.

Mercier, F., J.T. Kitasako, and G.I. Hatton. 2002. Anatomy of the brain neurogenic zones revisited: fractones and the fibroblast/macrophage network. *J Comp Neurol* **451:** 170-188.

Mercier, F., J.T. Kitasako, and G.I. Hatton. 2003. Fractones and other basal laminae in the hypothalamus. *J Comp Neurol* **455:** 324-340.

Meyers, J.A., D. Sanchez, L.P. Elwell, and S. Falkow. 1976. Simple agarose gel electrophoretic method for the identification and characterization of plasmid deoxyribonucleic acid. *J Bacteriol* **127:** 1529-1537.

Mirabeau, O., E. Perlas, C. Severini, E. Audero, O. Gascuel, R. Possenti, E. Birney, N. Rosenthal, and C. Gross. 2007. Identification of novel peptide hormones in the human proteome by hidden Markov model screening. *Genome Res* **17:** 320-327.

Miselis, R.R., T.M. Hyde, and R.E. Shapiro. 1984. Area postrema and adjacent solitary nucleus in water and energy balance. *Fed Proc* **43:** 2969-2971.

Moffett, M., L. Stanek, J. Harley, G. Rogge, M. Asnicar, H. Hsiung, and M. Kuhar. 2006. Studies of cocaine- and amphetamine-regulated transcript (CART) knockout mice. *Peptides* **27:** 2037-2045.

Mohamed, M. and P. Gader. 1996. Handwritten Word Recognition Using Segmentation-Free Hidden Markov Modeling and Segmentation-Based Dynamic Programming Techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18:** 548-554.

Mohler, E.R., 3rd. 2004. Mechanisms of aortic valve calcification. *Am J Cardiol* **94:** 1396-1402, A1396.

Mullis, K.B. and F.A. Faloona. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* **155:** 335-350.

Nakayama, K. 1997. Furin: a mammalian subtilisin/Kex2p-like endoprotease involved in processing of a wide variety of precursor proteins. *Biochem J* **327 (Pt 3):** 625-635.

Nathoo, A.N., R.A. Moeller, B.A. Westlund, and A.C. Hart. 2001. Identification of neuropeptide-like protein gene families in Caenorhabditiselegans and other species. *Proc Natl Acad Sci U S A* **98:** 14000-14005.

Nemeth, E., E.V. Valore, M. Territo, G. Schiller, A. Lichtenstein, and T. Ganz. 2003. Hepcidin, a putative mediator of anemia of inflammation, is a type II acute-phase protein. *Blood* **101:** 2461-2463.

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

Nielsen, H. and A. Krogh. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* **6:** 122-130.

Nikolova, G., B. Strilic, and E. Lammert. 2007. The vascular niche and its basement membrane. *Trends Cell Biol* **17:** 19-25.

Nilsson, C., M. Lindvall-Axelsson, and C. Owman. 1992. Neuroendocrine regulatory mechanisms in the choroid plexus-cerebrospinal fluid system. *Brain Res Brain Res Rev* **17:** 109-138.

Nishikawa, A. and S. Mizuno. 2001. The efficiency of N-linked glycosylation of bovine DNase I depends on the Asn-Xaa-Ser/Thr sequence and the tissue of origin. *Biochem J* **355:** 245-248.

Noda, M., Y. Teranishi, H. Takahashi, M. Toyosato, M. Notake, S. Nakanishi, and S. Numa. 1982. Isolation and structural organization of the human preproenkephalin gene. *Nature* **297:** 431-434.

Ogorek, C.P. and S. Cohen. 1989. Gastroesophageal reflux disease: new concepts in pathophysiology. *Gastroenterol Clin North Am* **18:** 275-292.

Oliver, J.A., O. Maarouf, F.H. Cheema, T.P. Martens, and Q. Al-Awqati. 2004. The renal papilla is a niche for adult kidney stem cells. *J Clin Invest* **114:** 795-804.

Olsson, G.F.a.R. 1959. The praeoptico-hypophysial system, Nucleus tuberis lateralis and the subcommissural organ of Gasterosteus aculeatus after changes in osmotic stimuli. *Cell and Tissue Research*.

Ozaki, O., K. Ito, K. Kobayashi, K. Toshima, H. Iwasaki, and T. Yashiro. 1990. Thyroid carcinoma in Graves' disease. *World J Surg* **14:** 437-440; discussion 440-431.

Ozmen, M.N., N. Aygun, I. Kilic, L. Kuran, B. Yalcin, and A. Besim. 1992. Wolman's disease: ultrasonographic and computed tomographic findings. *Pediatr Radiol* **22:** 541-542.

Pahapill, P.A. and A.M. Lozano. 2000. The pedunculopontine nucleus and Parkinson's disease. *Brain* **123 (Pt 9):** 1767-1783.

Palkovits, M., E. Monos, and J. Fachet. 1965. The Effect Of Subcommissural-Organ Lesions On Aldosterone Production In The Rat. *Acta Endocrinol (Copenh)* **48:** 169-176.

Parhami, F., Y. Tintut, A. Ballard, A.M. Fogelman, and L.L. Demer. 2001. Leptin enhances the calcification of vascular cells: artery wall as a target of leptin. *Circ Res* **88:** 954-960.

Park, K.J. and M. Kanehisa. 2003. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* **19:** 1656-1663.

Paul, M., A. Poyan Mehr, and R. Kreutz. 2006. Physiology of local renin-angiotensin systems. *Physiol Rev* **86:** 747-803.

Pedersen, A.G., P. Baldi, S. Brunak, and Y. Chauvin. 1996. Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol* **4:** 182-191.

Pedersen, C.A., J.A. Ascher, Y.L. Monroe, and A.J. Prange Jr. 1982. Oxytocin induces maternal behavior in virgin female rats. *Science* **216:** 648-650.

Pedersen, J.S. and J. Hein. 2003. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* **19:** 219-227.

Pencea, V., K.D. Bingaman, S.J. Wiegand, and M.B. Luskin. 2001. Infusion of brain-derived neurotrophic factor into the lateral ventricle of the adult rat leads to new neurons in the parenchyma of the striatum, septum, thalamus, and hypothalamus. *J Neurosci* **21:** 6706-6717.

Pepper, M.S. and M. Skobe. 2003. Lymphatic endothelium: morphological, molecular and functional properties. *J Cell Biol* **163:** 209-213.

Persson, E.K., G. Hagberg, and P. Uvebrant. 2005. Hydrocephalus prevalence and outcome in a population-based cohort of children born in 1989-1998. *Acta Paediatr* **94:** 726-732.

Peyron, C., D.K. Tighe, A.N. van den Pol, L. de Lecea, H.C. Heller, J.G. Sutcliffe, and T.S. Kilduff. 1998. Neurons containing hypocretin (orexin) project to multiple neuronal systems. *J Neurosci* **18:** 9996-10015.

Pfeiffer, R.F. 2003. Gastrointestinal dysfunction in Parkinson's disease. *Lancet Neurol* **2:** 107-116.

Pohle, K., R. Maffert, D. Ropers, W. Moshage, N. Stilianakis, W.G. Daniel, and S. Achenbach. 2001. Progression of aortic valve calcification: association with coronary atherosclerosis and cardiovascular risk factors. *Circulation* **104:** 1927-1932.

Puntervoll, P., R. Linding, C. Gemünd, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D.M.A. Martin, W.M. Hunter, G. Ausiello, B. Brannetti, A. Costantini, F. Ferree, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Küster, M. Helmer-Citterich, W.M. Hunte, R. Aasland, and T.J. Gibson. 2003. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research* **31:** 3625-3630.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77:** 257-286.

Rajamannan, N.M., M. Subramaniam, F. Caira, S.R. Stock, and T.C. Spelsberg. 2005. Atorvastatin inhibits hypercholesterolemia-induced calcification in the aortic valves via the Lrp5 receptor pathway. *Circulation* **112:** I229-234.

Rajamannan, N.M., M. Subramaniam, D. Rickard, S.R. Stock, J. Donovan, M. Springett, T. Orszulak, D.A. Fullerton, A.J. Tajik, R.O. Bonow, and T. Spelsberg. 2003. Human aortic valve calcification is associated with an osteoblast phenotype. *Circulation* **107:** 2181-2184.

Rajamannan, N.M., M. Subramaniam, M. Springett, T.C. Sebo, M. Niekrasz, J.P. McConnell, R.J. Singh, N.J. Stone, R.O. Bonow, and T.C. Spelsberg. 2002. Atorvastatin inhibits hypercholesterolemia-induced cellular proliferation and bone matrix production in the rabbit aortic valve. *Circulation* **105:** 2660-2665.

Reinscheid, R.K., H.P. Nothacker, A. Bourson, A. Ardati, R.A. Henningsen, J.R. Bunzow, D.K. Grandy, H. Langen, F.J. Monsma, Jr., and O. Civelli. 1995. Orphanin FQ: a neuropeptide that activates an opioidlike G protein-coupled receptor. *Science* **270:** 792-794.

Resin, H., D.H. Stern, R.A. Sturdevant, and J.I. Isenberg. 1973. Effect of the C-terminal octapeptide of cholecystokinin on lower esophageal sphincter pressure in man. *Gastroenterology* **64:** 946-949.

Reynolds, B.A. and S. Weiss. 1992. Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science* **255:** 1707-1710.

Riehle, M.A., S.F. Garczynski, J.W. Crim, C.A. Hill, and M.R. Brown. 2002. Neuropeptides and peptide hormones in Anopheles gambiae. *Science* **298:** 172-175.

Rodriguez, E.M., A. Oksche, S. Hein, S. Rodriguez, and R. Yulis. 1984a. Comparative immunocytochemical study of the subcommissural organ. *Cell Tissue Res* **237:** 427-441.

Rodriguez, E.M., A. Oksche, S. Hein, S. Rodriguez, and R. Yulis. 1984b. Spatial and structural interrelationships between secretory cells of the subcommissural organ and blood vessels. An immunocytochemical study. *Cell Tissue Res* **237:** 443-449.

Rodriguez, E.M., S. Rodriguez, and S. Hein. 1998. The subcommissural organ. *Microsc Res Tech* **41:** 98-123.

Rostene, W., P. Kitabgi, and S.M. Parsadaniantz. 2007. Chemokines: a new class of neuromodulator? *Nat Rev Neurosci* **8:** 895-903.

Saiki, R.K., D.H. Gelfand, S. Stoffel, S.J. Scharf, R. Higuchi, G.T. Horn, K.B. Mullis, and H.A. Erlich. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239:** 487-491.

Sakurai, T., A. Amemiya, M. Ishii, I. Matsuzaki, R.M. Chemelli, H. Tanaka, S.C. Williams, J.A. Richardson, G.P. Kozlowski, S. Wilson, J.R. Arch, R.E. Buckingham, A.C. Haynes, S.A. Carr, R.S. Annan, D.E. McNulty, W.S. Liu, J.A. Terrett, N.A. Elshourbagy, D.J. Bergsma, and M. Yanagisawa. 1998a. Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92:** 573-585.

Sakurai, T., A. Amemiya, M. Ishii, I. Matsuzaki, R.M. Chemelli, H. Tanaka, S.C. Williams, J.A. Richarson, G.P. Kozlowski, S. Wilson, J.R. Arch, R.E. Buckingham, A.C. Haynes, S.A. Carr, R.S. Annan, D.E. McNulty, W.S. Liu, J.A. Terrett, N.A. Elshourbagy, D.J. Bergsma, and M. Yanagisawa. 1998b. Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92:** 1 page following 696.

Sambrook, J., T. Maniatis, and E.F. Fritsch. 1989. *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-5467.

Sawamoto, K., H. Wichterle, O. Gonzalez-Perez, J.A. Cholfin, M. Yamada, N. Spassky, N.S. Murcia, J.M. Garcia-Verdugo, O. Marin, J.L. Rubenstein, M. Tessier-Lavigne, H. Okano, and A. Alvarez-Buylla. 2006. New neurons follow the flow of cerebrospinal fluid in the adult brain. *Science* **311:** 629-632.

Scarnati, E., E. Campana, and C. Pacitti. 1984. Pedunculopontine-evoked excitation of substantia nigra neurons in the rat. *Brain Res* **304:** 351-361.

Schacht, V., S.S. Dadras, L.A. Johnson, D.G. Jackson, Y.K. Hong, and M. Detmar. 2005. Up-regulation of the lymphatic marker podoplanin, a mucin-type transmembrane glycoprotein, in human squamous cell carcinomas and germ cell tumors. *Am J Pathol* **166:** 913-921.

Schuster-Bockler, B., J. Schultz, and S. Rahmann. 2004. HMM Logos for visualization of protein families. *BMC Bioinformatics* **5:** 7.

Seidah, N. and M. Chretien. 1999. Proprotein and prohormone convertases: a family of subtilases generating diverse bioactive polypeptides. *Brain Research* **848:** 45-62.

Seidah, N.G., S. Benjannet, S. Pareek, D. Savaria, J. Hamelin, B. Goulet, J. Laliberte, C. Lazure, M. Chretien, and R.A. Murphy. 1996. Cellular processing of the nerve growth factor precursor by the mammalian pro-protein convertases. *Biochem J* **314 (Pt 3):** 951-960.

Seoane, L.M., E. Carro, S. Tovar, F.F. Casanueva, and C. Dieguez. 2000. Regulation of in vivo TSH secretion by leptin. *Regul Pept* **92:** 25-29.

Shichiri, M., S. Ishimaru, T. Ota, T. Nishikawa, T. Isogai, and Y. Hirata. 2003. Salusins: newly identified bioactive peptides with hemodynamic and mitogenic activities. *Nat Med* **9:** 1166-1172.

Siegel, S.R., F.C. Brown, D.O. Castell, L.F. Johnson, and S.I. Said. 1979. Effects of vasoactive intestinal polypeptide (VIP) on lower esophageal sphincter in awake baboons: comparison with glucagon and secretin. *Dig Dis Sci* **24:** 345-349.

Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, and D. Haussler. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034-1050.

Singh, G. and A.P. Davenport. 2006. Neuropeptide B and W: neurotransmitters in an emerging G-protein-coupled receptor system. *Br J Pharmacol* **148:** 1033-1041.

Smith, P.M., W.K. Samson, and A.V. Ferguson. 2007. Cardiovascular actions of orexin-A in the rat subfornical organ. *J Neuroendocrinol* **19:** 7-13.

Sonnhammer, E.L., S.R. Eddy, E. Birney, A. Bateman, and R. Durbin. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26:** 320-322.

Southey, B.R., A. Amare, T.A. Zimmerman, S.L. Rodriguez-Zas, and J.V. Sweedler. 2006. NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. *Nucleic Acids Res* **34:** W267-272.

Spassky, N., F.T. Merkle, N. Flames, A.D. Tramontin, J.M. Garcia-Verdugo, and A. Alvarez-Buylla. 2005. Adult ependymal cells are postmitotic and are derived from radial glial cells during embryogenesis. *J Neurosci* **25:** 10-18.

Spechler, S.J. and D.O. Castell. 2001. Classification of oesophageal motility abnormalities. *Gut* **49:** 145-151.

Stache-Crain, B., J. Lee, Y. Tang, I. Lobal, R. Drmanac, and M. Nishikawa. 2008. Polypeptide having an activity to support proliferation or survival of hematopoietic stem cell and hematopoietic progenitor cell, and DNA coding for the same, United States.

Stark, M., O. Danielsson, W.J. Griffiths, H. Jornvall, and J. Johansson. 2001. Peptide repertoire of human cerebrospinal fluid: novel proteolytic fragments of neuroendocrine proteins. *J Chromatogr B Biomed Sci Appl* **754:** 357-367.

Stein, J., D.F. Steiner, and A. Dey. 2006. Processing of cocaine- and amphetamine-regulated transcript (CART) precursor proteins by prohormone convertases (PCs) and its implications. *Peptides* **27:** 1919-1925.

Steiner, D.F. 1998. The proprotein convertases. *Curr Opin Chem Biol* **2:** 31-39.

Steiner, D.F., D. Cunningham, L. Spigelman, and B. Aten. 1967. Insulin biosynthesis: evidence for a precursor. *Science* **157:** 697-700.

Sterba, G., A. Ermisch, K. Freyer, and G. Hartmann. 1967. Incorporation of sulphur-35 into the subcommissural organ and Reissner's fibre. *Nature* **216:** 504.

Sterba, G., C. Kiessig, W. Naumann, H. Petter, and I. Kleim. 1982. The secretion of the subcommissural organ. A comparative immunocytochemical investigation. *Cell Tissue Res* **226:** 427-439.

Subburaju, S. and G. Aguilera. 2007. Vasopressin mediates mitogenic responses to adrenalectomy in the rat anterior pituitary. *Endocrinology* **148:** 3102-3110.

Sun, Y., G. Hegamyer, and N.H. Colburn. 1993. PCR-direct sequencing of a GC-rich region by inclusion of 10% DMSO: application to mouse c-jun. *Biotechniques* **15:** 372-374.

Sutherland, R.J. 1982. The dorsal diencephalic conduction system: a review of the anatomy and functions of the habenular complex. *Neurosci Biobehav Rev* **6:** 1-13.

Szekely, M., E. Petervari, M. Balasko, I. Hernadi, and B. Uzsoki. 2002. Effects of orexins on energy balance and thermoregulation. *Regul Pept* **104:** 47-53.

Takano, T. 2007. Fetal cell carcinogenesis of the thyroid: theory and practice. *Semin Cancer Biol* **17:** 233-240.

Takeuchi, I.K., R. Kimura, M. Matsuda, and R. Shoji. 1987. Absence of subcommissural organ in the cerebral aqueduct of congenital hydrocephalus spontaneously occurring in MT/HokIdr mice. *Acta Neuropathol (Berl)* **73:** 320-322.

Tartaglia, L.A., M. Dembski, X. Weng, N. Deng, J. Culpepper, R. Devos, G.J. Richards, L.A. Campfield, F.T. Clark, J. Deeds, C. Muir, S. Sanker, A. Moriarty, K.J. Moore, J.S. Smutko,

G.G. Mays, E.A. Wool, C.A. Monroe, and R.I. Tepper. 1995. Identification and expression cloning of a leptin receptor, OB-R. *Cell* **83:** 1263-1271.

Tatemoto, K., S. Efendic, V. Mutt, G. Makk, G.J. Feistner, and J.D. Barchas. 1986. Pancreastatin, a novel pancreatic peptide that inhibits insulin secretion. *Nature* **324:** 476-478.

Temple, S. 2001. The development of neural stem cells. *Nature* **414:** 112-117.

Thomas, G. 2002. Furin at the cutting edge: from protein traffic to embryogenesis and disease. *Nat Rev Mol Cell Biol* **3:** 753-766.

Thomas, G., B.A. Thorne, L. Thomas, R.G. Allen, D.E. Hruby, R. Fuller, and J. Thorner. 1988. Yeast KEX2 endopeptidase correctly cleaves a neuroendocrine prohormone in mammalian cells. *Science* **241:** 226-230.

Tropepe, V., B.L. Coles, B.J. Chiasson, D.J. Horsford, A.J. Elia, R.R. McInnes, and D. van der Kooy. 2000. Retinal stem cells in the adult mammalian eye. *Science* **287:** 2032-2036.

Upton, P.D., F.W. Dunihue, and W.F. Chambers. 1961. Subcommissural organ and water metabolism. *Am J Physiol* **201:** 711-713.

Urabe, N., I. Naito, K. Saito, T. Yonezawa, Y. Sado, H. Yoshioka, S. Kusachi, T. Tsuji, A. Ohtsuka, T. Taguchi, T. Murakami, and Y. Ninomiya. 2002. Basement membrane type IV collagen molecules in the choroid plexus, pia mater and capillaries in the mouse brain. *Arch Histol Cytol* **65:** 133-143.

Valdar, W.S. 2002a. Scoring residue conservation. *Proteins* **48:** 227-241.

Valdar, W.S.J. 2002b. Scoring Residue Conservation. *PROTEINS: Structure, Function, and Genetics* **48:** 227–241.

Vale, W., J. Spiess, C. Rivier, and J. Rivier. 1981. Characterization of a 41-residue ovine hypothalamic peptide that stimulates secretion of corticotropin and beta-endorphin. *Science* **213:** 1394-1397.

van den Pol, A.N., X.B. Gao, K. Obrietan, T.S. Kilduff, and A.B. Belousov. 1998. Presynaptic and postsynaptic actions and modulation of neuroendocrine neurons by a new hypothalamic peptide, hypocretin/orexin. *J Neurosci* **18:** 7962-7971.

Vankelecom, H. 2007. Stem cells in the postnatal pituitary? *Neuroendocrinology* **85:** 110-130.

Vaudry, D., B.J. Gonzalez, M. Basille, L. Yon, A. Fournier, and H. Vaudry. 2000. Pituitary adenylate cyclase-activating polypeptide and its receptors: from structure to functions. *Pharmacol Rev* **52:** 269-324.

Vert, J.P. 2002. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. *Pac Symp Biocomput***:** 649-660.

Vesely, I. 1998. The role of elastin in aortic valve mechanics. *J Biomech* **31:** 115-123.

Vinson, G.P. 2003. Adrenocortical zonation and ACTH. *Microsc Res Tech* **61:** 227-239.

Vio, K., S. Rodriguez, E.H. Navarrete, J.M. Perez-Figares, A.J. Jimenez, and E.M. Rodriguez. 2000. Hydrocephalus induced by immunological blockage of the subcommissural organ-Reissner's fiber (RF) complex by maternal transfer of anti-RF antibodies. *Exp Brain Res* **135:** 41-52.

von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res* **14:** 4683-4690.

Vrang, N. 2006. Anatomy of hypothalamic CART neurons. *Peptides* **27:** 1970-1980.

Vrang, N., P.J. Larsen, J.T. Clausen, and P. Kristensen. 1999. Neurochemical characterization of hypothalamic cocaine- amphetamine-regulated transcript neurons. *J Neurosci* **19:** RC5.

Walker, J.R. and T. Wiltshire. 2006. Databases of free expression. *Mamm Genome* **17:** 1141-1146.

Warmuth-Metz, M., R. Klein, N. Sorensen, and L. Solymosi. 1999. Central neurocytoma of the fourth ventricle. Case report. *J Neurosurg* **91:** 506-509.

Waterston, R.H., K. Lindblad-Toh, E. Birney, J. Rogers, J. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M.C. Zody, and E.S. Lander. 2002. Initial sequencing and comparative analysis of the mouse genome. *nature* **420:** 520-562.

Wetterwald, A., W. Hoffstetter, M.G. Cecchini, B. Lanske, C. Wagner, H. Fleisch, and M. Atkinson. 1996. Characterization and cloning of the E11 antigen, a marker expressed by rat osteoblasts and osteocytes. *Bone* **18:** 125-132.

Wienke, J.R., W.K. Chong, J.R. Fielding, K.H. Zou, and C.A. Mittelstaedt. 2003. Sonographic features of benign thyroid nodules: interobserver reliability and overlap with malignancy. *J Ultrasound Med* **22:** 1027-1031.

Wise, R.A. 1996. Neurobiology of addiction. *Curr Opin Neurobiol* **6:** 243-251.

Wu, J. and D. Haussler. 2006. Coding exon detection using comparative sequences. *J Comput Biol* **13:** 1148-1164.

Wu, X.B., Y. Li, A. Schneider, W. Yu, G. Rajendren, J. Iqbal, M. Yamamoto, M. Alam, L.J. Brunet, H.C. Blair, M. Zaidi, and E. Abe. 2003. Impaired osteoblastic differentiation, reduced bone formation, and severe osteoporosis in noggin-overexpressing mice. *J Clin Invest* **112:** 924-934.

Xu, Y.L., R.K. Reinscheid, S. Huitron-Resendiz, S.D. Clark, Z. Wang, S.H. Lin, F.A. Brucher, J. Zeng, N.K. Ly, S.J. Henriksen, L. de Lecea, and O. Civelli. 2004. Neuropeptide S: a neuropeptide promoting arousal and anxiolytic-like effects. *Neuron* **43:** 487-497.

Xue, Y., H. Chen, C. Jin, Z. Sun, and X. Yao. 2006. NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm. *BMC Bioinformatics* **7**.

Yamashita, T., M. Ninomiya, P. Hernandez Acosta, J.M. Garcia-Verdugo, T. Sunabori, M. Sakaguchi, K. Adachi, T. Kojima, Y. Hirota, T. Kawase, N. Araki, K. Abe, H. Okano, and K. Sawamoto. 2006. Subventricular zone-derived neuroblasts migrate and differentiate into mature neurons in the post-stroke adult striatum. *J Neurosci* **26:** 6627-6636.

Yamron, J., I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. 1998. A hidden Markov model approach to text segmentation and eventtracking. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998.*

Yoshimichi, G., H. Yoshimatsu, T. Masaki, and T. Sakata. 2001. Orexin-A regulates body temperature in coordination with arousal status. *Exp Biol Med (Maywood)* **226:** 468-476.

Yue, C.M., D.J. Deng, M.X. Bi, L.P. Guo, and S.H. Lu. 2003. Expression of ECRG4, a novel esophageal cancer-related gene, downregulated by CpG island hypermethylation in human esophageal squamous cell carcinoma. *World J Gastroenterol* **9:** 1174-1178.

Zerbib, F., S. Bruley Des Varannes, C. Scarpignato, V. Leray, M. D'Amato, C. Roze, and J.P. Galmiche. 1998. Endogenous cholecystokinin in postprandial lower esophageal sphincter function and fundic tone in humans. *Am J Physiol* **275:** G1266-1273.

Zhang, J.V., P.G. Ren, O. Avsian-Kretchmer, C.W. Luo, R. Rauch, C. Klein, and A.J.W. Hsueh. 2005. Obestatin, a Peptide Encoded by the Ghrelin Gene, Opposes Ghrelin's Effects on Food Intake. *Science* **310:** 996-999.

10 Résumé détaillé de la thèse

10.1 Introduction

10.1.1 Contexte

Cette thèse a été effectuée à l'EMBL de Monterotondo dans les groupes de Nadia Rosenthal et Cornelius Gross. L'idée du projet a germé au cours d'une discussion entre Cornelius Gross et Ewan Birney au début de l'année 2002. L'idée de départ était qu'il serait possible de se servir des séquences génomiques pour tenter de trouver de nouvelles hormones peptidiques (PH), car de nouvelles hormones peptidiques venaient récemment d'être découvertes (ghréline et orexine) (Kojima et al. 1999; Sakurai et al. 1998a). Le fait que plusieurs nouvelles hormones aient été découvertes ces dix dernières années, qu'il soit difficile de les identifier en utilisant les techniques classiques de biochimie à cause de leur petite taille et de leur instabilité, et que de nombreux GPCR soient orphelins de leur ligand constituaient autant de bonnes raisons pour conduire une analyse exploratoire du génome humain pour tenter de découvrir de nouvelles PH. Enfin, l'explosion de la quantité de génomes de vertébrés disponibles et l'identification de leurs gènes, en particulier de la souris (Waterston et al. 2002), ont permis à ce projet de voir le jour.

10.1.2 Intérêt des hormones peptidiques

Les hormones peptidiques ou neuropeptides sont une classe importante de molécules puisqu'elles participent à un grand nombre de processus physiologiques chez les organismes complexes. On peut citer la régulation de la pression sanguine, du sommeil, de la faim, de la soif, de la douleur, du stress, du comportement sexuel, de la digestion, et des fonctions immunitaires. Nombre d'entre elles sont les messagers entre le cerveau et les organes périphériques, permettant aux organismes de produire des comportements adaptés aux situations et à l'environnement en informant le cerveau sur, par exemple, les réserves énergétiques du corps, les dégâts corporels et l'état général du métabolisme.

Parmi les systèmes peptidiques connus, l'on peut citer les hormones hypothalamiques qui agissent sur l'adénohypophyse (hormones hypophysiotropes telles que l'hormone de croissance, la thyrolibérine, la corticolibérine et la somatolibérine) pour réguler l'activité des principales glandes endocrines du corps (thyroïde, glande surrénale, gonades), le système angiotensine-rénine-aldosterone (RAAS) qui contribue au contrôle de la pression artérielle et de la soif, les hormones gastro-intestinales qui contrôlent la digestion (cholécystokinine,

gastrine, etc.), et les hormones sexuelles (gonadolibérine) qui contrôlent les comportements liés a la reproduction de l'espèce.

10.1.3 Définition d'une hormone peptidique

Une définition stricte pour désigner l'ensemble des hormones peptidiques n'est pas simple à établir. Un peptide est une courte séquence d'acides aminés (moins de 40 acides amines), et une hormone est une molécule qui est sécrétée par des cellules dites endocrines, et qui se lie a un récepteur afin d'exercer une action sur des cellules loin des cellules qui l'ont sécrétées. La distinction entre hormone peptidique et neuropeptide vient essentiellement du fait qu'un neuropeptide a une fonction dans le cerveau et active une voie cellulaire au travers de sa liaison avec un GPCR. Pour ce qui nous concerne, nous avons considéré que ces deux types de molécules avaient essentiellement les mêmes caractéristiques biochimiques et nous parlerons d'hormone peptidique lorsqu'il s'agit de neuropeptide. En effet bien des neuropeptides ont aussi des fonctions dans des organes périphériques, comme par exemple la pro-opiomélanocortine et la cholécystokinine. Les hormones peptidiques sont co-traductionnellement transloquées dans le réticulum endoplasmique, pour être ensuite modifiées dans l'appareil de Golgi. Enfin, elles seront emmagasinées dans des vésicules a coeur dense (DCG) qui, après fusion avec la membrane plasmique (exocytose) vont libérer ces molécules dans le milieu extracellulaire.

Avant de penser pouvoir en prédire de nouvelles il faut pouvoir donner une définition assez précise de ce qu'est une hormone peptidique. Nous avons considéré qu'une hormone peptidique était un peptide issu du clivage d'une protéine codée par le génome, appelée protéine précurseur, et possédant les caractéristiques fonctionnelles des hormones, c-a-d que ce sont de petites protéines secrétées exerçant leur action à distance sur d'autres cellules, après liaison avec un récepteur cellulaire. Nous avons pu dresser une liste d'environ 80 hormones peptidiques grâce aux mots clés de SwissProt suivant : « peptide », « hormone », « neuropeptide » et « clivage au niveau de paires de résidus basiques » (cleavage on pairs of basic residues). Il n'est pas aisé de définir une frontière entre les hormones peptidiques et les facteurs de. En effet, comme les précurseurs a hormones peptidiques, les précurseurs des facteurs de croissances possèdent bien souvent des sites de clivage dibasiques, qui, après découpage par des prohormones convertases, libèrent des protéines de plus petite taille. Ces facteurs de croissance (GF) agissent aussi à distance, de manière endocrine ou paracrine, tout comme les hormones peptidiques. Cependant, les GF ne se lient généralement pas à des

GPCR, contrairement au PH, et les GF produites après processing sont de taille plus grande que les PH (typiquement plus de 100 acides aminés). De plus, comme les cytokines, les GF agissent souvent sur le destin des cellules cibles (prolifération, différenciation) contrairement aux PH, qui peuvent jouer un rôle de neuromodulateur à la synapse. Pour certaines molécules, la catégorisation en PH ou GF n'est pas simple à établir, comme dans le cas de l'Insulin-Like Growth Factor-1 et 2 (IGF1-2) qui sont des peptides de 70 acides aminés environ et de structure proche de l'insuline et ayant des propriétés hormonales, mais qui sont généralement considérés comme des facteurs de croissance (comme leur nom l'indique).

10.1.4 Repères historiques

Le terme hormone fut utilisé pour la première fois en 1905 lors de la découverte par Ernest Starling de la sécrétine, une hormone gastro-intestinale. Starling fut le premier chercheur à utiliser le terme hormone pour désigner une substance produite par des organes, et agissant à distance sur d'autres organes. Parmi les hormones peptidiques les plus célèbres qui furent découvertes au siècle dernier, l'on doit citer la gastrine, qui ordonne la libération d'acide gastrique HCl dans l'estomac, l'insuline qui fut caractérisée par Banting et Macleod en 1923, l'oxytocine qui fut découverte par Du Vigneaud en 1955, et les hormones hypothalamiques libératrices (hypothalamic releasing hormones), la thyrolibérine, la gonadolibérine et la somatostatine, par Schally et Guillemin en 1977. Par la suite, d'autres stratégies furent utilisées pour isoler de nouveaux peptides, comme celle dite de pharmacologie inverse (« reverse pharmacology ») (Civelli 1998) qui vise à utiliser des récepteurs couplés aux protéines G (GPCR) dont on ne connaît pas de ligand (ils sont qualifiés de récepteurs « orphelins ») comme appât pour isoler et purifier le ligand endogène qui leur correspond. La nociceptine (Reinscheid et al. 1995), le neuropeptide S (Xu et al. 2004) et les neuropeptides B et W (Singh and Davenport 2006) ont ainsi été caractérisés.

Enfin, plus récemment, des chercheurs ont pu dresser des répertoires assez complets d'hormones peptidiques chez divers organismes, en s'appuyant sur les données génomiques et des outils bioinformatiques: on connaît ainsi la listes des hormones peptidiques chez la drosophile *D. melanogaster* (Hewes and Taghert 2001), l'anophèle (Riehle et al. 2002), le ver *C. elegans* (Nathoo et al. 2001), l'abeille *A. mellifera* (Hummon et al. 2006), et le coléoptère T. castaneum (Hauser et al. 2008). A notre connaissance, aucune étude poussée visant à identifier l'ensemble des hormones peptidiques chez les mammifères ou vertébrés n'a encore été publiée, en dehors des résultats de cette thèse.

10.1.5 Objectif

Les caractéristiques des hormones peptidiques que nous avons cherché à modéliser sont les suivantes : la présence d'un peptide signal (SP) dans la partie N-terminale de la protéine et la présence de sites de clivage reconnus par les convertases de prohormone.

10.2 Résultats bioinformatiques

Pour cela nous avons modélisé les séquences à l'aide de chaînes de Markov cachées (HMM). Un HMM est défini par son architecture, et ses matrices de transition et d'observation. L'architecture que nous avons implantée rend compte de la structure attendue chez les précurseurs à PH, à savoir l'existence d'un peptide signal au N-terminus, à la présence d'au moins un site de clivage.

10.2.1 Principales applications des HMM

Les fondations théoriques de HMM ont été posées par Baum et ses collaborateurs à la fin des années 1960 dans une série d'articles fondateurs (Baum and Eagon 1967; Baum and Petrie 1966; Baum et al. 1970). Si la théorie des HMM repose sur des concepts qui appartiennent à la théorie des probabilités, pratiquement, elle appartient au domaine de l'intelligence artificielle et de l'apprentissage automatique. Le succès des HMM doit beaucoup à l'efficacité informatique de ses principaux algorithmes.

Les articles fondateurs sont très techniques et difficilement accessibles au chercheur non-initié à la théorie des probabilités. Rabiner, dans son fameux rapport (Rabiner 1989) présente les HMM de manière didactique, permettant ainsi à l'ingénieur de comprendre et d'utiliser les algorithmes à bon escient, sans se préoccuper des aspects plus théoriques.

Traditionnellement, les HMM ont été utilisés pour résoudre les questions liées à la reconnaissance automatique de la parole (Kenny et al. 1990; Rabiner 1989) et de l'écriture manuscrite (Mohamed and Gader 1996), et de l'extraction automatique de l'information (Freitag and McCallum. 1999). Ils furent abondamment mis à profit dans les années 1990 pour résoudre les questions posées par la biologie moléculaire et l'explosion des données de séquences biologiques. Ses premières applications ont permis de s'attaquer aux problèmes de la prédiction de la structure secondaire des protéines (Asai et al. 1993), de la recherche automatique de gènes (Krogh et al. 1994b), et de la modélisation de familles de protéines

telles que les GPCR (Baldi and Chauvin 1994). Des HMM ont récemment été développés pour prédire la structure tridimensionnelle des protéines (Camproux et al. 2004), pour classer les promoteurs d'eucaryotes et de procaryotes (Pedersen et al. 1996), et distinguer peptides signaux et peptides ancres (Nielsen and Krogh 1998). Ils sont de plus utilisés dans des outils d'annotation automatique de génomes (Birney et al. 2001; Birney and Durbin 2000). Les HMM constituent maintenant une technique incontournable en analyse de séquences biologiques.

10.2.2 Présentation des HMM

Un modèle de Markov caché (HMM) d'ordre 1 est défini par :

Un ensemble de symboles que l'on observe. Dans notre cas, il est défini par l'ensemble des acides aminés usuels (alanine, arginine, cytosine, etc.).

Des états définissant des régions homogènes que l'on cherche à modéliser, ainsi que deux états START et END qui définissent l'origine et la fin d'une séquence.

Une matrice de transition définissant les probabilités de passage d'un état a un autre.

Une matrice d'observation qui définit, pour chaque état, les probabilités d'occurrence de chacun des symboles.

L'architecture du HMM, qui définit la connectivité des états entre eux. Choisir une architecture revient à spécifier lesquelles des transitions entre états sont possibles.

Un HMM peut être vu comme un processus qui génère des séquences. A chaque instant *t* on génère un état grâce à la connaissance de l'état à l'instant antérieur et la matrice de transition. A cet état on associe un symbole observé, avec une probabilité spécifiée par la matrice d'observation. Cependant, en pratique, on utilise les HMM pour résoudre les trois problèmes suivants :

Etant donnés un modèle (architecture, matrices d'observation et d'émission) et une séquence observée (suite de symboles) quelle est la séquence d'états (aussi nommée chemin) la plus probable ? Ce problème est typiquement résolu grâce à l'algorithme de Viterbi, qui utilise le principe de programmation dynamique.

Etant donnés un modèle et une séquence, quelle est la probabilité d'observer cette séquence étant donné le modèle ? Cette question est résolue grâce à l'algorithme avant-arrière (« forward-backward »), qui n'est autre qu'une application du principe de programmation dynamique.

Etant donnés une architecture et une séquence, comment estime-t-on les paramètres du modèle (matrices de transition et d'observation) de telle sorte que l'on maximise la probabilité d'observer la séquence ? cette question peut être résolue par l'algorithme Baum-Welch, qui utilise le principe EM (Expectation-Maximisation).

10.2.3 Application des HMM à la découverte de nouveaux précurseurs à PH

Pour l'apprentissage de notre HMM d'hormones peptidiques, les matrices de transition et d'observation ont été apprises à l'aide des annotations de SwissProt.

10.2.3.1 Différents types de modèles

Par ailleurs, deux types d'améliorations par rapport à un schéma classique de HMM ont été développés. D'une part, l'état cyclique modélisant la partie hydrophobe des peptides signaux a été « éclaté » en plusieurs états liés par une topologie fixe. Les paramètres (transitions entre ces états liés) ont été appris en utilisant le principe du maximum de vraisemblance, et en utilisant un ensemble connu de régions hydrophobes de peptides signaux. Une topologie optimale pour le sous-HMM modélisant les sites de clivage a également été développée, en utilisant le principe de maximum de vraisemblance, et le BIC (Bayesian Information Criterion) comme critère pour choisir le modèle.

Un deuxième type d'amélioration a consisté en la prise en compte de l'information contenue dans des alignements de protéines. L'HMM ne s'applique alors plus à des séquences simples, il s'applique désormais à des alignements de protéines. En particulier, dans les algorithmes avant-arrière et Viterbi, la probabilité d'observer le symbole à l'instant t a été remplacée par un terme qui incorpore un score de conservation de la colonne a l'instant t, dont la distribution pour chaque état aura été apprise grâce aux annotations de SwissProt et aux alignements générés grâce à Ensembl.

10.2.3.2 Différents systèmes de notation (« scoring ») de précurseurs

On a développé divers systèmes pour noter les precurseurs à PH candidats. Une première idée, sans doute la plus intuitive, consiste en l'évaluation de la probabilité d'observer la séquence sous le modèle (*overall score*) que l'on normalise par la longueur de la séquence. Un deuxième score, que l'on calcule en estimant l'espérance du nombre de clivages sur la séquence, normalisée par sa longueur (*densitycleavage score*), favorise les protéines ayant plusieurs sites de clivage potentiels. Un troisième score privilégie les séquences qui ont au moins un peptide très probable (*bestpeptide score*). Enfin, on a considéré un score global (*mixed score*), qui est une combinaison linéaire des trois.

On a mesuré la performance des prédictions (sensibilité et la spécificité) sur l'ensemble des protéines sécrétées du génome humain (environ 1350 protéines) dans un schéma expérimental de validation croisée (50X)

L'algorithme correspondant à la meilleure combinaison type de modèle/système de scoring (CSA-HMM/mixedscore) a été ensuite utilisé pour prédire de nouveaux précurseurs à PH dans le protéome humain.

10.2.3.3 Résultat du criblage sur le protéome humain

L'application de l'algorithme à des alignements de protéines pour l'ensemble du protéome humain (environ 25000) nous a permis d'établir une liste de gènes candidats codant potentiellement pour des précurseurs à PH. Parmi les 200 protéines qui arrivent en tête du classement, figure plus de 50% de précurseurs à hormones peptidiques. Parmi ces candidats, deux d'entre eux, que l'ont a nommés spexine et augurine, sont apparus comme étant particulièrement prometteurs.

10.3 Les candidats : spexine et augurine

En effet, l'examen des alignements de protéines de vertébrés correspondant aux gènes de la spexine et de l'augurine nous a révélé que les caractéristiques de PH attendues chez ces protéines sont clairement conservées chez les vertébrés : l'on peut reconnaître que la spexine, comme l'augurine, possède un peptide signal à son N-terminus et plusieurs sites dibasiques (Arg/Lys-Arg/Lys) qui sont potentiellement des sites de reconnaissance pour les prohormones convertases. De plus, l'on peut distinguer sans ambiguïté, dans ces deux protéines, une région fortement conservée par rapport au reste de la protéine, et dont une des frontières correspond

justement à un des sites de clivage potentiels. Ces régions fortement conservées correspondent aux peptides potentiels. Par ailleurs, l'un des peptide potentiels de la spexine serait amidé, comme l'est environ la moitié des hormones peptidiques.

10.3.1 Résultats

La spexine et l'augurine exhibent toutes les caractéristiques attendues d'une hormone : un peptide signal, des sites de clivage conservés chez les vertébrés, et des peptides potentiels qui sont hautement conservés, en comparaison avec les parties flanquantes (qui ont été désignées abusivement « propeptide » en reprenant la nomenclature de SwissProt).

10.3.1.1 Sécrétion et clivage des protéines spexine et augurine

De l'ADN codant pour le peptide FLAG (DYKDDDDK) a été introduit à différents endroits de l'ADN codant pour la spexine et l'augurine. Ces inserts ont ensuite été clonés dans un vecteur d'expression (PCDNA 3.1). Les constructions génétiques ainsi obtenues ont ensuite servi à exprimer les protéines de fusion spexine-FLAG et augurine-FLAG dans des lignées cellulaires dérivées de cellules bêta d'îlots de Langerhans pancréatiques. Deux séries d'expérience, pour chacun de ces deux gènes, ont ensuite été conduites. La première a consisté en la récupération de surnageant après transfection des constructions permettant l'expression des protéines de fusion (« conditioned supernatant »), et l'analyse par Western blot des fragments de protéines FLAG, en utilisant des anticorps contre le peptide FLAG (anticorps M1 et M2 développés par Sigma-Aldrich). Cette analyse nous a révélé que la spexine et l'augurine sont sécrétées *in vitro*, car l'on obtient des bandes spécifiques correspondant à l'échantillon de surnageant de cellules endocrines, après transfection des constructions génétiques spexine et augurine-FLAG.

De plus, la spexine, comme l'augurine, sont vraisemblablement découpées de manière analogue autres hormones peptidiques, comme le suggère la multiplicité de bandes détectées lors des Western Blot. La position exacte du ou des sites de clivage de la spexine n'a pas pu être déterminée avec certitude, mais l'on a cependant pu réduire les champs de possibilités à une région spécifique. Cependant, en utilisant le fait que l'anticorps M1 ne reconnaît le peptide FLAG que s'il est situé au N-terminus d'une protéine, nous avons pu déterminer avec exactitude la position d'un des sites de clivage de l'augurine. Cette première série d'expériences nous a permis de conclure que la spexine et l'augurine étaient des protéines sécrétées et découpées (« processed »), *in vitro.*

Dans une deuxième série d'expérience, où nous avons repris le même schéma expérimental (i.e. transfections dans des cellules pancréatiques de constructions FLAG), nous avons regardé par immunofluorescence la localisation sous-cellulaire des fragments de protéines FLAG exprimées par les cellules. Nous avons pu observer que les protéines de fusion étaient localisées dans des vésicules de sécrétion, appelées granules a cœur dense (DCG), car elles se retrouvent dans des structures dites ponctuelles (punctate structure) ou l'on a pu détecter l'insuline. En effet, l'insuline est co-localisée avec les protéines spexine et augurine-FLAG dans des proportions semblables au degré de co-localisation que l'on observe entre l'insuline et les protéines de fusion TKN-FLAG, où TKN-FLAG est un neuropeptide connu auquel on a flanqué un peptide FLAG (notre témoin positif).

10.3.1.2 Expression de la spexine et de l'augurine chez la souris

Chez la souris, elle est exprimée dans la couche muqueuse de système gastro-oesophagien, dont le sphincter oesophagien inférieur (LES) et le fundus gastrique. Elle est aussi présente dans le cerveau murin, dont les noyaux du tegment dorso-latéral (LDT), pedunculopontine (PPT) et l'habénula latérale.

L'augurine, quant à elle, est exprimée chez la souris dans diverses glandes endocrines, dont les glandes surrénales et l'adénohypophyse, et certainement la glande thyroïdienne. Dans le cerveau, elle est exprimée dans les organes circumventriculaires (CVO) du cerveau de la souris, en particulier le plexus choroïdien et l'organe sous-commissural. Dans tous ces organes, elle semble être produite par des cellules épithéliales. Par hybridation in situ (ISH) et immunohistochimie (IHC), nous avons pu montrer que l'augurine était abondamment présente dans la couche épendymaire des ventricules du cerveau. L'augurine est également présente dans le courant migratoire rostral (RMS) qui part de la zone médiane du troisième ventricule pour arriver jusqu'au bulbe olfactif. Ce résultat présente un intérêt particulier car cette région du cerveau a été récemment caractérisée comme étant le lieu privilégie de la neurogénèse (Alvarez-Buylla and Garcia-Verdugo 2002) chez les mammifères, avec l'hippocampe. Dans le modèle présenté dans cet article, des cellules précurseur de neurones partent des ventricules, et migrent vers le bulbe olfactif afin de repeupler les populations de neurones olfactifs. L'augurine est aussi présente en très grande quantité dans les ostéoblastes, les precurseurs du cartilage, la valve aortique. Ces tissus ont en point commun qu'ils sont soumis à des contraintes mécaniques particulières. Chez l'embryon de souris, au stade développement E17 (quatre jours après la naissance), l'augurine est présente de manière claire

dans pratiquement tous les tissus mentionnés ci-dessus, dont les glandes surrénales, hypophysaire, thyroïdienne, les primordium cartilagineux, les ventricules et CVO du cerveau.

10.3.1.3 Etude fonctionnelle d'un peptide de la spexine, *ex vivo*

L'un des résultats les plus importants de ma thèse est certainement que l'un des peptides issu du précurseur de la spexine (peptide amidé de 14 acides aminés), contracte l'estomac de rats isolés, et ce, de manière dose-dépendante et à des amplitudes similaires à celles enregistrées en présence d'acétylcholine.

10.3.2 Discussion

La spexine étant exprimée dans le sphincter LES et dans le système gastro-osophagien, et ayant une action dans les tissus du fundus gastrique dans notre modèle d'estomac de rat isolé, il est possible qu'elle ait une fonction dans la péristaltisme de l'œsophage et la vidange gastrique. Il est également tentant de spéculer sur l'implication de la spexine dans des maladies liées a un dysfonctionnement du LES ; tels que les GERD (gastro-esophageal reflux disease). Par ailleurs, les trois régions du cerveau où la spexine est exprimée (PPT, LDT et habenula) ont en commun le fait d'être associées au système dopaminergique du mésencephalon (Lodge and Grace 2006; Matsumoto and Hikosaka 2007; Mena-Segovia et al. 2004). Il n'est donc pas déraisonnable de penser que la spexine puisse avoir une fonction neuromodulatrice dans les synapses afférentes aux neurones dopaminergiques du mésencephalon. De plus le noyau PPT a été associé a la maladie de Parkinson (Hunter et al. 1997; Pahapill and Lozano 2000), dont plusieurs symptômes concernent un dérèglement gastro-intestinal (Pfeiffer 2003) (lieu d'expression de spexine) et un dérèglement du système peptidergique (potentiel) à spexine pourrait ainsi avoir des conséquences sur cette maladie.

Il est encore trop tôt pour proposer des hypothèses sérieuses quant à la fonction d'augurine, alors qu'aucune étude fonctionnelle n'a encore été présentée sur d'éventuels peptides issus de l'augurine et qui auraient une activité biologique. Cependant, l'on a plusieurs éléments qui nous laisse penser qu'elle pourrait avoir une fonction paracrine dans la maintenance des niches de cellules souches et le cancer (Fuchs et al. 2004). Certainement, sa structure primaire ressemble plus à celle d'IGF1, un peptide qui est aussi un facteur de croissance.

10.4 Conclusion

Nous avons développé des algorithmes spécifiques pour découvrir de nouvelles hormones peptidiques. Ces algorithmes permettent de prédire la grande majorité des hormones peptidiques connues. Elles nous ont permis de dégager deux gènes candidats, la spexine et l'augurine qui ont de fortes caractéristiques de précurseurs à hormones peptidiques (conservation chez les vertébrés restreinte à une partie seulement de la protéine, sécrétion et clivage *in vitro*, localisation sous-cellulaire dans les DCG). Nous ne pouvons cependant pas, à l'heure actuelle, affirmer que la spexine et l'augurine sont bien des précurseurs à PH. Une étape intermédiaire importante vers cette conclusion serait l'identification ces peptides *in vivo*. Cependant, seuls des tests pharmacologiques (e.g. injection de peptides dans l'animal IV ou ICV et contrôle de la vidange gastrique), des expériences génétiques (e.g. génération de souris transgéniques ou KO), et biochimiques (e.g. activation de GPCR, *in vitro*) apporteront un réponse définitive à cette question.

11 Appendix

APPENDIX A: Top 200 candidate PH

Ensembl peptide ID	rank	description	annotation
ENSP00000303452	1	Thyroliberin precursor	peptide hormone
ENSP00000176195	2	Secretin precursor	peptide hormone
ENSP00000356213	3	VIP peptides precursor	peptide hormone
ENSP00000356212	4	VIP peptides precursor	peptide hormone
ENSP00000165524	5	Prolactin-releasing peptide precursor	peptide hormone
ENSP00000325286	6	Tachykinin 4 isoform delta	peptide hormone
ENSP00000324248	7	Proenkephalin A precursor	peptide hormone
ENSP00000340461	8	tachykinin 4 isoform delta	peptide hormone
ENSP00000331358	9	Gastrin precursor	peptide hormone
ENSP00000289574	10	Protachykinin 1 precursor	peptide hormone
ENSP00000217305	11	Beta-neoendorphin-dynorphin precursor	peptide hormone
ENSP00000256969	12	Spexin	poorly annotated protein
ENSP00000293330	13	Orexin precursor	peptide hormone
ENSP00000218230	14	ProSAAS precursor	contains putative peptides
ENSP00000321106	15	Beta preprotachykinin I	peptide hormone
ENSP00000233604	16	Glucagon precursor	peptide hormone
ENSP00000353664	17	Neuromedin-B precursor	peptide hormone
ENSP00000307040	18	Relaxin-3 precursor	peptide hormone
ENSP00000278175	19	ADM precursor	peptide hormone
ENSP00000302724	20	Insulin-like peptide INSL5 precursor	peptide hormone
ENSP00000320951	21	Beta-defensin 103A precursor	defensin
ENSP00000324633	22	Beta-defensin 103A precursor	defensin
ENSP00000242152	23	Neuropeptide Y precursor	peptide hormone
ENSP00000353198	24	Peptide YY precursor	peptide hormone
ENSP00000332766	25	Neuropeptide B precursor	peptide hormone
ENSP00000225992	26	Pancreatic prohormone precursor	peptide hormone
ENSP00000334042	27	tachykinin 4 isoform delta	peptide hormone
ENSP00000297496	28	Sperm associated antigen 11B isoform H precursor	other secreted protein
ENSP00000354411	29	Sperm-associated antigen 11 precursor	other secreted protein
ENSP00000296099	30	Urocortin precursor	peptide hormone
ENSP00000369412	31	Urocortin precursor	peptide hormone
ENSP00000330070	32	Neuropeptide W precursor	peptide hormone
ENSP00000276571	33	Corticoliberin precursor	peptide hormone
ENSP00000201015	34	Parathyroid hormone-related protein precursor	peptide hormone
ENSP00000250971	35	Insulin precursor	peptide hormone
ENSP00000370720	36	Insulin precursor	peptide hormone
ENSP00000370731	37	Insulin precursor	peptide hormone
ENSP00000287641	38	Somatostatin precursor	peptide hormone
ENSP00000345487	39	Orexigenic neuropeptide QRFP precursor	peptide hormone
ENSP00000350269	40	Fibroblast growth factor 5 precursor	peptide hormone
ENSP00000365651	41	Natriuretic peptides B precursor	peptide hormone
ENSP00000346398	42	Parathyroid hormone-related protein precursor	peptide hormone
ENSP00000364176	43	Matrix-remodelling-associated protein 7	other secreted protein
ENSP00000307650	44	Metastasis-suppressor KiSS-1 precursor	peptide hormone
ENSP00000350236	45	Neurokinin-B precursor	peptide hormone
ENSP00000356162	46	Metastasis-suppressor KiSS-1 precursor	peptide hormone

ENSP00000300108	47	Neurokinin-B precursor	peptide hormone
ENSP00000367600	48	chemokine (C-C motif) ligand 4-like 2 precursor	cytokine and growth factor
ENSP00000307954	49	Transmembrane protein 157 precursor	membrane
ENSP00000369156	50	Transmembrane protein 157 precursor	membrane
ENSP00000365781	51	ProSAAS precursor	peptide hormone
ENSP00000296877	52	Liver-expressed antimicrobial peptide 2 precursor	peptide hormone
ENSP00000367924	53	uncharacterized	poorly annotated protein
ENSP00000345572	54	chemokine (C-C motif) ligand 4-like 2 precursor	cytokine and growth factor
ENSP00000289576	55	Protachykinin precursor	peptide hormone
ENSP00000267017	56	FMRFamide-related peptides precursor	peptide hormone
ENSP00000346017	57	Calcitonin gene-related peptide 2 precursor	peptide hormone
ENSP00000297498	58	Sperm-associated antigen 11 precursor	other secreted protein
ENSP00000295440	59	C-type natriuretic peptide precursor	peptide hormone
ENSP00000264218	60	Neuromedin-U precursor	peptide hormone
ENSP00000370717	61	Neuromedin-U precursor	peptide hormone
ENSP00000269200	62	Pituitary adenylate cyclase-activating polypeptide precursor	peptide hormone
ENSP00000256857	63	Gastrin-releasing peptide precursor	peptide hormone
ENSP00000350005	64	Gastric inhibitory polypeptide precursor	peptide hormone
ENSP00000365644	65	Natriuretic peptides B precursor	peptide hormone
ENSP00000287656	66	Appetite-regulating hormone precursor	peptide hormone
ENSP00000215781	67	Oncostatin-M precursor	cytokine and growth factor
ENSP00000308018	68	Prorelaxin H2 precursor	peptide hormone
ENSP00000264708	69	Corticotropin-lipotropin precursor	peptide hormone
ENSP00000370171	70	Corticotropin-lipotropin precursor	peptide hormone
ENSP00000223858	71	Prorelaxin_H1_	peptide hormone
ENSP00000335657	72	Cholecystokinins precursor	peptide hormone
ENSP00000334122	73	INT-2 proto-oncogene protein precursor	cytokine and growth factor
ENSP00000296777	74	Cocaine- and amphetamine-regulated transcript protein precursor	peptide hormone
ENSP00000355881	75	MOSC domain-containing protein 2, mitochondrial precursor	mitochondrial
ENSP00000238044	76	Esophageal cancer related gene 4 protein precursor/Augurin	poorly annotated protein
ENSP00000301908	77	Nociceptin precursor	peptide hormone
ENSP00000370716	78	Neuromedin-U precursor	peptide hormone
ENSP00000276414	79	Progonadoliberin precursor	peptide hormone
ENSP00000372331	80	CH455 (Fragment)	poorly annotated protein
ENSP00000351345	81	RPLK9433	poorly annotated protein
ENSP00000316012	82	sperm associated antigen 11B isoform H precursor	other secreted protein
ENSP00000361821	83	Collagen alpha-2(IX) chain precursor	other secreted protein
ENSP00000352663	84	Calcitonin gene-related peptide 1 precursor	peptide hormone
ENSP00000348050	85	Matrix-remodelling-associated protein 7	membrane
ENSP00000356968	86	Apolipoprotein A-II precursor	other secreted protein
ENSP00000348930	87	Chordin precursor	other secreted protein
ENSP00000308559	88	Chordin precursor	other secreted protein
ENSP00000374028	89	Lymphocyte antigen 6H precursor	membrane
ENSP00000352797	90	Sperm-associated antigen 11 precursor	other secreted protein
ENSP00000367952	91	Uncharacterized protein C10orf49 precursor	poorly annotated protein
ENSP00000354067	92	Protein CASC4	membrane
ENSP00000365511	93	null	poorly annotated protein
ENSP00000353693	94	Transmembrane protein 119 precursor	membrane
ENSP00000216492	95	Chromogranin A precursor	peptide hormone

ENSP00000369647	96	Vasopressin-neurophysin 2-copeptin precursor	peptide hormone
ENSP00000367916	97	Liver-expressed antimicrobial peptide 2 precursor	peptide hormone
ENSP00000265643	98	Galanin precursor	peptide hormone
ENSP00000339565	99	RGPG542	other secreted protein
ENSP00000311854	100	Endothelin-3 precursor	peptide hormone
ENSP00000369548	101	Uncharacterized protein C20orf116 precursor	poorly annotated protein
ENSP00000315764	102	Sperm-associated antigen 11 precursor	other secreted protein
ENSP00000322591	103	Sperm-associated antigen 11 precursor	other secreted protein
ENSP00000371662	104	Transmembrane protein 41A precursor	membrane
ENSP00000366738	105	Urotensin-2 precursor	peptide hormone
ENSP00000226284	106	Bone sialoprotein-2 precursor	cytokine and growth factor
ENSP00000345317	107	DMC	cytokine and growth factor
ENSP00000259631	108	Small inducible cytokine A27 precursor	cytokine and growth factor
ENSP00000365663	109	Atrial natriuretic factor precursor	peptide hormone
ENSP00000354204	110	TIMM9	mitochondrial
ENSP00000318437	111	Follicular dendritic cell secreted peptide precursor	cytokine and growth factor
ENSP00000356737	112	N-terminal kinase-like-binding protein 1	cytoplasmic
ENSP00000371040	113	Prorelaxin H1 precursor	peptide hormone
ENSP00000373371	114	FAM132B protein (Fragment)	poorly annotated protein
ENSP00000223862	115	Prorelaxin H1 precursor	peptide hormone
ENSP00000360064	116	Endothelin-3 precursor	peptide hormone
ENSP00000215530	117	Fibroblast growth factor 22 precursor	cytokine and growth factor
ENSP00000355163	118	Urotensin-2 precursor	peptide hormone
ENSP00000302924	119	Interleukin-17	cytokine and growth factor
ENSP00000331746	120	Calcitonin gene-related peptide 1 precursor	peptide hormone
ENSP00000337128	121	Endothelin-3 precursor	peptide hormone
ENSP00000360067	122	Endothelin-3 precursor	peptide hormone
ENSP00000356736	123	N-terminal kinase-like-binding protein 1	cytoplasmic
ENSP00000346483	124	Uncharacterized protein C20orf116 precursor	poorly annotated protein
ENSP00000369561	125	Uncharacterized protein C20orf116 precursor	poorly annotated protein
ENSP00000222304	126	Hepcidin precursor	peptide hormone
ENSP00000364962	127	G6b protein precursor	membrane
ENSP00000335481	128	EMI domain-containing protein 1 precursor	other secreted protein
ENSP00000354286	129	Calcitonin gene-related peptide 1 precursor	peptide hormone
ENSP00000364967	130	G6b protein precursor	membrane
ENSP00000311697	131	Fibroblast growth factor 5 precursor	cytokine and growth factor
ENSP00000263273	132	Nucleobindin-1	membrane
ENSP00000295619	133	Prokineticin-2	peptide hormone
ENSP00000302362	134	null	poorly annotated protein
ENSP00000368721	135	Neurokinin-B precursor	peptide hormone
ENSP00000360660	136	Neural proliferation differentiation and control protein 1 precursor	poorly annotated protein
ENSP00000361726	137	WAP four-disulfide core domain protein 10A precursor	other secreted protein
ENSP00000260595	138	Cell growth regulator with EF hand domain protein 1	poorly annotated protein
ENSP00000354955	139	ADM2 precursor	peptide hormone
ENSP00000339703	140	Glial cell line-derived neurotrophic factor precursor	cytokine and growth factor
ENSP00000365246	141	Novel protein similar to peptide YY	peptide hormone
ENSP00000365247	142	Novel protein similar to peptide YY	peptide hormone
ENSP00000246551	143	hematopoietic cell signal transducer isoform 1 precursor	cytokine and growth factor

ENSP00000313140	144	Full-length cDNA clone CS0DC026YP23 of Neuroblastoma	poorly annotated protein
ENSP00000349365	145	interleukin 27	cytokine and growth factor
ENSP00000302400	146	Submaxillary gland androgen-regulated protein 3 homolog B precursor	other secreted protein
ENSP00000355008	147	Thioredoxin-like selenoprotein M precursor	poorly annotated protein
ENSP00000340526	148	Urotensin-2B precursor	peptide hormone
ENSP00000217386	149	Oxytocin-neurophysin 1 precursor	peptide hormone
ENSP00000374172	150	Cell growth regulator with EF hand domain protein 1	cytokine and growth factor
ENSP00000371264	151	Basic salivary proline-rich protein 3 precursor	other secreted protein
ENSP00000290953	152	Agouti-related protein precursor	cytokine and growth factor
ENSP00000287020	153	Growth/differentiation factor 6 precursor	cytokine and growth factor
ENSP00000327506	154	Transmembrane protein C16orf54	membrane
ENSP00000302648	155	Neurturin precursor	cytokine and growth factor
ENSP00000301263	156	Lymphocyte antigen 6D precursor	membrane
ENSP00000370783	157	Insulin precursor	peptide hormone
ENSP00000364968	158	G6b protein precursor	membrane
ENSP00000342711	159	Lymphocyte antigen 6H precursor	membrane
ENSP00000317145	160	Glial cell line-derived neurotrophic factor precursor	cytokine and growth factor
ENSP00000371249	161	Glial cell line-derived neurotrophic factor precursor	cytokine and growth factor
ENSP00000257724	162	MyoD family inhibitor domain-containing protein	nuclear
ENSP00000054668	163	Urotensin-2 precursor	peptide hormone
ENSP00000328729	164	15 kDa selenoprotein precursor	ER lumen
ENSP00000338297	165	Insulin-like growth factor II precursor	peptide hormone
ENSP00000370786	166	Insulin-like growth factor II precursor	peptide hormone
ENSP00000370799	167	Insulin-like growth factor II precursor	peptide hormone
ENSP00000302938	168	Protein WFDC13 precursor	other secreted protein
ENSP00000368683	169	Endothelin-1 precursor	peptide hormone
ENSP00000360597	170	leucine rich repeat containing 26	poorly annotated protein
ENSP00000339067	171	Fibroblast growth factor-binding protein 3 precursor	cytokine and growth factor
ENSP00000313362	172	similar to MYC-associated zinc finger protein	nuclear
ENSP00000271331	173	Prokineticin-1	peptide hormone
ENSP00000245810	174	Persephin precursor	other secreted protein
ENSP00000272224	175	Growth/differentiation factor 7 precursor	cytokine and growth factor
ENSP00000370489	176	Growth/differentiation factor 7 precursor	cytokine and growth factor
ENSP00000300632	177	Insulin-like growth factor II precursor	peptide hormone
ENSP00000348986	178	Insulin-like growth factor II precursor	peptide hormone
ENSP00000370796	179	Insulin-like growth factor II precursor	peptide hormone
ENSP00000370802	180	Insulin-like growth factor II precursor	peptide hormone
ENSP00000370813	181	Insulin-like growth factor II precursor	peptide hormone
ENSP00000279573	182	Basic salivary proline-rich protein 3 precursor	other secreted protein
ENSP00000364971	183	G6b protein precursor	membrane
ENSP00000364206	184	Advanced glycosylation end product-specific receptor precursor	membrane
ENSP00000373069	185	Carbohydrate sulfotransferase 13	other secreted protein
ENSP00000228938	186	Matrix Gla-protein precursor	cytokine and growth factor
ENSP00000362447	187	small adipocyte factor 1	cytokine and growth factor
ENSP00000335397	188	reticulon 4 receptor-like 2	membrane
ENSP00000349884	189	Galanin-like precursor	peptide hormone
ENSP00000347547	190	proline rich 7 (synaptic)	other secreted protein

ENSP00000270294	191	Kin of IRRE-like protein 2 precursor	membrane
ENSP00000282479	192	Dentin matrix acidic phosphoprotein 1 precursor	other secreted protein
ENSP00000364201	193	Advanced glycosylation end product-specific receptor precursor	membrane
ENSP00000221498	194	Dickkopf-like protein 1 precursor	other secreted protein
ENSP00000324870	195	Serine protease inhibitor Kazal-type 6 precursor	other secreted protein
ENSP00000364982	196	lymphocyte antigen 6 complex	membrane
ENSP00000301411	197	Neurotrophin-5	cytokine and growth factor
ENSP00000348862	198	Sperm-associated antigen 11 precursor	other secreted protein
ENSP00000297533	199	Transmembrane and ubiquitin-like domain-containing protein 1	membrane
ENSP00000364092	200	Agouti-signaling switch protein	cytokine and growth factor
ENSP00000298743	201	Growth-arrest-specific protein 1 precursor	cytokine and growth factor
ENSP00000321767	202	Growth-arrest-specific protein 1 precursor	cytokine and growth factor
ENSP00000199448	203	Mammalian ependymin-related protein 1 precursor	cytokine and growth factor
ENSP00000297439	204	Beta-defensin 1 precursor	defensin
ENSP00000252809	205	Growth/differentiation factor 15 precursor	cytokine and growth factor
ENSP00000217425	206	WAP four-disulfide core domain protein 2 precursor	other secreted protein
ENSP00000368365	207	dermokine isoform beta	other secreted protein
ENSP00000353280	208	CD44 antigen precursor	membrane
ENSP00000354426	209	15 kDa selenoprotein precursor	ER lumen
ENSP00000246794	210	Transmembrane gamma-carboxyglutamic acid protein 2 precursor	membrane
ENSP00000293825	211	Tumor necrosis factor ligand superfamily member 12	cytokine and growth factor
ENSP00000369929	212	Tumor necrosis factor ligand superfamily member 12	cytokine and growth factor
ENSP00000327168	213	proline_rich_7	poorly annotated protein
ENSP00000363137	214	CD164 sialomucin-like 2 protein precursor	membrane
ENSP00000346283	215	LGLL338	membrane
ENSP00000357922	216	Uncharacterized protein C1orf56	poorly annotated protein
ENSP00000257383	217	Acrosomal protein SP-10 precursor	poorly annotated protein
ENSP00000354175	218	Uncharacterized protein C1orf56	poorly annotated protein
ENSP00000289823	219	Fibroblast growth factor 17	cytokine and growth factor
ENSP00000352414	220	Fibroblast growth factor 17	cytokine and growth factor
ENSP00000340719	221	Transmembrane protein 149 precursor	membrane
ENSP00000366965	222	Vascular_endot	poorly annotated protein
ENSP00000328325	223	Pituitary tumor-transforming gene 1 protein-interacting protein precursor	cytokine and growth factor
ENSP00000291934	224	MDAC1	membrane
ENSP00000318234	225	Collagen alpha-1(XXVI) chain precursor	other secreted protein
ENSP00000346934	226	CDNA_FLJ41553_	poorly annotated protein
ENSP00000256010	227	Neurotensin	peptide hormone
ENSP00000216085	228	Rhomboid domain-containing protein 3	membrane
ENSP00000342162	229	uncharacterized	poorly annotated protein
ENSP00000301464	230	Insulin-like growth factor-binding protein 6 precursor	other secreted protein
ENSP00000311127	231	Vascular endothelial growth factor B precursor	cytokine and growth factor
ENSP00000257504	232	Golgi phosphoprotein 2	membrane
ENSP00000365191	233	Golgi phosphoprotein 2	membrane
ENSP00000314144	234	Rhomboid domain-containing protein 2	membrane
ENSP00000294796	235	Fc receptor-like and mucin-like 1 precursor	membrane
ENSP00000367608	236	Carbonic anhydrase 9 precursor	membrane

ENSP00000371054	237	Insulin-like peptide 6	peptide hormone
ENSP00000295618	238	Prokineticin-2 precursor	cytokine and growth factor
ENSP00000304590	239	Immunoglobulin iota chain precursor	other secreted protein
ENSP00000256733	240	serum amyloid A2	other secreted protein
ENSP00000242465	241	Serglycin precursor	other secreted protein
ENSP00000294485	242	Uncharacterized protein C1orf187 precursor	poorly annotated protein
ENSP00000259989	243	Fibroblast growth factor-binding protein 2 precursor	cytokine and growth factor
ENSP00000371761	244	Fibroblast growth factor-binding protein 2 precursor	cytokine and growth factor
ENSP00000371765	245	Fibroblast growth factor-binding protein 2 precursor	cytokine and growth factor
ENSP00000373363	246	Golgi phosphoprotein 2	membrane
ENSP00000373364	247	Golgi phosphoprotein 2	membrane
ENSP00000365486	248	Beta-defensin 123 precursor	defensin
ENSP00000370473	249	Insulin-like growth factor-binding protein 3 precursor	other secreted protein
ENSP00000233813	250	Insulin-like growth factor-binding protein 3 precursor	other secreted protein
ENSP00000371901	251	Uncharacterized protein C19orf24	poorly annotated protein
ENSP00000367601	252	chemokine (C-C motif) ligand 4-like 2 precursor	cytokine and growth factor
ENSP00000204615	253	Thrombopoietin	cytokine and growth factor
ENSP00000368629	254	CDNA FLJ42060 fis	poorly annotated protein
ENSP00000359585	255	15 kDa selenoprotein precursor	ER lumen
ENSP00000327766	256	Granulocyte colony-stimulating factor precursor	cytokine and growth factor
ENSP00000290015	257	Protein Wnt-9b precursor	cytokine and growth factor
ENSP00000219782	258	Myc-associated zinc finger protein	nuclear
ENSP00000348918	259	Serum amyloid A protein precursor	other secreted protein
ENSP00000341335	260	Thrombopoietin	cytokine and growth factor
ENSP00000302657	261	PPWG6510	poorly annotated protein
ENSP00000274625	262	Fibroblast growth factor 18 precursor	cytokine and growth factor
ENSP00000370472	263	Insulin-like_g	cytokine and growth factor
ENSP00000259608	264	Signaling threshold-regulating transmembrane adapter 1 precursor	membrane
ENSP00000346773	265	mucin 1 isoform 5 precursor	other secreted protein
ENSP00000363693	266	tapasin isoform 1 precursor	membrane
ENSP00000371489	267	Complement C1q tumor necrosis factor-related protein 9 precursor	other secreted protein
ENSP00000371498	268	Complement C1q tumor necrosis factor-related protein 9 precursor	other secreted protein
ENSP00000278222	269	Serum_amyloid_	other secreted protein
ENSP00000252489	270	Apolipoprotein	other secreted protein
ENSP00000337065	271	Small inducible cytokine B14 precursor	cytokine and growth factor
ENSP00000368102	272	Coiled-coil domain-containing protein 3 precursor	other secreted protein
ENSP00000006053	273	Fractalkine precursor	membrane
ENSP00000344881	274	Linker for activation of T-cells family member 2	membrane
ENSP00000259870	275	Uncharacterized protein C6orf15 precursor	poorly annotated protein
ENSP00000322084	276	Beta-1,3-galactosyltransferase 6	membrane
ENSP00000368496	277	Beta-1,3-galactosyltransferase 6	membrane
ENSP00000361875	278	WAP four-disulfide core domain protein 5 precursor	other secreted protein
ENSP00000257386	279	Acrosomal protein SP-10 precursor	other secreted protein
ENSP00000359584	280	15 kDa selenoprotein precursor.	ER lumen
ENSP00000314455	281	Tumor necrosis factor ligand superfamily member 12	cytokine and growth factor

ENSP00000225474	282	Granulocyte colony-stimulating factor precursor	cytokine and growth factor
ENSP00000374458	283	MDAC1	membrane
ENSP00000374459	284	MDAC1	membrane
ENSP00000293826	285	Tumor necrosis factor ligand superfamily member 12	cytokine and growth factor
ENSP00000363696	286	tapasin isoform 1 precursor	membrane
ENSP00000331751	287	Signal transducer CD24 precursor	membrane
ENSP00000372291	288	Signal transducer CD24 precursor	membrane
ENSP00000275635	289	Linker for activation of T-cells family member 2	membrane
ENSP00000369798	290	Urocortin-3 precursor	other secreted protein
ENSP00000246532	291	transmembrane_	membrane
ENSP00000300177	292	Gremlin-1 precursor	other secreted protein
ENSP00000370725	293	Resistin precursor	other secreted protein
ENSP00000304422	294	Kremen protein 2 precursor	membrane
ENSP00000333988	295	resistance to inhibitors of cholinesterase 3 homolog	poorly annotated protein
ENSP00000370680	296	SPARC-related modular calcium-binding protein 1 precursor	other secreted protein
ENSP00000296695	297	Pancreatic secretory trypsin inhibitor precursor	other secreted protein
ENSP00000342411	298	neuritin 1-like	poorly annotated protein
ENSP00000329040	299	CDNA PSEC0264 fis	poorly annotated protein
ENSP00000315370	300	AVLV472	poorly annotated protein

APPENDIX B: Primers for cloning FLAG constructs

Oligonucleotide primers for spexin constructs:

Spexin-N-FLAG (c-terminal of first PC/furin cleavage site)
PCR#1
56#1F: 5’-caccatgaagggactcagaagtctg-3’
56#1R: 5’-**gtcgtcgtccttgtagtc**ccttctctccaacagtctctgc-3’

PCR#2
56#2F: 5’-**gactacaaggacgacgacgacaag**aactggactcctcaagctatgc-3’
56#2R: 5’-gacataatccagtatattttcac-3’

PCR#3
56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-gacataatccagtatattttcac-3’

Spexin-S-FLAG (behind signal peptide)
PCR#1
56#1F: 5’-caccatgaagggactcagaagtctg-3’
56#1R: 5’-**gtcgtcgtccttgtagtc**gcagctggagtttcccag-3’

PCR#2
56#2F: 5’-**gactacaaggacgacgacgacaag**gctccgcagagactgttg-3’
56#2R: 5’-gacataatccagtatattttcac-3’

PCR#3
56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-gacataatccagtatattttcac-3’

Spexin-C-FLAG (c-terminal)
PCR#1
56#1F: 5’-caccatgaagggactcagaagtctg-3’
56ctermR: 5’-tta**cttgtcgtcgtccttgtagtc**ccagttaagcagactgtcttc-3’

Spexin-N-FLAG/Δ51 (STOP at second putative PC/furin cleavage site)
PCR#1
56#1F: 5’-caccatgaagggactcagaagtctg-3’
56Stop1R1: 5’-ctaaccctgtgccccttcagg-3’

PCR#2
56Stop2F2: 5’-acctgaaaggggcacagggttagcgcttcatctccgaccagag-3’
56#2R: 5’-gacataatccagtatattttcac-3’

PCR#3
56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-gacataatccagtatattttcac-3’

Spexin-mutation N-FLAG/R35G

PCR#1

56#1F: 5’-caccatgaagggactcagaagtctg-3’
56R35GR1: 5’-cgtcgtcgtccttgtagtcccc-3’

PCR#2

56R35GF2: 5’-gactacaaggacgacgacg-3’
56#2R: 5’-gacataatccagtatattttcac-3’

PCR#3

56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-gacataatccagtatattttcac-3’

Spexin-mutation N-FLAG/R51L

PCR#1

56#1F: 5’-caccatgaagggactcagaagtctg-3’
56R51LR1: 5’-gaggcgaccctgtgccc-3’

PCR#2

56R51LF2: 5’-tgaaaggggcacagggt ctc cgcttcatctccgaccag-3’
56#2R: 5’-gacataatccagtatattttcac-3’

PCR#3

56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-gacataatccagtatattttcac-3’

Spexin-mutation N-FLAG/R60S

PCR#1

56#1F: 5’-caccatgaagggactcagaagtctg-3’
56R60SR1: 5’-gctccggctctggtcgga-3’

PCR#2

56R60SF2: 5’-atctccgaccagagccggagcaaggacctctccgaccgg-3’
56#2R: 5’-gacataatccagtatattttcac-3’

PCR#3

56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-gacataatccagtatattttcac-3’

Spexin-mutation N-FLAG/R72G

PCR#1

56#1F: 5’-caccatgaagggactcagaagtctg-3’
56R72G1R: 5’-tagtagttggggatttgggcttcctctttccggcagtggccg-3’

PCR#2

56R72G2F: 5’-agcccaaatccccaactacta-3’

56#2R: 5’-gacataatccagtatattttcac-3’

PCR#3

56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-gacataatccagtatattttcac-3’

Spexin-mutation N-FLAG/N36Q

PCR#1

56#1F: 5’-caccatgaagggactcagaagtctg-3’
56N361R: 5’-ttgcttgtcgtcgtcgtccttg-3’

PCR#2

56N36QF2: 5’-**aggacgacgacgacaag**caatggactcctcaagctatgc-3’
56#2R: 5’-gacataatccagtatattttcac-3’

PCR#3

56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-gacataatccagtatattttcac-3’

Spexin-mutation C-FLAG/R51G

PCR#1

56#1F: 5’-caccatgaagggactcagaagtctg-3’
56R35GR1: 5’-cgtcgtcgtccttgtagtcccc-3’

PCR#2

56R35GF2: 5’-gactacaaggacgacgacg-3’
56#2R: 5’-tta**cttgtcgtcgtcgtccttgtagtc**ccagttaagcagactgtcttc-3’

PCR#3

56#3F: 5’-caccatgaagggactcagaagtctg-3’
56#3R: 5’-tta**cttgtcgtcgtcgtccttgtagtc**ccagttaagcagactgtcttc-3’

Oligonucleotide primers for augurin constructs:

Augurin-N-FLAG (internal)

PCR#1

99#1F: 5’-caccatgagcacctcgtctgcg-3’
99#1R: 5’-**gtcgtcgtccttgtagtc**tcgtttggcacgcttcag-3’

PCR#2

99#2F: 5’-**gactacaaggacgacgacgacaag**cagctgtgggaccgtacg-3’
99#2R: 5’-tctgtgggcacctcagga-3’

PCR#3

99#3F: 5’-caccatgagcacctcgtctgcg-3’
99#3R: 5’-tctgtgggcacctcagga-3’

Augurin-C-FLAG (c-terminal)

PCR#1

99#1F: 5’-caccatgagcacctcgtctgcg-3’
99#1R: 5’-tta**cttgtcgtcgtcgtccttgtagtc**atagtcatcatagttgacact-3’

Augurin-mutation N-FLAG/R70L

PCR#1

99#1F: 5’-caccatgagcacctcgtctgcg-3’
99R70LR1: 5’- tagtttggcacgcttcaggcc-3’

PCR#2

99R70LF2: 5’-gcctgaagcgtgccaaactagactacaaggacgacgacg-3’
99#2R: 5’-tctgtgggcacctcagga-3’

PCR#3

99#3F: 5’-caccatgagcacctcgtctgcg-3’
99#3R: 5’-tctgtgggcacctcagga-3’

Augurin-mutation N-FLAG/R67G & Augurin mutation N-FLAG/R70L/R67G

PCR#1

99#1F: 5’-caccatgagcacctcgtctgcg-3’
99R67GR: 5’- cccttcaggccacctagga-3’

PCR#2

99R67GF: 5’-attcctaggtggcctgaagggtgccaaactagactacaagga-3’
99#2R: 5’-tctgtgggcacctcagga-3’

PCR#3

99#3F: 5’-caccatgagcacctcgtctgcg-3’
99#3R: 5’-tctgtgggcacctcagga-3’

Oligonucleotide-primers for other candidates:

36-Flag (c-terminal)
PCR#1
36#1F: 5’-caccatgtcctggaaggcgctga-3’
36#1R: 5’-tta**cttgtcgtcgtccttgtagtc**ggatgacacagcagggctc-3’

44-Flag (internal)
PCR#1
44#1F: 5’-cgggatccagaagatggtgctgcggc-3’
44#1R: 5’-**gtcgtcgtccttgtagtc**acgtcggataggccgcgact-3’

PCR#2
44#2F: 5’-**gactacaaggacgacgacgacaag**gcggcgcagctcacggcctt-3’
44#2R: 5’-gctctagacttgtccaatcacaagacttcaac-3’

PCR#3
44#3F: 5’-cgggatccagaagatggtgctgcggc-3’
44#3R: 5’-gctctagacttgtccaatcacaagacttcaac-3’

10-Flag (internal)
PCR#1
10#1F: 5’-cgggatcccttactccttcactatgagacttc-3’
10#1R: 5’-gctctagatta**cttgtcgtcgtcgtccttgtagtc**aggcaacatagtgtgtatg-3’

PCR#2
10#2F: 5’-**gactacaaggacgacgacgacaag**tagggcctcctccgcccag-3’
10#2R: 5’-gctctagaaggcaacatagtgtgtatg-3’

PCR#3
10#3F: 5’-ttacttgtcgtcgtcgtccttgtagtctatctgtgggagagccccag-3’
10#3R: 5’-tatctgtgggagagccccag-3’

51-Flag (c-terminal)
PCR#1
51#1F: 5’-cacccatgaagatcccaattcttcc-3’
51#1R: 5’-tta**cttgtcgtcgtcgtccttgtagtc**ctgggaatcaggagcggca-3’

APPENDIX C: A survey of vertebrate peptide hormone precursors

This list of vertebrate peptide hormone precursors includes information about the sequence, function and receptors of the peptides they contain. Whenever possible, they were grouped according to a relevant structure/organ of the body. This information was gathered using mainly annotations from SwissProt, data from recent publications on peptide hormones, and information from the medical biochemistry webpage at the Indiana University School of Medecine.

Note that most peptide hormones may have additional functions in other tissues than the one used in this classification. In particular many more peptide hormones could have been classified as neuropeptides. Note also that only peptide hormones were included in this list and not hormones/growth factors such as growth hormone, follicle-stimulating hormone, gonadotropin-stimulating hormone, luteinizing hormone, and placental hormones including chorionic gonadotropin and the placental lactogen.

The main features of human endogenous precursor sequences were highlighted : signal peptides in blue, PC cleavage sites (usually pairs of basic residues) in red and amidated c-terminal residues of mature peptides in cyan. The fully processed bioactive peptides were underlined, numbered and annotated.

Note that with the exception of amidation, post-translational modifications of peptides are not indicated.

Pituitary hormones

HORMONE	PRECURSOR STRUCTURE (Human) and ASSOCIATED RECEPTORS	FUNCTION
Oxytocin precursor (NEU1)	Processing of the precursor yields a polypeptide of 9 amino acids. The vasopressin peptide is the endogenous ligand of three GPCRs (V1AR, V1BR and V2R)	Causes uterine contraction, milk ejection in lactating females, responds to suckling reflex and estradiol and lowers steroid synthesis in testes.

MAGPSLACCLLGLLALTSA*CYIQNCPLG*$_1$GKRAAPDLDVRKCLPCGPGGKGRCFGPNICCAEELGCFVGTAEALRCQEENYLPSPCQSGQKACGSGGRCAVLGLCCSPDGCHADPACDAEATFSQR 1. oxytocin.		
Arginine-vasopressin precursor (AVP)	Processing of the precursor yields a polypeptide of 9 amino acids, homologous to the oxytocin peptide. Agonist of GPCR OXYR	Responds to osmoreceptor which senses extracellular [Na+], involved in blood pressure regulation and increases H_2O readsorption from distal tubules in kidney.
MPDTMLPACFLGLLAFSSA*CYFQNCPRG*$_1$GKRAMSDLELRQCLPCGPGGKGRCFGPSICCADELGCFVGTAEALRCQEENYLPSPCQSGQKACGSGGRCAAFGVCCNDESCVTEPECREGFHRRARASDRSNATQLDGPAGALLLRLVQLAGAPEPFEPAQPDAY 1. vasopressin.		
Proopiomelanocortin precursor (COLI)	Processing of the precursor yields several polypeptides, including the melanocyte-stimulating hormones (MSH) α (13 aa), β (18 aa), and γ (11 aa) and the adrenocorticotropin (ACTH, 39 aa). GPCRs ACTHR, MC3R, MC4R MC5R are receptors for ACTH, and MSH α, β and γ.	ACTH stimulates the adrenal glands to release cortisol (steroid). MSH (melanocyte-stimulating hormone) increases the pigmentation of skin by increasing melanin production in melanocytes. β-endorphin and Met-enkephalin are endogenous opiates.

MPRSCCSRSGALLLALLLQASMEVRGWCLESSQCQDLTTESNLLECIRACKPDLSAETP
MFPGNGDEQPLTENPRK*YVMGHFRWDRF*$_1$GRRNSSSSGSSGAGQKR*EDVSAGEDC*
GPLPEGGPEPRSDGAKPGPRE$_2$GKR*SYSMEHFRWGKPV*$_4$*GKKRRPVKVYPNG*
AEDESAEAFPLEF$_3$KR*ELTGQRLREGDGPDGPADDGAGAQADLEHSLLVAAE*
KKDEGPYRMEHFRWGSPPKD$_{6_5}$KR*YGGFM*$_8$*TSEKSQTPLVTLFKNAIIKNAYK*
KGE$_7$

1. MSH γ.
2. putative peptide.
3. ACTH.
4. MSH α.
5. lipotropin γ.
6. MSH β.
7. β-endorphin.
8. Met-enkephalin.

Hypothalamic hormones

HORMONE	STRUCTURE and RECEPTORS	FUNCTION
Corticotropin-releasing factor (CRF)	protein of 41 amino acids. 2 GPCRs splicing variants CRFR1 and CRFR2.	Acts on corticotrope to release ACTH and β-endorphin.
MRLPLLVSAGVLLVALLPCPPCRALLSRGPVPGARQAPQHPQPLDFFQPPPQSEQPQQP QARPVLLRMGEEYFLRLGNLNKSPAAPLSPASSLLAGGSGSRPSPEQATANFFRVLLQQL LLPRRSLDSPAALAERGARNALGGHQEAPERERR*SEEPPISLDLTFHLLREVLEMAR* *AEQLAQQAHSNRKLMEII*$_1$GK 1. corticoliberin.		
Gonadotropin-releasing factor-1 precursor (gonadoliberin-1, GON1)	Processing of the precursor yields a peptide of 10 amino acids. The receptor associated to gonadoliberin-1 is a GPCR (GNRHR).	Acts as a gonadotrope to release LH and FSH.
MKPIQKLLAGLILLTWCVEGCSS*QHWSYGLRPG*$_1$GKRDAENLIDSFQEIVKEVGQLAET QRFECTTHQPRSPLRDLKGALESLIEEETGQKKI 1. gonadoliberin-1.		

Gonadotropin-releasing factor-2 precursor (gonadoliberin-2, GON2)	Processing of the precursor yields a peptide of 10 amino acids. The specific receptor associated to gonadoliberin-2 is thought to be GPCR (GNRR2).	Stimulates the secretion of gonadotropins, luteinizing and follicle-stimulating hormones (LH and FSH).
MASSRRGLLLLLLLTAHLGPSEA*QHWSHGWYPG*$_1$GKRALSSAQDPQNALRPPGRALD TAAGSPVQTAHGLPSDALAPLDDSMPWEGRTTAQWSLHRKRHLARTLLTAAREPRPAP PSSNKV 1. gonadoliberin-2.		
Prolactin-releasing factor precursor (PRF or PRRP)	Processing of the precursor yields 2 homologous peptides of 20 and 31 aa that are endogenous ligands to GPCR PRLHR.	Acts as a lactotrope to release prolactin.
MKVLRAWLLCLLMLGLALRGAA*SRTHRHSMEIRTPDINPAWYASRGIRPVGRF*$_2$$_1$ GRRRATLGDVPKPGLRPRLTCFPLEGGAMSSQDG 1. PrRP-31. 2. PrRP-20.		
somatoliberin precursor (GHRH)	contains a peptide of 44 amino acids. Activates GPCR GHRHR.	Somatoliberin is released by the hypothalamus and acts on the adenohypophyse to stimulate the secretion of growth hormone.
MPLWVFFFVILTLSNSSHCS PPPPLTLRM RR*YADAIFTNSYRKVLGQLSARKLLQDI MSRQQGESNQERGARARL*$_1$GRQVDSMWAEQKQMELESILVALLQKHSRNSQG 1. somatoliberin		
Somatostatin precursor (SMS)	Processing of this precursor yields peptides of 14 and 28 amino acids. Those peptides are associated to 5 GPCRs (SSR1, SSR2, SSR3, SSR4, SSR5).	Somatostatin peptides inhibit growth hormone (GH) and thyroid stimulating hormone (TSH) secretion.

<table>
<tr><td colspan="3">MLSCRLQCALAALSIVLALGCVTGAPSDPRLRQFLQKSLAAAAGKQELAKYFLAELLSE
PNQTENDALEPEDLSQAAEQDEMRLELQRSANSNPAMAPRERK<u>AGCKNFFWKTF</u>
<u>TSC</u>$_{2_1}$
1. somatostatin-28.
2. somatostatin-14.</td></tr>
<tr><td>Corticostatin precursor (CORT)</td><td>Contains polypeptide of 17 or 29 amino acids. Binds to all 5 GPCRs of the somatostatin family (cf. somatostatin precursor).</td><td>inhibits cAMP production induced by forskolin through somatostatin receptors.</td></tr>
<tr><td colspan="3">MPLSPGLLLLLLSGATATAALPLEGGPTGRDSEHMQEAAGIRKSSLLTFLAWWFEWTSQ
ASAGPLIGEEAREVARR<u>QEGAPPQQSARR</u><u>DRMPCRNFFWKTFSSCK</u>$_{1_2}$
1. cortistatin-29.
2. cortistatin-17.</td></tr>
<tr><td>Thyroliberin precursor (TRH)</td><td>Contains multiple repeats of a 3aa-long amidated peptide. Thyroliberin is the ligand for a GPCR (TRFR).</td><td>Functions as a regulator of the biosynthesis of TSH and prolactin in the anterior pituitary gland and as a neurotransmitter and neuromodulator in the central and peripheral nervous systems.</td></tr>
<tr><td colspan="3">MPGPWLLLALALTLNLTGVPGGRAQPEAAQQEAVTAAEHPGLDDFLRQVERLLFLREN
IQRLQGDQGEHSASQIFQSDWLSKR<u>QHP</u>$_1$GKREEEEEEGVEEEEEEEGGAVGPHKR
<u>QHP</u>$_1$GRREDEASWSVDVTQHKR<u>QHP</u>$_1$GRRSPWLAYAVPKR<u>QHP</u>$_1$GRRLADPKAQRS
WEEEEEEEEREEDLMPEKR<u>QHP</u>$_1$GKRALGGPCGPQGAYGQAGLLLGLLDDLSRSQGA
EEKR<u>QHP</u>$_1$GRRAAWVREPLEE
1. thyroliberin.</td></tr>
<tr><td>Pituitary adenylate cyclase-activating polypeptide precursor (PACAP)</td><td>precursor of a polypeptide of 27 or 38 amino acids. Endogenous ligand for the VIP GPCR VIPR1.</td><td>Acts as a hypophysiotropic hormone by stimulating adenylate cyclase in pituitary cells. Acts as a neuromodulator and neurotransmitter in the CNS, and on several organs of the body, including the adrenal glands.</td></tr>
</table>

MTMCSGARLALLVYGIIMHSSVYSSPAAAGLRFPGIRPEEEAYGEDGNPLPDFDGSEPPG AGSPASAPRAAAAWYRPAGDVAHGILNEAYRKVLDQLSAGKHLQSLVARGVGGSLGGG AGDDAEPLSKRHSDGIFTDSYSRYRKQMAVKKYLAAVLGKRYKQRVKNK *HSDGIFTDSYSRYRKQMAVKKYLAAVL*$_2$*GKRYKQRVKNK*$_1$ GKRIAYLQRFECTT HQPRSPLRDLKGALESLIEEETGQKKI 1. PACAP-38. 2. PACAP-27.		
Metastasis-suppressor KISS-1 precursor (KISS1)	several homologous peptides of 10, 13, 14 and 54 aa. KISS peptides are agonists of GPCR KISSR/GPR54.	Activation of the receptor inhibits cell proliferation and cell migration, key characteristics of tumor metastasis. The hypothalamic KISS1/GPR54 system is a pivotal factor in central regulation of the gonadotropic axis at puberty and in adulthood.
MNSLVSWQLLLFLCATHFGEPLEKVASVGNSRPTGQQLESLGLLAPGEQSLPCTERKPA ATARLSRR*GTSLSPPPESSGSRQQPGLSAPHSRQIPAPQGAVLVQREKDLPNYN WNSFGLRF*$_4$$_3$$_2$$_1$ GKREAAPGNHGRSAGRGWGAGAGQ 1. metastin. 2. kisspeptin-14. 3. kisspeptin-13. 4. kisspeptin-10.		
Agouti gene-related protein precursor (AGRP)	Contains a protein of 112 amino acids and processing by prohormone convertases may release a peptide of 43 amino acids (AGRP(83-132)). AGRP has been shown to mediate its action by antagonising the melanocortin receptor MC4R.	Plays a role in weight homeostasis. May play a role in the regulation of melanocortin receptors within the hypothalamus and adrenal gland, and therefore in the central control of feeding.

MLTAAVLSCALLLALPATRG*AQMGLAPMEGIRRPDQALLPELPGLGLRAPLKKTTA EQAEEDLLQEAQALAEVLDLQDREPRSSRRCVRLHESCLGQQVPCCDPCATCY CRFFNAFCYCRKLGTAMNPCSRT*$_{2}$$_{1}$

1. agouti-gene related peptide.
2. agouti-gene related peptide (83-132).

Thyroid hormones

HORMONE	STRUCTURE AND RECEPTORS	FUNCTION
Calcitonin precursor (CALC).	Contains a protein of 32 amino acids which is the endogenous ligand of GPCR CALCR.	Produced in parafollicular C cells of the thyroid, regulates calcium and phosphate metabolism.

MGFQKFSPFLALSILVLLQAGSLHAAPFRSALESSPADPATLSEDEARLLLAALVQDYVQM KASELEQEQEREGSSLDSPRSKR*CGNLSTCMLGTYTQDFNKFHTFPQTAIGVGAP*$_{1}$ GKKRDMSSDLERDHRPHVSMPQNAN

1. calcitonin.

Parathyroid hormones

HORMONE	STRUCTURE AND RECEPTORS	FUNCTION
Parathyroid hormone precursor (PTH)	Proteolytic cleavage releases a protein of 84 amino acids associated to GPCR PTHR2.	PTH elevates calcium level by dissolving the salts in bone and preventing their renal excretion.

MIPAKDMAKVMIVMLAICFLTKSDGKSVKKR*SVSEIQLMHNLGKHLNSMERVEWL RKKLQDVHNFVALGAPLAPRDAGSQRPRKKEDNVLVESHEKSLGEADKADV NVLTKAKSQ*$_{1}$

1. parathyroid hormone.

Hormones and peptides of the Gut

HORMONE	STRUCTURE	FUNCTION
Gastric inhibitory polypeptide precursor (GIP)	Proteolytic cleavage of precursor produces a polypeptide of 42 amino acids, the endogenous ligand of GPCR GIPR.	It is a potent stimulator of insulin secretion and a relatively poor inhibitor of gastric acid secretion.
MVATKTFALLLLSLFLAVGLGEKKEGHFSALPSLPVGSHAKVSSPQPRGPR*YAEGTFISDYSIAMDKIHQQDFVNWLLAQKGKKNDWKHNIT*$_1$QREARALELASQANRKEEEAVEPQSSPAKNPSDEDLLRDLLIQELLACLLDQTNLCRLRSR 1. gastric inhibitory polypeptide.		
Secretin precursor (SECR)	Contains a peptide of 27 amino acids which belongs to the glucagon family. Its action is mediated by the transduction of GPCR SCTR.	Secreted from the duodenum at pH values below 4.5. Stimulates formation of $NaHCO_3$-rich pancreatic juice and secretion of $NaHCO_3$-rich bile and inhibits hydrochloric acid production by the stomach.
MAPRPLLLLLLLLGGSAARPAPPRARR*HSDGTFTSELSRLREGARLQRLLQGLV*$_1$GKRSEQDAENSMAWTRLSAGLLCPSGSNMPILQAWMPLDGTWSPWLPPGPMVSEPAGAAAEGTLRPR 1. secretin.		
Gastrin precursor (GAST)	Contains several peptides sharing the same C-terminus of 52, 34, 17, 14 and 6 aa (the most potent form is gastrin-6). The peptides Activate GPCR GASR (receptor shared with cholecystokinin).	Gastrin stimulates the stomach mucosa to produce and secrete hydrochloric acid and the pancreas to secrete its digestive enzymes. It also stimulates smooth muscle contraction and increases blood circulation and water secretion in the stomach and intestine.

MQRLCVYVLIFALALAAFSEASWKPRSQQPDAPLGTGANR*DLELPWLEQQGPASHH RRQLGPQGPPHLVADPSKKQGPWLEEEEEAYGWMDF*$_{5}$$_{4}$$_{3}$$_{2}$$_{1}$ GRRSAEDEN 1. gastrin-52 2. big gastrin (34). 3. gastrin-17. 4. gastrin-14. 5. gastrin-6.		
Cholecystokinin precursor (CCKN)	Contains several peptides sharing the same C-terminus of 39, 33, 12 and 8 aa (the most potent form is CCK-8). CCK peptides act through GPCRs GASR and CCKAR.	They stimulate gallbladder contraction and bile flow, increase secretion of digestive enzymes from pancreas. They are widely expressed in the brain and their function in the CNS is still unclear.
MNSGVCLCVLMAVLAAGALTQPVPPADPAGSGLQRAEEAPRRQLR*VSQRTDGESRAH LGALLARYIQQARKAPSGRMSIVKNLQNLDPSHRISDRDYMGWMDF*$_{5}$$_{4}$$_{3}$$_{2}$$_{1}$ GRRS AEEYEYPS 1. cholecystokinin-58. 2. cholecystokinin-39. 3. cholecystokinin-33. 4. cholecystokinin-12. 5. cholecystokinin-8.		
Ghrelin precursor (GHRL)	Contains a bioactive peptide named ghrelin, of 27 or 28 amino acids. Has been recently reported to contain a second endogenous peptide, obestatin, which is amidated and 23 aa-long. Ghrelin activates GPCR GHSR.	Ghrelin is the ligand for the growth hormone secretagogue receptor type 1 (GHSR) inducing the release of growth hormone from the pituitary. It has an appetite-stimulating effect, induces adiposity and stimulates gastric acid secretion. Involved in growth regulation. Obestatin is a novel peptide hormone which has been reported to contrast the effect of ghrelin peptide and reduce appetite in rats.

MPSPGTVCSLLLLGMLWLDLAMA*GSSFLSPEHQRVQQRKESKKPPAKLQP*$_2$*R*$_1$ ALA GWLRPEDGGQAEGAEDELEVR*FNAPFDVGIKLSGVQYQQHSQAL*$_3$GKFLQDILWEE AKEAPADK

1. ghrelin-28.
2. ghrelin-27.
3. obestatin.

Motilin precursor (MOTI)	Contains a peptide of 22 amino acids. Acts through GPCR MLNR (GPR38).	Plays an important role in the regulation of interdigestive gastrointestinal motility and indirectly causes rhythmic contraction of duodenal and colonic smooth muscle. Has antibiotics properties.

MVSRKAVAALLVVHVAAMLASQTEA*FVPIFTYGELQRMQEKERNKGQ*$_1$KK*SLSVW QRSGEEGPVDPAEPIREEENEMIKLTAPLEIGMRMNSRQLEKYPATLEGLLSE MLPQHAAK*$_2$

1. motilin.
2. motilin-associated peptide.

Vasoactive intestinal peptide (VIP)	Contains a peptide of 28 amino acids. Belongs to the glucagon family. Acts through GPCRs VIPR1 and VIPR2.	Produced by the hypothalamus and gastrointestinal tract, relaxes the gut, inhibits acid and pepsin secretion, acts as a neurotransmitter in peripheral autonomic nervous system, increases secretion of H_2O and electrolytes from pancreas and gut.

MDTRNKAQLLVLLTLLSVLFSQTSAWPLYRAPSALRLGDRIPFEGANEPDQVSLKEDIDM LQNALAENDTPYYDVSRNAR*HADGVFTSDFSKLLGQLSAKKYLESLM*$_2$*GKRVSSN ISEDPVPV*$_1$KR*HSDAVFTDNYTRLRKQMAVKKYLNSIL*N_3GKRSSEGESPDFPEE LEK

1. intestinal peptide PHM-42.
2. intestinal peptide PHM-27.
3. vasoactive intestinal peptide.

Protackykinin-1 (TKN1) precursor	Peptides occurring from this precursor ar homologous and include substance P (11 aa), neurokinin A (or substance K, 10 aa) and Neuropeptide K (NPK, 36 aa). Substance P activates mainly the GPCR NK1R, while NKA is associated to the GPCR NK2R.	substance P is involved in pain (nociception), vomit reflex, it stimulates salivary secretions, induces vasodilation antagonists, has anti-depressant properties, causes rapid contractions of the gastrointestinal smooth muscle, and modulates inflammatory and immune responses. All tachykinins excite neurons, evoke behavioral responses, are potent vasodilators and secretagogues, and contract (directly or indirectly) many smooth muscles.
MKILVALAVFFLVSTQLFAEEIGANDDLNYWSDWYDSDQIKEELPEPFEHLLQRIAR *RPKPQQFFGLM*$_1$GKR*DADSSIEKQVALLKALYGHGQISHKRHKTDSFVGLM*$_{3_2}$G KRALNSVAYERSAMQNYERRR 1. substance P. 2. neuropeptide K. 3. neurokinin A.		
Neuropeptide Tyrosine (NPY, NEUY)	Contains a peptide of 36 amino acids. At least 4 verified GPCR associated (NPY1R, NPY2R, NPY4R, NPY5R) and 1 probable (GPR83)	Effects on hypothalamic function in appetite, controls feeding behavior and energy homeostasis. Levels of NPY increase during starvation to induce food intake.
MLGNKRLGLSGLTLALSLLVCLGALAEA*YPSKPDNPGEDAPAEDMARYYSALRHYI* *NLITRQRY*$_1$GKR*SSPETLISDLLMRESTENVPRTRLEDPAMW*$_2$ 1. neuropeptide Y. 2. C-flanking peptide of NPY.		

Pancreatic hormone precursor (PAHO)	Contains a peptide which belongs to the pancreatic polypeptide family of 36 amino acid peptides (with NPY and PPY). Acts through the same GPCRs as NPY (NPY1R, NPY2R, NPY4R, NPY5R)	Pancreatic hormone is synthesized in pancreatic islets of Langerhans and acts as a regulator of pancreatic and gastrointestinal functions.
MAAARLCLSLLLLSTCVALLLQPLLGAQGA*PLEPVYPGDNATPEQMAQYAADLRRY INMLTRPRY*$_1$GKR*HKEDTLAFSEWGSPHAAVPR*$_2$*ELSPLDL* 1. pancreatic hormone. 2. pancreatic icosapeptide.		
peptide tyrosine-tyrosine precursor (PYY)	Contains a peptide of 36 amino acids. Acts through the same GPCRs as NPY.	Inhibits gastric motility by inhibiting cholinergic neurotransmission, inhibits gastric acid secretion.
MVFVRRPWPALTTVLLALLVCLGALVDA*YPIKPEAPGEDASPEELNRYYASLRHYL NLVTRQRY*$_2$$_1$GKRDGPDTLLSKTFFPDGEDRPVRSRSEGPDLW 1. peptide YY (34). 2. peptide YY.		
Gastrin-releasing peptide precursor (GRP)	Processing of the precursor yields two peptides, gastrin-releasing peptide (27 aa) and neuromedin C (10 aa). The peptides are associated to GPCR GRPR.	GRP peptides stimulate gastrin release as well as other gastrointestinal hormones. They Operate as a negative feedback regulating fear and established a causal relationship between GRP-receptor gene expression, long-term potentiation, and amygdala-dependent memory for fear.
MRGSELPLVLLALVLCLAPRGRA*VPLPAGGGTVLTKMYPRGNHWAVGHLM*$_2$$_1$GKK STGESSSVSERGSLKQQLREYIRWEEAARNLLGLIEAKENRNHQPPQPKALGNQQPSWD SEDSSNFKDVGSKGKVGRLSAPGSQREGRNPQLNQQ 1. gastrin-releasing peptide. 2. neuromedin C.		

Neuromedin U precursor (NMU)	The precursor contains 1 peptide of 25 amino acids. NMU binds two GPCRs, NMUR1 and NMUR2.	Stimulates muscle contractions of specific regions of the gastrointestinal tract. In humans, NMU stimulates contractions of the ileum and urinary bladder.
MLRTESCRPRSPAGQVAAASPLLLLLLLLAWCAGACRGAPILPQGLQPEQQLQLWNEIDDTCSSFLSIDSQPQASNALEELCFMIMGMLPKPQEQDEKDNTKRFLFHYSKTQKLGKSNVVSSVVHPLLQLVPHLHERR*MKRFRVDEEFQSPFASQSRGYFLFRPRN*$_1$GRRSAGFI 1. neuromedin U.		
Neuromedin B precursor (NMB)	Upon proteolytic processing of the precursor 2 peptides of 32 (NMB-32) and 10 (NMB) amino acids are released . The NMB peptides are associated to the GPCR NMBR.	NMB peptides stimulate smooth muscle contraction in a manner similar to that of bombesin.
MARRAGGARMFGSLLLFALLAAGV*APLSWDLPEPRSRASKIRVHSRGNLWATGHFM*$_2$$_1$GKKSLEPSSPSPLGTATHTSLRDQRLQLSHDLLGILLLKKALGVSLSRPAPQIQYRRLLVQILQK 1. neuromedin B-32. 2. neuromedin B.		

Pancreatic hormones

HORMONE	STRUCTURE AND RECEPTORS	FUNCTION
Insulin precursor (INS)	Releases disulfide bonded dipeptide of 21 and 30 amino acids. Receptor is a tyrosine-protein kinase (INSR)	produced by β-cells of the pancreas, increases glucose uptake and utilization, increases lipogenesis. Accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver. General anabolic effects.
MALWMRLLPLLALLALWGPDPAAA*FVNQHLCGSHLVEALYLVCGERGFFYTPKT*$_1$ *RREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKR*GIVEQCCTSICSLYQLENYCN$_2$ 1. insulin B chain. 2. insulin A chain.		
Glucagon pecursor (GLUC).	Contains glucagon and glucagon-like peptides (GLPs). Glucagon binds to GPCR GLR, glucagon-like peptides 1 and 2 are the endogenous ligands of respectively GPCRs GLP1R and GLP2R.	Glucagon plays a key role in glucose metabolism and homeostasis. It regulates blood glucose by increasing gluconeogenesis and decreasing glycolysis and counteracts the effect of insulin. GLP-1 is a potent stimulator of glucose-dependent insulin release and plays important roles in gastric motility and suppression of plasma glucagon levels. GLP-2 plays a key role in nutrient homeostasis, enhancing nutrient assimilation through enhanced gastrointestinal function and increasing nutrient disposal.
MKSIYFVAGLFVMLVQGSWQ*QDTEEKSRSLRSFSASQADPLSDPDQMNEDKR*<u>*HSQGTFTSDYSKYLDSRRAQDFVQWLMNT*</u>$_3$*KRNRNNIA* KR<u>*HDEFERHAEGTFTSDVSSYLEGQAAKEFIAWLVKGR*</u>$_5$$_4$ GRR<u>*DFPEEVAIVEELGRRHADGSFSDEMNTILDNLAARDFINWLIQTKITD*</u>$_6$RK 1. glicentin. 2. oxyntomodulin. 3. glucagon. 4. glucagon-like peptide 1. 5. glucagon-like peptide 1(7-36). 5. glucagon-like peptide 2.		

Islet amyloid polypeptide precursor (IAPP).	Processing of the precursor yields a peptide of 37 amino acids. Unequivocal evidence of a known IAPP receptor still lacks.	Selectively inhibits insulin-stimulated glucose utilization and glycogen deposition in muscle, while not affecting adipocyte glucose metabolism.
MGILKLQVFLIVLSVALNHLKATPIESHQVEKR*KCNTATCATQRLANFLVHSSNNFG AILSSTNVGSNTY*$_1$GKRNAVEVLKREPLNYLPL 1. islet amyloid polypeptide.		

Placental hormones

HORMONE	STRUCTURE	FUNCTION
Relaxins (e.g. REL1)	multigenic family of 2 proteins of 22 and 32 amino acids bound by disulfide bridges (like insulin). Activate GPCRs (ex. RL3R1).	Relaxins are an ovarian hormones that act with estrogen to produce dilatation of the birth canal in many mammals. May be involved in remodeling of connective tissues during pregnancy, promoting growth of pubic ligaments and ripening of the cervix.
MPRLFLFHLLEFCLLLNQFSRA*VAAKWKDDVIKLCGRELVRAQIAICGMSTWS*$_1$KR SLSQEDAPQTPRPVAEIVPSFINKDTETIIIMLEFIANLPPELKAALSERQPSLPELQQYVP ALKDSNLSFEEFKKLIRNRQSEAADSNPSELKYLGLDTHSQKKRR*PYVALFEKCCLIG CTKRSLAKYC*$_2$ 1. relaxin B chain. 2. relaxin A chain.		

Liver hormones

HORMONE	STRUCTURE AND RECEPTORS	FUNCTION
Angiotensinogen precursor (ANGT)	Contains angiotensin II, a polypeptide of 8 amino acids, after processing of the precursor by the proteases renin and angiotensin-converting enzyme (ACE).	It is a potent vasoconstrictor and is responsible for essential hypertension through stimulated synthesis and release of aldosterone from adrenal cells. Part of the renin- angiotensin-aldosterone system (RAAS) central to water-salt homeostasis.
MRKRAPQSEMAPAGVSLRATILCLLAWAGLAAG*DRVYIHPF*$_{3_2}$*HL* VIHNESTCEQLAK ANAGKPKDPTFIPAPIQAKTSPVDEKALQDQLVLVAAKLDTEDKLRAAMVGMLANFLG FRIYGMHSELWGVVHGATVLSPTAVFGTLASLYLGALDHTADRLQAILGVPWKDKNCT SRLDAHKVLSALQAVQGLLVAQGRADSQAQLLLSTVVGVFTAPGLHLKQPFVQGLALYT PVVLPRSLDFTELDVAAEKIDRFMQAVTGWKTGCSLMGASVDSTLAFNTYVHFQGKM KGFSLLAEPQEFWVDNSTSVSVPMLSGMGTFQHWSDIQDNFSVTQVPFTESACLLLIQP HYASDLDKVEGLTFQQNSLNWMKKLSPRTIHLTMPQLVLQGSYDLQDLLAQAELPAILH TELNLQKLSNDRIRVGEVLNSIFFELEADEREPTESTQQLNKPEVLEVTLNRPFLFAVYD QSATALHFLGRVANPLSTA 1. angiotensin I. 2. angiotensin II. 3. angiotensin III.		
Insulin-like growth factor family precursors (example: IGF1A and IGF2)	The IGF1A precursor contains a polypeptide of 70 amino acids. IGF1A and IGF2 are associated to tyrosine kinase receptor IGF1R.	IGF1A is a potent growth factor that can be secreted by the liver and is under the endocrine control of the growth hormone. Belongs to a large family of proteins structurally related to insulin, some of them generated from alternatively spliced form of the same genes IGF-1 and IGF-2. It can act as a paracrine molecule in several tissues and has been shown to be involved in skeletal muscle aging, regeneration, and disease.
MGKISSLPTQLFKCCFCDFLKVKMHTMSSSHLFYLALCLLTFTSSATA*GPETLCGAELV DALQFVCGDRGFYFNKPTGYGSSSRRAPQTGIVDECCFRSCDLRRLEMYCAPL KPAKSA*$_1$RSVRAQRHTDMPKTQKEVHLKNASRGSAGNKNYRM 1. insulin-like growth factor 1A (IGF1A).		

Cardiac hormones

HORMONE	STRUCTURE AND RECEPTORS	FUNCTION
Atrial natriuretic peptide precursor (ANP)	The precursor mainly contains a peptide of 28 amino acids (ANP peptide). ANP is the cognate ligand of at least three single-pass type I membrane receptors with guanylate cyclase activity (ANPRA, ANPRB, ANPRC). Binds ANPRA with a higher affinity than other natriuretic peptides.	Released from heart atria in response to hypovolemia, acts on outer adrenal cells to decrease aldosterone production; provokes smooth muscle relaxation. Defends against mineralocorticoid-induced and salt-induced hypertension. It is thought to play a key role in cardiovascular homeostasis. Has antimitogenic properties.
MSSFSTTTVSFLLLLAFQLLGQTRA*NPMYNAVSNADLMDFKNLLDHLEEKMPLED*$_1$EVVPPQVLSEPNEEAGAALSPLPEVPPWTGEVSPAQRDGGALGRGPWDSSDRSALLKSKLRALLTAPR*SLRRSSCFGGRMDRIGAQSGLGCNSFRY*$_2$RR 1. cardiodilatin-related peptide. 2. atrial natriuretic peptide.		
Brain natriuretic peptide precursor (BNP)	The precursor contains a peptide of 32 amino acids (BNP peptide). BNP is the cognate ligand of at least three single-pass type I membrane receptors with guanylate cyclase activity (ANPRA, ANPRB, ANPRC).	Acts as a cardiac hormone with a variety of biological actions including natriuresis, diuresis, vasorelaxation, and inhibition of renin and aldosterone secretion. Helps restore the body's salt and water balance. Improves heart function. Has antimitogenic properties.
MDPQTAPSRALLLLLFLHLAFLGGRSHPLGSPGSASDLETSGLQEQRNHLQGKLSELQVEQTSLEPLQESPRPTGVWKSREVATEGIRGHRKMVLYTLRAPR*SPKMVQGSGCFGRKMDRISSSSGLGCKVLRRH*$_1$ 1. brain natriuretic peptide.		

Neuropeptides

HORMONE	STRUCTURE AND RECEPTORS	FUNCTION
Galanin precursor (GALA)	The precursor contains 2 peptides of 30 and 59 amino acids. The galanin peptide binds to 3 GPCRs (GALR1, GALR2 and GALR3).	Contracts smooth muscle of the gastrointestinal and genito-urinary tract, regulates growth hormone release, modulates insulin release, and may be involved in the control of adrenal secretion.
MARGSALLLASLLLAAALSASAGLWSPAKEKR*GWTLNSAGYLLGPHAVGNHRSFSDKNGLTS*$_1$KR*ELRPEDDMKPGSFDIERSDRSIPENNIMRTIIEFLSFLHLKEAGALDRLLDLPAAASSE*$_2$ 1. galanin. 2. galanin message-associated peptide.		
Neuropeptide B precursor (NPB)	The precursor contains 2 peptides of 24 and 29 amino acids, corresponding to alternative processing of precursor. 2 GPCRs (GPR7, GPR8).	May be involved in the regulation of feeding, the neuroendocrine system, memory, learning and in the afferent pain pathway.
MARSATLAAAALALCLLLAPPGLA*WYKPAAGHSSYSVGRAAGLLSGLR*$_2$*RSPYA*$_1$RR SQPYRGAEPPGGAGASPELQLHPRLRSLAVCVQDVAPNLQRCERLPDGRGTYQCKANV FLSLRAADCLAA 1. neuropeptide B-29. 2. neuropeptide B-23.		
Neuropeptide W precursor (NPW)	The precursor contains 2 peptides of 23 and 30 amino acids, corresponding to alternative processing of precursor. Same family as NPB. 2 known GPCRs (GPR7, GPR8).	Plays a regulatory role in the organization of neuroendocrine signals accessing the anterior pituitary gland. Stimulates water drinking and food intake. May play a role in the hypothalamic response to stress. NPW23 activates GPR7 and GPR8 more efficiently than NPW30.

<table>
<tr><td colspan="3">MAWRPGERGAPASRPRLALLLLLLLLPLPSGAWYKHVASPRYHTVGRAAGLLMGL₂
RRSPYLW₁RRALRAAAGPLARDTLSPEPAAREAPLLLPSWVQELWETRRRSSQAGIPV
RAPRSPRAPEPALEPESLDFSGAGQRLRRDVSRPAVDPAANRLGLPCLAPGPF
1. neuropeptide W-30.
2. neuropeptide W-23.</td></tr>
<tr><td>Melanin-concentrating hormone precursor (MCH)</td><td>this precursor contains three non-homologous neuropeptides. The best studied one, MCH is 19 aa long, and act through GPCRs (MCHR1).</td><td>MCH may act as a neurotransmitter or neuromodulator in a broad array of neuronal functions directed toward the regulation of goal-directed behavior, such as food intake, and general arousal. May also have a role in spermatocyte differentiation.</td></tr>
<tr><td colspan="3">MAKMNLSSYILILTFSLFSQGILLSASKSIRNLDDDMVFNTFRLGKGFQKEDTAEKSVIA
PSLEQYKNDESSFMNEEENKVSKNTGSKHNFLNHGLPLNLAIKPY-
LALKGSVDFPAENGVQNTESTQE₁KREIGDEENSAKFPI₂GRR
DFDMLRCMLGRVYRPCWQV₃
1. neuropeptide-glycine-glutamic acid (NGE).
2. neuropeptide-glutamic acid-isoleucine (NEI).
3. melanin-concentrating hormone (MCH peptide).</td></tr>
<tr><td>Enkephalin precursor (PENK)</td><td>The precursor sequence contains repeated homologous peptide sequences of 5, 7 or 8 amino acids flanked by dibasic residues.</td><td>Met-and Leu-enkephalins compete with and mimic the effects of opiate drugs. They play a role in a number of physiologic functions, including pain perception and responses to stress.</td></tr>
<tr><td colspan="3">MARFLTLCTWLLLLGPGLLATVRAECSQDCATCSYRLVRPADINFLACVMECEGKLPSL
KIWETCKELLQLSKPELPQDGTSTLRENSKPEESHLLAKRYGGFM₁KRYGGFM₁KKM
DELYPMEPEEEANGSEILAKRYGGFM₁KKDAEEDDSLANSSDLLKELLETGDNRERSH
HQDGSDNEEEVSKRYGGFMRGL₂KRSPQLEDEAKELQKRYGGFM₁RRVGRPEWWM
DYQKRYGGFL₃KRFAEALPSDEEGESYSKEVPEMEKRYGGFMRF₄
1. Met-enkephalin.
2. Met-enkephalin-Arg-Gly-Leu.
3. Leu-enkephalin.
4. Met-enkephalin-Arg-Phe.</td></tr>
</table>

Beta-neoendorphin-dynorphin precursor (PDYN)	this precursor contains at least 6 neuropeptides, homologous to the enkephalin peptides. They act through the kappa-type opioid GPCR (OPRK)	They are peptides produced by the pituitary gland and the hypothalamus in vertebrates, and they resemble the opiates in their abilities to produce analgesia and a sense of well-being.

MAWQGLVLAACLLMFPSTTADCLSRCSLCAVKTQDGPKPINPLICSLQCQAALLPSEEW
ERCQSFLSFFTPSTLGLNDKEDLGSKSVGEGPYSELAKLSGSFLKELEKSKFLPSISTKEN
TLSKSLEEKLRGLSDGFREGAESELMRDAQLNDGAMETGTLYLAEEDPKEQVKR
YGGFL$_3$*RKYP*$_2$*K* RSSEVAGEGDGDSMGHEDLYKR*YGGFL*$_3$*RRIRPKLKWDNQ*$_4$KR
YGGFL$_3$*RRQFKVVT*$_6$*RSQEDPNAYSGELFDA*$_5$

1. alpha-neoendorphin.
2. beta-neoendorphin.
3. Leu-enkephalin.
4. dynorphin a.
5. leumorphin.
6. rimorphin.

Nociceptin precursor (PNOC)	this precursor contains the Nociceptin neuropeptide, and probably two more neuropeptides. Nociceptin acts through the GPCR (OPRK, which has some homology with the opioid receptors (cf. Enkephalins)).	They are peptides produced by the pituitary gland and the hypothalamus in vertebrates, and they resemble the opiates in their abilities to produce analgesia and a sense of well-being.

MKVLLCDLLLLSLFSSVFSSCQRDCLTCQEKLHPALDSFDLEVCILECEEKVFPSPLWTP
CTKVMARSSWQLSPAAPEHVAAALYQPRASEMQHLRR*MPRVRSLFQEQEEPEPGM*
EEAGEMEQKQLQ$_1$KR*FGGFTGARKSARKLANQ*$_2$KR*FSEFMRQYLVLSMQSSQ*$_3$
RRRTLHQNGNV

1. probable neuropeptide 1 (unknown function).
2. nociceptin.
3. probable neuropeptide 2 (unknown function).

FMRFamide-related peptides precursor (NPFF)	This precursor contains 3 peptides of 8, 11, and 18 aa. 2 known GPCRs associated NPFF1(formerly GPR147) and NPFF2.	Morphine modulating peptides. Have wide-ranging physiologic effects, including the modulation of morphine-induced analgesia, elevation of arterial blood pressure, and increased somatostatin secretion from the pancreas. Neuropeptide FF potentiates and sensitizes ACCN2 and ACCN3 channels.
MDSRQAAALLVLLLLIDGGCAEGPGGQQEDQLSAEEDSEPLPPQDAQTSGSLLHYLLQAMERPGR*SQAFLFQPQRF*$_{2_1}$ GRNTQGSWRNEWLSPR*AGEGLNSQFWSLAAPQRF*$_3$GKK 1. neuropeptide SF. 2. neuropeptide FF. 3. neuropeptide AF.		
FMRFamide-related peptides precursor (RFRP)	This precursor contains 2 peptides of 8 and 37 aa.	Neuropeptides NPSF and NPVF efficiently inhibit forskolin-induced production of cAMP. Neuropeptide NPSF induces secretion of prolactin in rats. Neuropeptide NPVF blocks morphine-induced analgesia.
MEIISSKLFILLTLATSSLLTSNIFCADELVMSNLHSKENYDKYSEPRGYPKGER*SLNFEELKDWGPKNVIKMSTPAVNKMPHSFANLPLRF*$_1$GRNVQEER*SAGATANLPLRS*$_2$GRNMEVSLVRR*VPNLPQRF*$_3$GRTTTAKSVCRMLSDLCQGSMHSPCANDLFYSMTCQHQEIQNPDQKQSRRLLFKKIDDAELKQEK 1. neuropeptide NPSF. 2. potential neuropeptide RFRP-2 (unknown function). 3. neuropeptide NPVF.		
Orexin or Hypocretin precursor(OREX)	This precursor contains 2 homologous peptides of 28 and 33 aa. 2 known GPCRs associated OX1R and OX2R	Neuropeptides that play a significant role in the regulation of food intake and sleep and arousal.
MNLPSTKVSWAAVTLLLLLLLLPPALLSSGAAA*QPLPDCCRQKTCSCRLYELLHGAGNHAAGILTL*$_1$GKR*RSGPPGLQGRLQRLLQASGNHAAGILTM*$_2$GRRAGAEPAPRPCLGRRCSAPAAASVAPGGQSGI 1. orexin-A. 2. orexin-B.		

Orexigenic neuropeptide QRFP precursor(OX26)	This precursor contains 1 known neuropeptide QRF-amide, of 26 aa. Activates GPCR QRFPR (GPR103).	Has orexigenic activity in mice. Promotes aldosterone secretion by the adrenal gland when administered to rats.
MVRPYPLIYFLFLPLGACFPLLDRREPTDAMGGLGAGERWADLAMGPRPHSVWGSSR WLRASQPQALLVIARGLQTSGREHAGCRFRFGRQDEGSEATGFLPAAGEK*TSGPLGNL AEELNGYSRKKGGFSFRF*$_1$GRR 1. QRF-amide.		
Cocaine and amphetamine-related neuropeptide precursor (short isoform) (CART)	This precursor contains 2 known bioactive neuropeptides, of 48 aa (CART I) and 41 aa (CART II). No CART receptor has been identified to date.	Satiety factor closely associated with the actions of leptin and neuropeptide Y; these anorectic peptides inhibit both normal and starvation-induced feeding and completely block the feeding response induced by neuropeptide Y and regulated by leptin in the hypothalamus. They promote neuronal development and survival *in vitro*.
MESSRVRLLPLLGAALLLMLPLLGTRAQEDAELQPRALDIYSAVDDASHEKELIEALQEV LKKLKSKR*VPIYEKKYGQVPMCDAGEQCAVRKGARIGKLCDCPRGTSCNSFLL KCL*$_2$$_1$ 1. CART I. 2. CART II.		
Urocortin precursor (UCN1)	Encoded by a gene paralogous to the corticoliberin (CRF) gene. Contains a peptide of 41 amino acids. GPCR CRFR2.	Acts in vitro to stimulate the secretion of adrenocorticotropic hormone (ACTH).
MRQAGRAALLAALLLLVQLCPGSSQRSPEAAGVQDPSLRWSPGARNQGGGARALLLLL AERFPRRAGPGRLGLGTAGERPRR*DNPSLSIDLTFHLLRTLLELARTQSQRERAEQ NRIIFDSV*$_1$GK 1. urocortin.		

Urocortin II precursor (UCN2)	Encoded by a gene paralogous to the corticoliberin (CRF) and UCN1 genes. Contains a peptide of 41 amino acids. GPCR CRFR2.	Suppresses food intake, delays gastric emptying and decreases heat-induced edema. Might represent an endogenous ligand for maintaining homeostasis after stress.
MTRCALLLLMVLMLGRVLVVPVTPIPTFQLRPQNSPQTTPRPAASESPSAAPTWPWAA QSHCSPTRHPGSR*IVLSLDVPIGLLQILLEQARARAAREQATTNARILARVGHC*$_1$ 1. urocortin-2.		
Urocortin III precursor (UCN3)	Encoded by a gene paralogous to the corticoliberin (CRF) and UCN1 genes. Contains a peptide of 38 amino acids. GPCR CRFR2.	Suppresses food intake, delays gastric emptying and decreases heat-induced edema. Might represent an endogenous ligand for maintaining homeostasis after stress.
MLMPVHFLLLLLLLLGGPRTGLPHKFYKAKPIFSCLNTALSEAEKGQWEDASLLSKRSF HYLRSRDASSGEEEEGKEKKTFPISGARGGAGGTRYRYVSQAQPRGKPRQDTAKSPHR TK*FTLSLDVPTNIMNLLFNIAKAKNLRAQAAANAHLMAQI*$_1$GRKK 1. urocortin-3.		
C-type natriuretic peptide precursor (CNP)	The CNP precursor, a paralog of ANP and BNP precursors, contains three peptides that share the same C-terminal part, of 53, 29 and 22 aa amino acids. CNP-22 precursor is the natural agonist of receptor ANPRB and also binds ANPRC.	CNP peptides have vasorelaxant and cGMP-stimulating activities and antimitogenic properties. CNP is predominantly expressed in the CNS, kidney and pituitary.

MHLSQLLACALLLTLLSLRPSEAKPGAPPKVPRTPPAEELAEPQAAGGGQKKGDKAPG
GGGANLKGDRSRLLR*DLRVDTKSRAAWARLLQEHPNARKYKGANKKGLSKGCF*
GLKLDRIGSMSGLGC$_{3_{2_1}}$

1. CNP-53.
2. CNP-29.
3. CNP-22.

Calcitonin gene-related peptide I precursor (CGRP1, CALCA)	Contains a protein of 37 amino acids, product of the calcitonin gene derived by alternative splicing of the precursor mRNA in the brain. Its receptor is thought to be GPCR CALRL.	CGRP1 induces vasodilatation. It dilates a variety of vessels including the coronary, cerebral and systemic vasculature. Its abundance in the CNS also points toward a neurotransmitter or neuromodulator role. It also elevates platelet cAMP.

MGFQKFSPFLALSILVLLQAGSLHAAPFRSALESSPADPATLSEDEARLLLAALVQDYVQM
KASELEQEQEREGSRIIAQKR*ACDTATCVTHRLAGLLSRSGGVVKNNFVPTNVGS*
KAF$_1$GRRRRDLQA

1. calcitonin gene-related peptide 1.

Calcitonin gene-related peptide II precursor (CGRP2, CALCB)	Encoded by a gene paralogous to the calcitonin/CGRP1 gene. Contains a protein of 37 amino acids highly homologous to the CGRP1 peptide (almost identical). Its receptor is thought to be GPCR CALRL.	CGRP2 induces vasodilatation. It dilates a variety of vessels including the coronary, cerebral and systemic vasculature. Its abundance in the CNS also points toward a neurotransmitter or neuromodulator role. It also elevates platelet cAMP.

MGFRKFSPFLALSILVLYQAGSLQAAPFRSALESSPDPATLSKEDARLLLAALVQDYVQM
KASELKQEQETQGSSSAAQKR*ACNTATCVTHRLAGLLSRSGGMVKSNFVPTNVGS*
KAF$_1$GRRRRDLQA

1. calcitonin gene-related peptide 2.

Tuberoinfundibular peptide of 39 residues precursor (PTH2, TIP39)	Same family as parathyroid hormone. Contains a peptide of 39 amino acids. Its receptor is a GPCR of the parathyroid hormone receptor family PTH2R.	TIP39 may inhibit cell proliferation via its action on PTH2R activation. TIP39 is a neuropeptide which may activate nociceptive circuits and may also have a role in spermatogenesis.
METRQVSRSPRVRLLLLLLLLLVVPWGVRTASGVALPPVGVLSLRPPGRAWADPATPRP RR*SLALADDAAFRERARLLAALERRHWLNSYMHKLLVLDAP*$_1$ 1. tuberoinfundibular peptide of 39 residues.		
Tachykinin-3 precursor (TKNK)	Contains neurokinin-B, a peptide of 10 amino acids, which is the endogenous agonist ligand of GPCR NK3R.	Hypothalamic arcuate nucleus NKB-expressing neurons may be responsive to sex steroid signals such as estrogen.
MRIMLLFTAILAFSLAQSFGAVCKEPQEEVVPGGGRSKRDPDLYQLLQRLFKSHSSLEGL LKALSQASTDPKESTSPEKR*DMHDFFVGLM*$_1$GKRSVQPDSPTDVNQENVPSFGILKY PPRAE 1. neurokinin B (NKB).		
Apelin precursor (APEL)	Processing of the precursor releases several peptides sharing the same c-terminus. Apelin is the endogenous ligand for GPCR APJ.	APJ is an alternative coreceptor with CD4 for HIV-1 infection. May have a role in the modulation of the immune responses in neonates. May also have a role in the central control of body fluid homeostasis by influencing AVP release and drinking behavior.
MNLRLCVQALLLLWLSLTAVCGGSLMPLPDGNGLEDGNVRH*LVQPRGSRNGPGPWQ GGRRKFRRQRPRLSHKGPMPF*$_4$$_3$$_2$$_1$ 1. apelin-36. 2. apelin-31. 3. apelin-28. 4. apelin-13.		

Secretogranin-I precursor (SCG1)	Processing of the precursor yields at least two bioactive peptides, GAWK (74 aa) and CCB (60 aa). No receptor to these peptides have been identified to date.	The precise function of GAWK and CCB peptides is not understood at present. Secretogranin-1 is a neuroendocrine secretory granule protein, which may be the precursor for several other biologically active peptides.
MQPTLLLSLLGAVGLAAVNSMPVDNRNHNEGMVTRCIIEVLSNALSKSSAPPITPECRQV LKTSRKDVKDKETTENENTKFEVRLLRDPADASEAHESSSRGEAGAPGEEDIQGPTKAD TEKWAEGGGHSRERADEPQWSLYPSDSQVSEEVKTRHSEKSQREDEEEEEGENYQKGE RGEDSSEEKHLEEPGETQNAFLNERKQASAIKKEELVARSETHAAGHSQEKTHSREKSS QESGEEAGSQENHPQESKGQPRSQEESEEGEEDATSEVDKRRTRPRHHHGRSRPDRSSQ GGSLPSEEKGHPQEESEESNVSMASLGEKRDHHSTHYRASEEEPEYGEEIKGYPGVQAP EDLEWERYRGRGSEEYRAPRPQSEESWDEEDKRNYPSLELDKMAHGYGEESEEERGLE PGKGRHHRGRGGEPRAYFMSDTREEKR*FLGEGHHRVQENQMDKARRHPQGAWK ELDRNYLNYGEEGAPGKWQQQGDLQDTKENREEARFQDKQYSSHHTAE*$_1$KRK RLGELFNPYYDPLQWKSSHFERRDNMNDNFLEGEEENELTLNEKNFFPEYNYDWWEK KPFSEDVNWGYEKRNLARVPKLDLKRQYDRVAQLDQLLHYRKK*SAEFPDFYDSEEP VSTHQEAENEKDRADQTVLTEDEKKELENLAAMDLELQKIAEKFSQR*$_2$G 1. GAWK peptide. 1. CCB peptide.		
Secretogranin-II precursor (SCG2)	Processing of the precursor yields secretoneurin, a bioactive peptide of 33 amino acids. No receptor to secretoneurin has been identified to date.	Secretoneurin is a potent chemoattractant for human eosinophils. It has been shown to inhibit the proliferation and stimulation of migration of endothelial cells.

MPRSCCSRSGALLLALLLQASMEVRGMAEAKTHWLGAALSLIPLIFLISGAEAASFQRNQ
LLQKEPDLRLENVQKFPSPEMIRALEYIENLRQQAHKEESSPDYNPYQGVSVPLQQKEN
GDESHLPERDSLSEEDWMRIILEALRQAENEPQSAPKENKPYALNSEKNFPMDMSDDYE
TQQWPERKLKHMQFPPMYEENSRDNPFKR*TNEIVEEQYTPQSLATLESVFQELGK*
LTGPNNQ$_1$KRERMDEEQKLYTDDEDDIYKANNIAYEDVVGGEDWNPVEEKIESQTQE
EVRDSKENIEKNEQINDEMKRSGQLGIQEEDLRKESKDQLSDDVSKVIAYLKRLVNAAGS
GRLQNGQNGERATRLFEKPLDSQSIYQLIEISRNLQIPPEDLIEMLKTGEKPNGSVEPERE
LDLPVDLDDISEADLDHPDLFQNRMLSKSGYPKTPGRAGTEALPDGLSVEDILNLLGME
SAANQKTSYFPNPYNQEKVLPRLPYGAGRSRSNQLPKAAWIPHVENRQMAYENLNDKD
QELGEYLARMLVKYPEIINSNQVKRVPGQGSSEDDLQEEEQIEQAIKEHLNQGSSQETDK
LAPVSKRFPVGPPKNDDTPNRQYWDEDLLMKVLEYLNQEKAEKGREHIAKRAMENM

1. secretoneurin.

Neurosecretory protein VGF precursor (VGF)	Several potential endogenous peptides have been identified, including TLQP-21, NERP-1 and NERP-2 (neuroendocrine regulatory peptides). No receptors to these peptides have yet been identified.	TLQP-21 is involved in energy homeostasis and neuronal survival. NERPs may modulate body fluid homeostasis. VGF peptides may be involved in the regulation of cell-cell interactions or in synaptogenesis during the maturation of the nervous system.

MKALRLSASALFCLLLINGLGAAPPGRPEAQPPPLSSEHKEPVAGDAVPGPKDGSAPEV
RGARNSEPQDEGELFQGVDPRALAAVLLQALDRPASPPAPSGSQQGPEEEAAEALLTET
VRSQTHSLPAAGEPEPAAPPRPQTPENGPEASDPSEELEALASLLQELRDFSPSSAKRQQ
ETAAAETETRTHTLTRVNLESPGPERVWRASWGEFQARVPERAPLPPPAPSQFQARMP
DSGPLPETHKFGEGVSSPKTHLGEALAPLSKAYQGVAAPFPKARR*AESALLGGSEAGE*
RLLQQGLAQVEA$_1$GRR*QAEATRQAAAQEERLADLASDLLLQYLLQGGARQRGLG*$_2$
GRGLQEAAEERESAREEEEAEQERRGGEERVGEEDEEAAEAAEAEADEAERARQNALL
FAEEEDGEAGAEDKRSQEETPGHRRKEAEGTEEGGEEEDDEEMDPQTIDSLIELSTKLH
LPADDVVSIIEEVEEKRNRKKKAPPEPVPPPRAAPAPTHVRSPQPPPPPPSARDELPDW
NEVLPPWDREEDEVYPPGPYHPFPNYIRPR*TLQPPSALRRRHYHHALPPSR*$_3$HYPG
REAQARHAQQEEAEAEERRLQEQEELENYIEHVLLRRP

1. NERP-1.
2. NERP-2.
3. TLQP-21.

Neuropeptide S precursor (NPS)	Processing of the precursor releases a peptide of 20 amino acids. NPS is the endogenous ligand for GPCR NPSR1.	Modulates arousal and anxiety. May play an important anorexigenic role.
MISSVKLNLILVLSLSTMHVFWCYPVPSSKVSGKSDYFLILLNSCPTRLDRSKELAFLKPILEKMFVKR*SFRNGVGTGMKKTSFQRAKS*$_1$ 1. NPS.		
ODN precursor (ACBP)	Atypical precursor (without signal peptide) contains a neuropeptide of 18 amino acids (octaneuropeptide). Has been shown to bind to a GPCR.	ODN is able to displace diazepam from the benzodiazepine (BZD) recognition site located on the GABA type A receptor. It may act as a neuropeptide to modulate the action of the GABA receptor. ICV injection of ODN stimulates feeding and corresponding receptor has been shown to be involved in anxiety.
MSQAEFEKAAEEVRHLKTKPSDEEMLFIYGHYK*QATVGDINTERPGMLDFT*$_1$GKAKWDAWNELKGTSKEDAMKAYINKVEELKKKYGI 1. octadecaneuropeptide (ODN).		

Peptide hormones involved in cardiovascular homeostasis

HORMONE	STRUCTURE AND RECEPTORS	FUNCTION
Adrenomedullin precursor (ADML)	Processing of the precursor yields two amidated peptides, proadrenomedullin N-20 terminal peptide (PAMP,20 aa) and adrenomedullin (AM, 52 amino acids). Shares GPCR CALRL with calcitonin gene-related peptide and interaction with receptors RAMP2 and RAMP3 confer specificity.	AM and PAMP are potent hypotensive and vasodilatator agents. Numerous actions have been reported most related to the physiologic control of fluid and electrolyte homeostasis. In the kidney, AM is diuretic and natriuretic, both AM and PAMP inhibit aldosterone secretion by direct adrenal actions. In pituitary gland, both peptides at physiologically relevant doses inhibit basal ACTH secretion. Both peptides appear to act in brain and pituitary gland to facilitate the loss of plasma volume, actions which complement their hypotensive effects in blood vessels.
MKLVSVALMYLGSLAFLGADT*ARLDVASEFRKKWNKWALSR*$_1$GKRELRMSSSYPTG LADVKAGPAQTLIRPQDMKGASRSPEDSSPDAARIRVKR*YRQSMNNFQGLRSFGCRF GTCTVQKLAHQIYQFTDKDKDNVAPRSKISPQGY*$_2$GRRRRRSLPEAGPGRTLVSSK PQAHGAPAPPSGSAPHFL 1. proadrenomedullin N-20 terminal peptide (PAMP). 2. adrenomedullin (AM).		
Adrenomedullin 2 precursor (ADM2)	Processing of the precursor yields at least one bioactive peptide, adrenomedullin-2, of 47 amino acids. Shares GPCR CALRL with calcitonin gene-related peptide and interaction with receptors RAMP2 and RAMP3 confer specificity.	intermedin-2 may play a role as a physiological regulator of gastrointestinal, cardiovascular bioactivities mediated by the CALCRL/RAMPs receptor complexes. Activates the cAMP-dependent pathway.

<table>
<tr><td colspan="3">MARIPTAALGCISLLCLQLPGSLSRSLGGDPRPVKPREPPARSPSSSLQPRHPAPRPVVW
KLHRALQAQRGAGLAPVMGQPLRDGGRQHSGPRRHSGPRR<u>TQAQLLRVGCVLGTCQ
VQNLSHRLWQLMGPAGRQDSAPVDPSSPHSY</u>$_1$G
1. intermedin-2.</td></tr>
<tr><td>Urotensin-2 precursor (UTS2)</td><td>Processing of the precursor yields one bioactive peptide, urotensin-2, of 11 amino acids, that is associated with GPCR UR2R.</td><td>Urotensin-2 is a highly potent vasoconstrictor.</td></tr>
<tr><td colspan="3">MYKLASCCLLFIGFLNPLLSLPLLDSRREISFQLSAPHEDARLTPEELERASLLQILPEMLGA
ERGDILRKADSSTNIFNPRGNLRKFQDFSGQDPNILLSHLLARIWKPYKKR<u>ETPDCFW
KYCV</u>$_1$
1. urotensin-2.</td></tr>
<tr><td>Urotensin-2B precursor (UTS2)</td><td>Processing of the precursor yields one bioactive peptide, urotensin-2B, of 8 amino acids, that is associated with GPCR UR2R.</td><td>Urotensin-2B is a highly potent vasoconstrictor (related to UTS-2).</td></tr>
<tr><td colspan="3">MNKILSSTVCFGLLTLLSVLIFLQSVHGRPYLTQGNEIFPDKKYTNREELLLALLNKNFD
FQRPFNTDLALPNKLEELNQLEKLKEQLVEEKDSETSYAVDGLFSSHPSKR<u>ACFWKYC
V</u>$_1$
1. urotensin-2B.</td></tr>
<tr><td>Prosalusin precursor (TOR2X)</td><td>Processing of the precursor derived from an isoform of the torsin-2A gene yields two bioactive peptides, salusin-α (28 aa) and salusin-β (20 aa). To date no salusin-α/β receptor has been found (mouse receptor MRGA1 is a surrogate receptor for human salusin-β).</td><td>Salusins -alpha and -beta are endocrine and/or paracrine factors able to increase intracellular calcium concentrations and induce cell mitogenesis. Salusins are potent hypotensive peptides.</td></tr>
</table>

MAAATRGCRPWGSLLGLLGLVSAAAAAWDLASLRCTLGAFCECDFRPDLPGLECDLAQ
HLAGQHLAKALVVKALKAFVRDPAPTKPLVLSLHGWTGTGKSYVSSLLAHYLFQGGLR
SPRVHHFSPVLHFPHPSHIERYKKDLKSWVQGNLTACGRSLFLFDEMDKMPPGLMEVLR
PFLGSSWVVYGTNYRK*AIFIFIRWLLKLGHHGRAPP*$_1$RR*SGALPPAPAAPRPALRA*
QRAGPAGPGAK$_2$*G*

1. salusin-β.
2. salusin-α.

Peptide hormones involved in bone and cartilage homeostasis

HORMONE	STRUCTURE AND RECEPTORS	FUNCTION
Parathyroid hormone-related protein precursor (PTHR, PTHR)	Contains three peptides of 36, 57 and 33 amino acids. Receptors are GPCRs of the parathyroid hormone receptor family PTHR1.	Neuroendocrine peptide which is a critical regulator of cellular and organ growth, development, migration, differentiation and survival and of epithelial calcium ion transport. Regulates endochondral bone development and epithelial-mesenchymal interactions during the formation of the mammary glands and teeth.

MQRRLVQQWSVAVFLLSYAVPSCGRSVEGLSRRLKR*AVSEHQLLHDKGKSIQDLRRR*
FFLHHLIAEIHTAEI$_1$R*ATSEVSPNSKPSPNTKNHPVRFGSDDEGRYLTQETNK*
VETYKEQPLKTPGKKKKGKP$_2$GKRKEQEKKKRR*TRSAWLDSGVTGSGLEGDHL*
SDTSTTSLELDSR$_3$RH

1. PTHrP[1-36].
2. PTHrP[38-94].
3. osteostatin.

Main sources : *SwissProt and Indiana University School of Medecine*

Identification of novel peptide hormones in the human proteome by hidden Markov model screening

Olivier Mirabeau, Emerald Perlas, Cinzia Severini, Enrica Audero, Olivier Gascuel, Roberta Possenti, Ewan Birney, Nadia Rosenthal and Cornelius Gross

Genome Res. 2007 17: 320-327; originally published online Feb 6, 2007;
Access the most recent version at doi:10.1101/gr.5755407

Methods

Identification of novel peptide hormones in the human proteome by hidden Markov model screening

Olivier Mirabeau,[1] Emerald Perlas,[1] Cinzia Severini,[2] Enrica Audero,[1] Olivier Gascuel,[3] Roberta Possenti,[2,4] Ewan Birney,[5] Nadia Rosenthal,[1] and Cornelius Gross[1,6]

[1]*Mouse Biology Unit, EMBL, 00016 Monterotondo, Italy;* [2]*INMM, 00143 Rome, Italy;* [3]*LIRMM-CNRS, 34392 Montpellier, France;* [4]*Department of Neuroscience, University Tor Vergata Rome, 00133 Rome, Italy;* [5]*European Bioinformatics Institute, EBI-EMBL, CB10 1SD Hinxton, United Kingdom*

Peptide hormones are small, processed, and secreted peptides that signal via membrane receptors and play critical roles in normal and pathological physiology. The search for novel peptide hormones has been hampered by their small size, low or restricted expression, and lack of sequence similarity. To overcome these difficulties, we developed a bioinformatics search tool based on the hidden Markov model formalism that uses several peptide hormone sequence features to estimate the likelihood that a protein contains a processed and secreted peptide of this class. Application of this tool to an alignment of mammalian proteomes ranked 90% of known peptide hormones among the top 300 proteins. An analysis of the top scoring hypothetical and poorly annotated human proteins identified two novel candidate peptide hormones. Biochemical analysis of the two candidates, which we called spexin and augurin, showed that both were localized to secretory granules in a transfected pancreatic cell line and were recovered from the cell supernatant. Spexin was expressed in the submucosal layer of the mouse esophagus and stomach, and a predicted peptide from the spexin precursor induced muscle contraction in a rat stomach explant assay. Augurin was specifically expressed in mouse endocrine tissues, including pituitary and adrenal gland, choroid plexus, and the atrio-ventricular node of the heart. Our findings demonstrate the utility of a bioinformatics approach to identify novel biologically active peptides. Peptide hormones and their receptors are important diagnostic and therapeutic targets, and our results suggest that spexin and augurin are novel peptide hormones likely to be involved in physiological homeostasis.

[Supplemental material is available online at www.genome.org and at http://bioinfo.embl.it/.]

The study of peptide hormones has received considerable attention because of their role in modulating a wide range of physiological functions (Kastin 2006). A large group of peptide hormones serve as both hormones and neurotransmitters, being secreted into the bloodstream by endocrine cells and released into the synapse by neurons (Hökfelt 1991). Because of this dual function, peptide hormones often play important roles in the coordination of behavioral and somatic responses to environmental stimuli, and understanding their biology has helped advance our understanding of interactions between brain and body.

Peptide hormones are short peptides (<100 amino acids) produced by the proteolytic cleavage of pre-pro-hormone precursors. Following signal peptide removal by the signal peptidase complex, the pro-hormone undergoes cleavage at specific sites by pro-hormone convertases (Steiner 1998) or furin (Thomas 2002). In many cases, processed peptides undergo post-translational modification, with >50% of peptide hormones becoming amidated at their C terminus (Eipper et al. 1992). Mature peptides pass through the secretory pathway and are released into the extracellular space, where they can bind to specific cell surface receptors and modulate cellular functions.

[6]Corresponding author.
E-mail gross@embl.it; fax 39-06-90091272.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5755407.

Most peptide hormones are ligands for G-protein-coupled receptors (GPCR), via which they modulate intracellular signaling pathways and regulate cellular homeostasis. GPCRs belong to the seven transmembrane receptor family and share a high degree of sequence homology. As a result, in many organisms the complete set of GPCRs has been identified and classified (Vassilatis et al. 2003). The fraction of human GPCRs with known peptide ligands has been used to estimate that 27 orphan GPCRs are expected to have endogenous peptide ligands (Vassilatis et al. 2003). Although some of these missing ligands may turn out to be either previously characterized peptide hormones or novel peptides produced by known genes (Shichiri et al. 2003), including known peptide hormone genes (Zhang et al. 2005), some of these are likely to be produced by as yet uncharacterized genes.

Several methods have been used to identify new peptide hormones. Biochemical purification coupled with functional assays has been the predominant discovery method (for example, see Braun-Menendez et al. 1939; Burgus et al. 1969; Schmidt et al. 1991; Katafuchi et al. 2003). More recently, with the advent of genomic sequence, bioinformatics search strategies have been developed. Bioinformatics search strategies have the advantage over biochemical approaches that they are not biased against proteins with low or highly restricted expression and can be equally well applied to organisms in which biochemical purification of sufficient peptides is prohibitive. Most of these bioinformatics strategies relied on searches for single sequence features

17:320–327 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

common to peptide hormones, such as C-terminal RF-amide (Hinuma et al. 2000; Chartrel et al. 2003; Jiang et al. 2003), dibasic cleavage sites (Duckert et al. 2004; Zhang et al. 2005), homology with known peptide hormones (Hsu 1999; Park et al. 2002), and other shared motifs (Baggerman et al. 2005). In at least one case, a combination of features was used to identify candidate peptide hormones (Shichiri et al. 2003). In this study, several independent sequence requirements including presence of signal peptide and pro-hormone cleavage sites, subcellular location, and precursor length were applied to retrieve a novel peptide hormone precursor from human cDNA databases. These studies demonstrated the success of sequence-based approaches to peptide hormone discovery and inspired us to develop a more systematic bioinformatics approach to address this problem.

We present here the development of a hidden Markov model (HMM) based search algorithm that integrates several peptide hormone sequence features for the discovery of novel peptide hormones. HMM techniques are well adapted to address sequence analysis problems because of their ability to handle variable sequence length signals and to implicitly integrate information from multiple dispersed signals in a sequence. As a result, HMMs have been applied successfully to both gene prediction (Burge and Karlin 1997; Birney et al. 2004) and protein domain finding, leading to domain databases such as Pfam (Krogh et al. 1994; Finn et al. 2006). In this study, we provide an HMM for all peptide hormones, more akin to gene prediction models that integrate biological processing signals than protein domain models that integrate homology signals. Application of our peptide hormone HMM to the human proteome allowed us to identify two novel candidate peptide hormones. Biochemical characterization of the candidates demonstrates that they are processed and secreted as predicted, and one of them has biological activity in a stomach contractility assay. These results demonstrate the power of a bioinformatics approach to find novel biologically active peptides.

Results

Development of peptide hormone search algorithm

Peptide hormones contain several common sequence features that distinguish them from other proteins. Peptide hormones all carry a signal peptide sequence and cleavage site at their N terminus, at least one pro-hormone cleavage site (generally occurring at a pair of basic residues), and amino acid residues that are typical for extracellular proteins and frequently include aromatic amino acids. Finally, peptide hormones do not contain transmembrane domains, and their processed products are short (<100 amino acids). We reasoned that these features could be used to identify novel peptide hormone genes.

Our strategy was to build a hidden Markov model (HMM) that would score proteins according to the likelihood that they encode a peptide hormone. An HMM assigns states to each amino acid in a protein sequence. Each state is associated with a probability distribution over amino acids and a set of transition probabilities to other states. Generally, these states correspond to protein sequence motifs, and as a result HMMs can be used to determine whether a protein contains a specific motif or series of motifs. The two main advantages of HMMs are the ability to handle variable length regions and the ability to integrate multiple signals in a biologically constrained manner.

In our case, two steps were involved in using HMM for protein analysis. First, the HMM was trained on a set of proteins with well-characterized motifs in order to determine the amino acid frequencies and transition probabilities for each state. Second, the HMM was used to assign states to uncharacterized proteins and calculate a score based on how well the protein fits the HMM. The state architecture of our peptide hormone HMM is shown in Figure 1A. The HMM was assembled from three com-

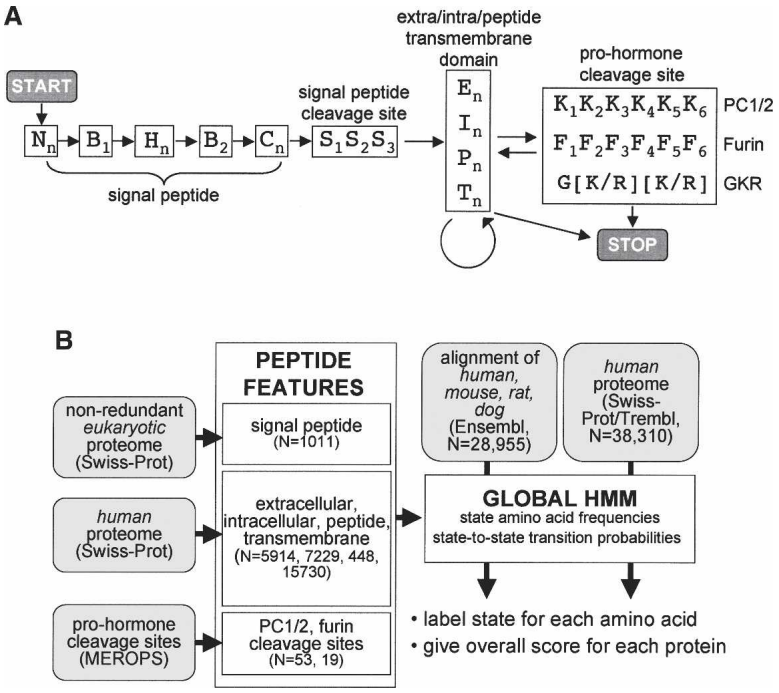

Figure 1. Hidden Markov model (HMM) for the identification of peptide hormones. (*A*) State structure of the peptide hormone HMM with states indicated by letters and transitions between states indicated by arrows. States with numerical subscripts are single amino acid states, while states with the "n" subscript are multiple amino acid states whose length is determined by the transition probability between that state and other permitted states. N_n, B_1, H_n, B_2, C_n, and S_{1-3}, are N terminus, border, hydrophobic, C terminus, and cleavage site states, respectively, of the signal peptide feature. E_n, I_n, P_n, and T_n are extracellular, intracellular, peptide, and transmembrane states, respectively, while K_{1-6} and F_{1-6} are pro-hormone cleavage site states and G[K/R][K/R] is a simple sequence motif. START and STOP mark entry and exit points of the HMM. (*B*) Protocol for building and running the peptide hormone HMM. HMM states for individual sequence features were built by learning amino acid frequencies and transition probabilities from sets of proteins or motifs with known features (*N* = size of training set). Signal peptide states were built using a previously curated set of eukaryotic SWISS-PROT proteins; extra/intra/peptide/transmembrane states were built using selected sets of human SWISS-PROT proteins; and pro-hormone cleavage sites were built using a set of PC1/2 and furin sites from the MEROPS database. The peptide hormone HMM was assembled with the state-to-state transition constraints outlined in *A*. Finally, the HMM was used to assign states and scores to either a set of alignments of human, mouse, rat, and dog proteins from Ensembl or a set of human proteins from SWISS-PROT/TrEMBL.

ponents each of which contains one or more states: (1) signal peptide, (2) extracellular/intracellular/peptide/transmembrane region, and (3) pro-hormone cleavage site. Several constraints were imposed on transitions between states so that, for example, the states for the signal peptide cleavage site ($S_1S_2S_3$) had to follow the C-terminal signal peptide state (C_n). The scheme for building and running the peptide hormone HMM is shown in Figure 1B.

For the signal peptide feature, our state architecture was based on previous work (Nielsen and Krogh 1998; Zhang and Wood 2002) and comprised N-terminal, hydrophobic, and C-terminal states followed by a three-state cleavage site. To improve predictive accuracy, we added two intermediate boundary states (B_1 and B_2) (Fig. 1A). Frequencies and transition probabilities for each state were derived by training the HMM on a previously curated set of 1011 signal peptide containing proteins from SWISS-PROT downloaded from the Center for Biological Sequence Analysis (http://www.cbs.dtu.dk/ftp/signalp/euksig.red).

For the extracellular/intracellular/peptide/transmembrane features, frequencies and transition probabilities were built from sets of 5914, 7229, 448, and 15,730 sequences derived from human SWISS-PROT entries annotated as "extracellular," "cytoplasmic," "peptide," and "transmembrane," respectively. These features were modeled by a single state of variable length where the length distribution was encoded by the transition probability out of the state. Because the first-order HMM formalism produces length distributions that are geometric and that may not be best suited to model actual protein feature lengths, we used a modified HMM formalism that retains the efficiency of first-order HMM, while being able to model lengths more accurately (Ramesh and Wilpon 1992).

Finally, states for the pro-hormone cleavage site feature used three different cleavage site models. The first two included 6 states each (from −6 to −1 relative to the cleavage site) and were built from a training set of 53 pro-hormone convertase 1/2 (PC1/2) cleavage sites and 19 furin cleavage sites, respectively, derived from known and predicted eukaryotic cleavage sites collected in the MEROPS database (Rawlings et al. 2006). The third cleavage site model simply required the presence of the amino acid sequence G[K/R][K/R] and was added to ensure that this common cleavage site motif was not overlooked by the other two models.

Labeling by the peptide hormone HMM was achieved using the Viterbi algorithm, and scoring was performed by the forward–backward algorithm (Rabiner 1989). To ensure the exclusion of transmembrane domain–containing proteins, a maximal score was assigned to any protein containing a transmembrane state. As a result, our algorithm was not expected to detect peptides that are proteolytically released from transmembrane-containing pro-peptides, such as occurs for tumor necrosis factor (TNF) and CX3CL1.

Screening the human proteome for novel peptide hormones

The peptide hormone HMM was applied to an alignment of nonredundant known and hypothetical proteins derived from the Ensembl database (human, mouse, rat, dog; N_{total} = 28,955), and proteins were ranked according to HMM scores. For each member of the set, multiple alignments of the human, rat, mouse, and dog orthologs were built using the Clustal W program with default settings (Thompson et al. 1994). An examination of the highest ranked sequences revealed that 90% of known peptide hormone precursors (66/75 proteins) (see Supplemental Material) were found in the top 300 proteins (Fig. 2A). An alternative HMM in which the PC1/2 and furin cleavage site models included only four states (from −4 to −1 relative to the cleavage site) performed slightly less well, returning 85% of known peptide hormones among the top 300 proteins (data not shown). Application of the peptide hormone HMM to human proteins from the SWISS-PROT/TrEMBL database (N_{total} = 38,310) was somewhat less efficient at recovering known peptide hormones, suggesting that screening the pre-aligned human proteome resulted in the recovery of fewer false positives (Fig. 2A). These findings demonstrate that our HMM successfully identified most known peptide hormones and suggests that as yet uncharacterized peptide hormones are likely to be found among the top scoring proteins in our list. The HMM Java application and supporting material are available at http://bioinfo.embl.it/.

At least 61% of the top 300 proteins belonged to several families of well-characterized secreted proteins, including peptide hormones, growth factors, cytokines, defensins, and antimicrobial peptides (Fig. 2B). A further 19% of the proteins were well-characterized membrane, mitochondrial, cytoplasmic, nuclear, or other nonsecreted proteins. The remaining 20% were hypothetical or poorly annotated proteins. In addition, we found four proteins (KISS1, TIP39, QRFP, OSTN) that had been recently reported to encode peptide hormones (Usdin et al. 1999;

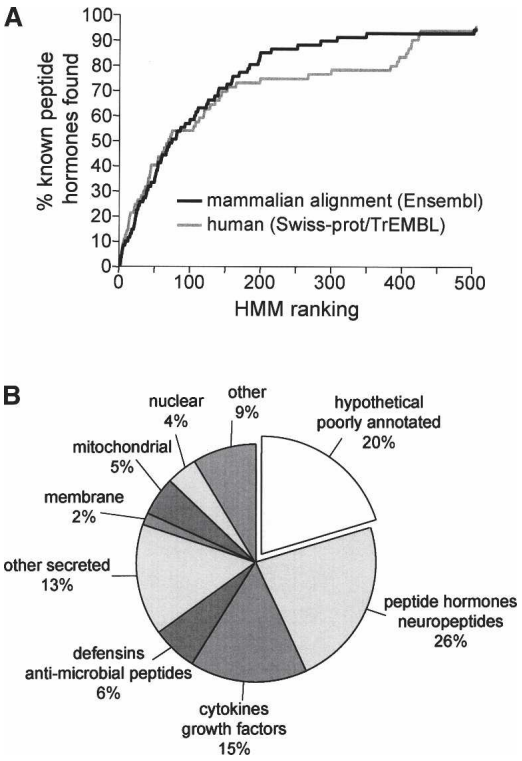

Figure 2. HMM successfully identified known peptide hormones. (*A*) Plot showing cumulative fraction of known peptide hormones (N_{total} = 77) (see Supplemental Table 1) identified among the top scoring 500 proteins; (solid line) aligned mammalian proteome as substrate; (gray line) human proteome as substrate. For the aligned proteome, 90% of known peptide hormones were found among the top 300 proteins. (*B*) Pie chart indicating percent composition of the top 300 proteins. Known peptide hormones, cytokines, growth factors, defensins, and other secreted proteins make up 61% of the proteins. Hypothetical and poorly annotated proteins make up 20% of the proteins and were submitted to further analysis to identify candidate novel peptide hormones.

Ohtaki et al. 2001; Chartrel et al. 2003; Thomas et al. 2003) despite lacking annotations as processed and secreted proteins in SWISS-PROT at the time of our search (March 2004).

Next, we applied three additional criteria to the 61 hypothetical and poorly annotated proteins to determine whether novel peptide hormones might be included among this group. First, proteins in which at least one of the amino acids at each putative pro-hormone cleavage site was not conserved among orthologs were removed from the list. Second, proteins in which labeled cleavage sites formed part of a longer stretch of basic residues were removed. These regions were likely to be nuclear localization signals or other basic amino acid domains rather than pro-hormone cleavage sites. Finally, we required a significant change in amino acid homology surrounding at least one putative cleavage site. In known peptide hormones, pro-hormone cleavage sites typically separate highly from poorly conserved regions. For this calculation, a significant change was defined as >30% change in average homology index (Livingstone and Barton 1993) for the five amino acids preceding and following the putative cleavage site. Two out of 61 proteins satisfied all three criteria. These candidate peptide hormones ranked 41 and 276 in our list and were called spexin (Ensembl: ENSP00000256969) and augurin (Ensembl: ENSP00000238044), respectively. Spexin carries a SWISS-PROT annotation as containing a putative amidated peptide (http://www.expasy.org/uniprot/Q9BT56), and augurin was previously identified as a gene expressed in esophageal cancer cell lines (Su et al. 1998). However, to the best of our knowledge, our study is the first to argue that augurin encodes a peptide hormone.

Characterization of candidate peptide hormones

The primary structure of human spexin and augurin precursors is shown in Figure 3. Both proteins contain obvious signal peptide sequences and cleavage sites as well as at least one putative dibasic residue pro-hormone cleavage site. Spexin contains a small, 15 amino acid region flanked by putative dibasic pro-hormone cleavage sites that is highly conserved in mammals, birds, and fish (for a full alignment of spexin and augurin orthologs, see Supplemental Fig. S1). The presence of a glycine residue at the end of this putative peptide suggests that it is processed and amidated, a common feature of peptide hormones (Eipper et al. 1992). Augurin, on the other hand, contains a single putative pro-hormone cleavage site followed by a single, long putative peptide that is highly conserved in mammals and fish. Both spexin and augurin peptides contain many aromatic amino acids, a feature typical of peptide hormones. Finally, there is a significant increase in sequence conservation that coincides with the N-terminal putative pro-hormone cleavage site in both spexin and augurin, further supporting a biological role for these features.

To study intracellular trafficking of spexin and augurin, the eight amino acid Flag antigen sequence (DYKDDDDK) was inserted just upstream and downstream of the putative spexin and augurin peptides (Fig. 3). As a control, Flag antigen was also inserted just upstream of neuropeptide K (NPK) in the human beta-preprotachykinin (*TAC1*) gene. Previous studies have shown that Flag sequences are compatible with proteolytic cleavage just N-terminal to the Flag sequence (Duguay et al. 1995). Immunocytochemistry with Flag antibodies following transfection of Flag-NPK, Flag-spexin, and Flag-augurin into a rat pancreatic cell line demonstrated colocalization of Flag antigen with endogenous insulin in punctate intracellular bodies (Fig. 4). This colocalization suggested that spexin and augurin, like neuropeptide K, underwent trafficking into dense core granules of the secretory pathway, a hallmark of peptide hormones.

To determine whether spexin and augurin were processed and secreted, cell supernatants were collected from rat pancreatic cells transfected with Flag-NPK, N-Flag-spexin, C-Flag-spexin, and Flag-augurin. Western blotting of supernatant from N-Flag-spexin transfected cells, revealed three Flag-immunoreactive bands (13, 12, and 6 kDa), consistent with secretion of processed spexin products (Fig. 5A,B). Western blotting of supernatant from Flag-NPK transfected cells revealed processing and secretion of neuropeptide K, consistent with previous studies (Fig. 5A; Conlon et al. 1988). To determine whether the 6-kDa band could represent completely processed spexin peptide, we transfected cells with a truncated spexin protein, called N-Flag-Δspexin, in which a stop codon had been engineered to replace the C-terminal putative pro-hormone cleavage site (Fig. 3A). Western blotting of supernatants from N-Flag-Δspexin transfected cells revealed a 4-kDa band (Fig. 5B), suggesting that the 6-kDa band seen in N-Flag-spexin reflected cleavage at a site significantly C-terminal to the predicted GRR site (Fig. 3A). Processing of spexin was further assessed by Western blotting of supernatant from C-Flag-spexin transfected cells that revealed bands at 12 kDa and 8 kDa (Fig. 5C). The 12-kDa band corresponds to the 12-kDa band seen for N-Flag-spexin (Fig. 5C), while the 8-kDa band represents C-terminally cleaved spexin. The absence of the 13-kDa band for C-Flag-spexin supports the argument that processing N-terminal of spexin peptide occurred in both N-Flag- and C-Flag-spexin and that this appears to proceed more efficiently for C-Flag-spexin.

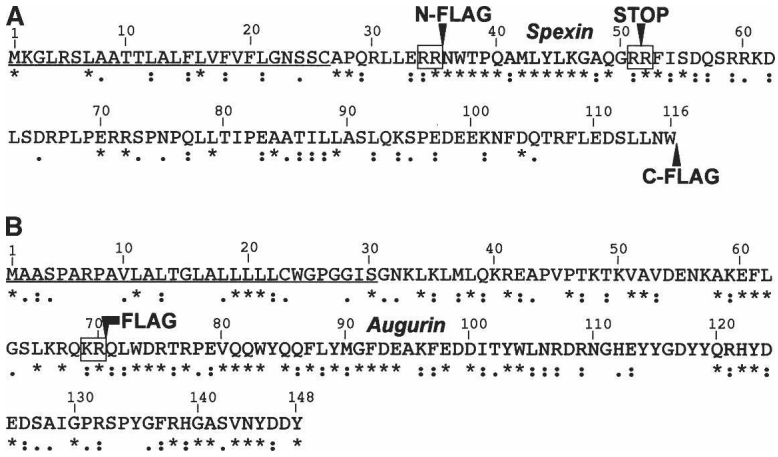

Figure 3. Primary structure of spexin and augurin. Sequence of spexin and augurin with the signal peptide underlined, pro-hormone cleavage sites boxed, and predicted processed peptide indicated in gray. Arrows indicate where the Flag antigen sequence (DYKDDDDK) was inserted to facilitate immunochemical detection of peptide products. Conservation among orthologs is shown *below*: (*) identity, (:) high homology; (·) low homology. The C-terminal glycine residue of the predicted spexin peptide is likely to be removed and the peptide amidated, a feature common to known peptide hormones. Both spexin and augurin peptides are enriched in aromatic amino acids.

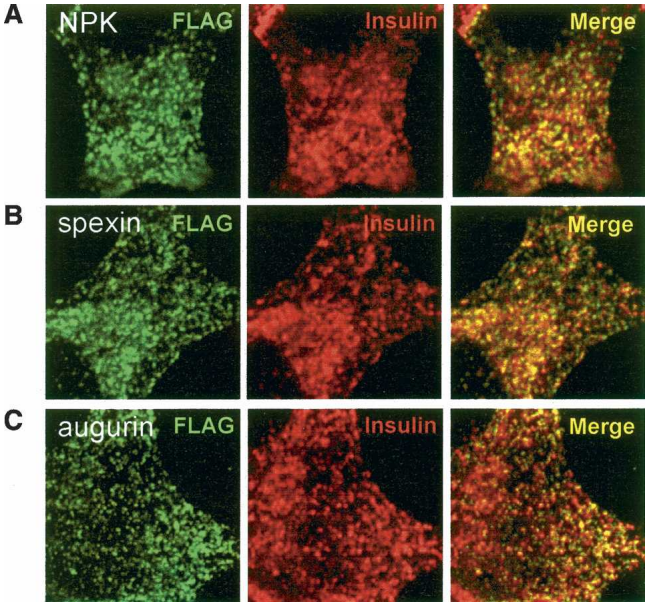

Figure 4. Colocalization of spexin and augurin with insulin in endocrine cells. Flag-tagged NPK, spexin, and augurin were transfected into rat pancreatic cells and fixed cells subjected to double immunofluorescence with Flag and insulin antibodies. (*A*) Neuropeptide K, (*B*) spexin, and (C) augurin show colocalization with insulin in small, cytoplasmic punctate structures. In all cases the majority of Flag immunoreactive puncta are also positive for insulin.

Western blotting of supernatant from Flag-augurin transfected cells revealed a pair of Flag-immunoreactive bands consistent with secretion of the pro-peptide and a processed variant (10 and 8 kDa) (Fig. 5D). Recognition of the Flag-augurin products by a Flag antibody that binds only N-terminal Flag antigen (M1) suggests that cleavage occurred at the predicted dibasic cleavage site just upstream of the Flag tag and supports the argument that the 8-kDa band reflects cleavage at a site near the C terminus of augurin. We speculate that this cleavage may occur at the non-canonical cleavage motif surrounding Arg132 (Fig. 3A). As expected, immunoblotting of the same supernatant with a Flag antibody that recognizes both N-terminal and embedded Flag epitopes (M2) revealed a high-molecular-weight product not recognized by the M1 antibody and corresponding to the full-length pro-peptide (Fig. 5D). Secretion and processing of Flag-augurin was confirmed in a second rat pancreatic cell line, RINm5f, that expresses high levels of insulin and forms distinct β-islet-like cell clusters (data not shown). These findings demonstrate that both spexin and augurin are processed and secreted when expressed in endocrine cells.

In situ hybridization localized spexin mRNA to the submucosal layer of esophagus and stomach fundus (Fig. 6A), a tissue containing the submucous plexus of the enteric nervous system and known to express several peptide hormones (e.g., gastrin-releasing peptide, vasoactive intestinal peptide) involved in the control of smooth muscle contractility (Costa et al. 2000). To examine whether the predicted peptide product of spexin could moderate smooth muscle contractility, a synthetic amidated spexin peptide, NWTPQAMLYLKGAQ-amide (Fig. 3A), was tested in a stomach explant contractility assay (Severini et al. 2000). The spexin peptide dose-dependently induced contraction of stomach muscle with an EC_{50} of 0.75 μM (Fig. 7). These findings demonstrate a biological activity for spexin and strongly support our hypothesis that spexin is a novel peptide hormone.

In situ hybridization revealed prominent augurin expression in mouse endocrine tissues, including the intermediate lobe of the pituitary, glomerular layer of the adrenal cortex, choroid plexus, and atrio-ventricular node of the heart (Fig. 6B). The intermediate lobe of the pituitary contains melanotrophs that produce alpha-melanocyte-stimulating hormone and beta-endorphin and whose role in mammalian physiology remains poorly understood (Mains and Eipper 1979; Saland 2001), while the glomerular layer of the adrenal cortex produces aldosterone and is involved in the regulation of salt homoeostasis (Connell and Davies 2005). In the heart, augurin mRNA localizes to a distinct set of cells that lie on either side of the ventricular valve and are likely to contribute to the cardiac conduction system. In situ hybridization on embryonic day 18.5 (E18.5) mouse revealed augurin mRNA in adrenal cortex, choroid plexus, and bone (data not shown). The expression pattern of augurin suggests that it is likely to express a secreted protein with a role in the modulation of salt and energy homeostasis, cardiovascular function, and cerebral spinal fluid composition.

Discussion

We have used a sequence-based approach to identify two candidate novel peptide hormones, which we called spexin and augurin. Both spexin and augurin were colocalized with insulin in the secretory pathway and were processed and secreted following

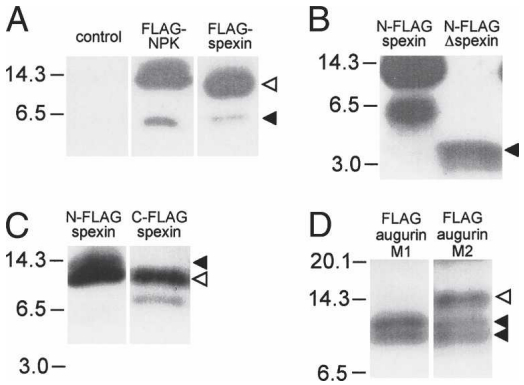

Figure 5. Identification of spexin and augurin in cell supernatants. Vector control and Flag-tagged NPK, spexin, and augurin were transfected into rat pancreatic cells in culture, and cell supernatants were harvested and submitted to immunoblotting with a Flag antibody. (*A*) Supernatants from Flag-NPK and N-Flag-spexin transfected cells contained high (solid arrow) and low (open arrow) mobility bands that reflected processing of Flag-tagged products from these constructs. (*B*) Supernatant from N-Flag-Δspexin transfected cells contained a 4-kDa band, suggesting that the 6-kDa product seen for N-Flag-spexin was the result of cleavage significantly C-terminal to the spexin peptide. (*C*) Supernatant from C-Flag-spexin contained two bands (12 and 8 kDa), confirming C-terminal cleavage of spexin pro-peptide. A 12-kDa product is seen for both N-Flag and C-Flag spexin (open arrow), while a 13-kDa product is seen only in N-Flag-spexin and corresponds to incompletely N-terminally processed spexin (closed arrow). (*D*) The presence of two bands (10 and 8 kDa, solid arrows) in supernatant from Flag-augurin transfected cells probed with M1 Flag antibody demonstrated cleavage of augurin at the putative pro-hormone cleavage site as well as close to the C terminus of the pro-peptide. The same immunoblot probed with M2 Flag antibody revealed an additional low-mobility product, confirming cleavage at the predicted dibasic cleavage site immediately adjacent to the Flag tag (open arrow).

transfection in endocrine cells. Furthermore, both spexin and augurin mRNA were expressed in endocrine tissues, and a predicted spexin peptide induced smooth muscle contractility in a stomach explant assay. Our findings confirm that most previously identified peptide hormones in the human proteome can be identified using a sequence-based screening approach. Our discovery of two novel peptide hormones suggests that our method is useful for the systematic screening of proteomes for biologically active peptides.

Several factors are likely to have prevented us from identifying additional candidate peptide hormones. First, we based our search on annotated protein databases that depend heavily on ESTs and full-length cDNAs for gene prediction. Given the poor expression level and restricted expression pattern of many known peptide hormones, it is possible that some peptide hormone genes are not present in these databases. Second, we decided to focus on hypothetical or poorly annotated proteins and did not apply our HMM to search for novel peptides produced by previously characterized, well-known genes. Recently, the peptide hormones obestatin and salusin were discovered to be produced by the ghrelin (Zhang et al. 2005) and Torsin 2A (Shichiri et al. 2003) genes, respectively, and further examples of such peptide hormone symbiosis are likely to exist. Finally, we used stringent criteria based on sequence conservation to identify spexin and augurin from among 61 high scoring candidates. Many proteins in this group failed the criterion requiring a shift in conservation across at least one cleavage site simply because they lacked distant orthologs and thus sufficient information to draw conclusions based on sequence conservation. Thus, it is possible that additional peptide hormones were overlooked among the top 61 high scoring proteins.

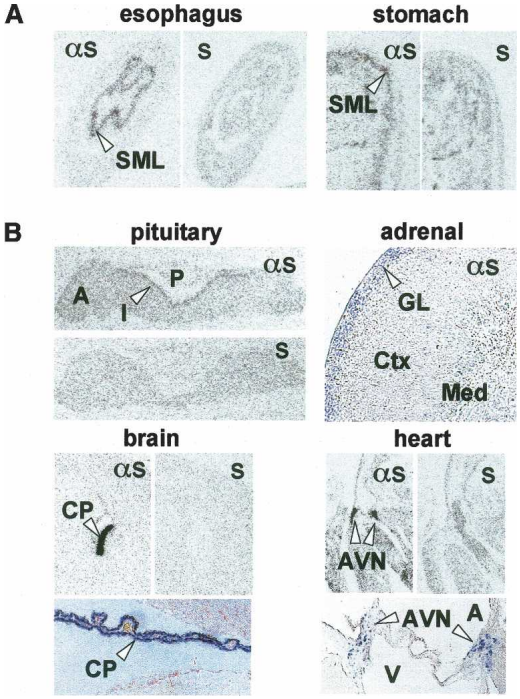

Figure 6. Expression of spexin and augurin mRNA in mouse tissues. (*A*) In situ hybridization with antisense (αs) and sense (s) probes detected spexin mRNA in the submucosal layer (SML) of the esophagus and stomach fundus. (*B*) In situ hybridization with antisense (αs) and sense (s) probes detected augurin mRNA in the intermediate lobe (I) of the pituitary (A, anterior; P, posterior), glomerular layer (GL) of the adrenal cortex (Ctx, cortex; Med, medulla), choroid plexus (CP), and atrio-ventricular node of the heart (AVN) (A, aorta; V, ventricle).

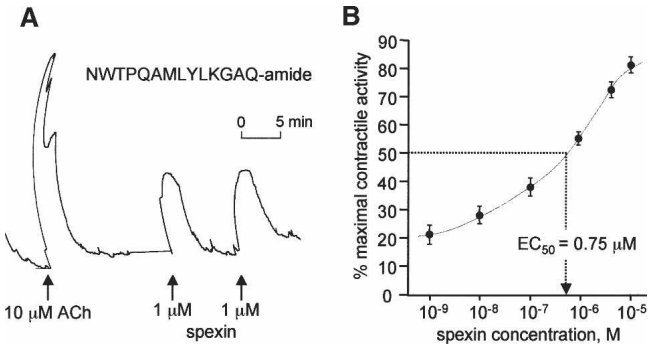

Figure 7. Spexin is a biologically active peptide hormone. (*A*) Representative muscle contractile response to 10 μM acetylcholine (ACh) and 1 μM spexin peptide (NWTPQAMLYLKGAQ-amide) in a rat stomach explant assay. Repeated administration of spexin peptide produced similar contactile responses. (*B*) Cumulative dose-response curve for contractile activity of spexin peptide on rat stomach explants (EC_{50} = 0.75 μM, N = 6). Error bars indicate standard error.

We believe that the HMM approach presented here could be extended to provide better sensitivity and specificity. First, the peptide hormone HMM could be combined with a DNA sequence HMM to create a peptide hormone-specific gene prediction method. Second, our use of orthology information was somewhat ad hoc, and integrating protein homology data internally into each state in the manner of phylogenetic HMMs in DNA sequence (Pedersen and Hein 2003; Siepel et al. 2005) could be envisioned. Third, it is clear that our background model (in this case, a simple one-state distribution of average amino acid content) is not rich enough to capture other features in real protein sequences, which may mislead the HMM. Nevertheless, this initial HMM, coupled with some downstream computational screens, has already provided several candidates for further biochemical screens and compares favorably to other experimental screening approaches.

Although there is considerable scope for improvement of the HMM, our initial results suggest that there is a low number (<15) of undiscovered peptide hormone precursors in the existing set of cDNA- and EST-supported genes (26% of 61 hypothetical or poorly annotated top scoring proteins) (see Fig. 2B). A more sophisticated HMM with less reliance on cDNA/EST based predictions will allow us to more confidently establish whether we have captured most peptide hormones with this biological model. The combination of computational screens and targeted biochemical verification will be a main route for further discoveries of peptide hormones.

Methods

Bioinformatics

Protein sequence data sets for the training of HMM states were retrieved from public databases using SRS (Sequence Retrieval System, http://www.expasy.org/srs5/) and Perl scripts—*signal peptide*: a previously curated set of 1011 nonredundant eukaryotic signal peptide-containing proteins (http://www.cbs.dtu.dk/ftp/signalp/euksig.red); *extracellular*: 5914 human SWISS-PROT entries with FtDescription = "extracellular"; *intracellular*: 7229

human SWISS-PROT entries with FtDescription = "cytoplasmic"; *peptide*: 448 human SWISS-PROT entries with FtKey = "peptide"; *transmembrane*: 15,730 human SWISS-PROT entries with FtKey = "transmem." Signal peptides were aligned using their hydrophobic and predicted cleavage sites features using a custom Perl script. The hydrophobic region was defined as the stretch of amino acids where the number of hydrophobic residues (AILFVMWY)/length was maximal. Amino acid frequencies and lengths for the signal peptide states were derived from this alignment. For pro-hormone convertase 1/2 and furin cleavage sites, data sets were retrieved from the MEROPS database and aligned at the cleavage site using a custom Perl script. Amino acid frequencies and lengths for the other feature states were directly derived from the relevant protein sets. This information was used to build the observation and transition matrices. Labeling and scoring were performed using Viterbi and forward–backward algorithms (Rabiner 1989), respectively, in Java. For the I_n, E_n, P_n, and T_n states, we modified the Viterbi algorithm to allow transition probabilities to depend on current state duration. This modification enabled us to model nongeometric transition probabilities (Ramesh and Wilpon 1992). Selection of candidate peptide hormones from the top 300 proteins was carried out by hand with the aid of a custom Java tool that displayed the score and assigned states of each protein. All custom scripts are available at http://bioinfo.embl.it/.

Cell culture and secretion assays

Unless otherwise noted, all cell culture was carried out in rat pancreatic β-TC3 cells (Efrat et al. 1988) in growing media (DMEM, 15% horse serum, 2.5% FBS). Forty-eight hours after transfection, cells were switched to serum-free RPMI media, and supernatant was collected for 24 h. In the case of Figure 5A, growing media was replaced immediately following transfection, and supernatant was collected for 48 h and immunoprecipitated with Flag antibodies. RINm5f cells (Gazdar et al. 1980) were grown in RPMI, 10% FBS. Supernatants were precipitated with acetone prior to immunoblotting. Transfection was carried out using Lipofectamine 2000 (Invitrogen) with transfection efficiency controlled by spiking 1:20 with a GFP expression plasmid (pLP-EGFP-C1 plasmid; Clontech). Human spexin (IMAGp958N21321), mouse augurin (IRAVp968F095D6), and human *TAC1* (IRATp970E0722D6) cDNA were obtained from RZPD. Flag-tagged expression constructs were designed with the Flag sequence (DYKDDDDK) inserted precisely at the beginning or end of the putative processed peptide. For neuropeptide K, Flag was inserted at residue 72 just before the first amino acid of neuropeptide K. For Flag-Δspexin, a stop codon was engineered just following the glycine residue of the putative spexin peptide. M2, and where indicated M1, Flag antibody was used (Sigma).

Immunocytochemistry

Rat pancreatic RINm5f cells were cultured in serum-containing RPMI medium, transfected with Flag-tagged NPK, spexin, and augurin, and grown for 48 h before fixation. Double fluorescent immunolabelling was performed with M2 Flag (Sigma) and insulin antibodies (Dako) following established protocols and visualized by confocal microscopy. Goat anti-mouse Alexa-488 and Goat anti-guinea pig Alexa-568 (Invitrogen) secondary antibodies were used.

In situ hybridization

Tissues and E18.5 embryos were dissected, fixed overnight in 4% paraformaldehyde, and embedded in paraffin. In situ hybridization using digoxigenin-labeled or ^{35}S-CTP-labeled probes on 8-µm paraffin sections was performed according to procedures previously described (Neubuser et al. 1995; Niederreither and Dolle 1998). Briefly, sections were dewaxed, rehydrated, digested with proteinase K, and hybridized with probe at 65°C. Post-hybridization washes in 20% formamide, 0.5× SSC were done at 60°C. The spexin probe was a 0.3-kb cDNA fragment cloned from mouse brain RNA (primers: 5′-ACAGGGTCGGAACATGAAGGG, 3′-AAGAGTCTGTCTTCCAAGAGTTCGC). The augurin probe was a 0.4-kb fragment amplified from mouse adrenal RNA (primers: 5′-CACCATGAGCACCTCGTCTGCG, 3′-TCTGTGGGCACC TCAGGG).

Explant assay

Albino Wistar female rats (250–350 g; Charles River) were sacrificed by inspiration of 75% CO_2, and stomach fundus muscles strips were isolated, washed in fresh Tyrode's solution (137 mM NaCl, 5.4 mM KCl, 0.5 mM $MgCl_2$, 1.8 mM $CaCl_2$, 10 mM glucose, 11.9 mM $NaHCO_3$, 0.4 mM NaH_2PO_4 at pH 7.4), mounted vertically in a 5-mL organ bath in oxygenated (95% O_2, 5% CO_2) Tyrode's solution, and maintained at 37°C. The segments were stretched to a tension of 2.0 g and allowed to equilibrate for 30–60 min, with the superfusion buffer changed every 15–20 min. At the beginning of each experiment, acetylcholine chloride (ACh 10^{-5} M) was applied to achieve a maximal control contraction. The potency of contractions was recorded isometrically by a strain gauge transducer (DY 1; Ugo Basile) and displayed on a recording microdynamometer (Unirecord; Ugo Basile). When reproducible responses to ACh were obtained, increasing concentrations (from 10^{-9} to 10^{-5} M) of synthetic amidated spexin peptide (NWTPQAMLYLKGAQ-amide; Primm) were applied every 2 min to establish a cumulative dose-response curve followed by washing and recovery for minimum 20 min. The EC_{50} was calculated by interpolation from the cumulative dose-response curve. Consecutively, single doses of spexin 10^{-6} M were applied until reproducible responses were obtained.

Acknowledgments

We thank W. Witke and F. Jönsson for antibodies and immunocytochemistry expertise, E. Lara-Pezzi for help with cell culture, D. Tosh for the gift of RINm5f cells, and S. Kang and P. Pilo Boyl for helpful suggestions and discussions. This work was supported by funds from the European Commission (N.R.). Manuscript charges were covered by EMBL.

References

Baggerman, G., Liu, F., Wets, G., and Schoofs, L. 2005. Bioinformatic analysis of peptide precursor proteins. *Ann. N. Y. Acad. Sci.* **1040:** 59–65.

Birney, E., Clamp, M., and Durbin, R. 2004. Genewise and genomewise. *Genome Res.* **14:** 988–995.

Braun-Menendez, E., Fasciolo, J.C., Leloir, L.F., and Muñoz, J.M. 1939. Hypertensin: The substance causing renal hypertension. *Nature* **144:** 980–981.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Burgus, R., Dunn, T.F., Desiderio, D., and Guillemin, R. 1969. Molecular structure of the hypothalamic hypophysiotropic TRF factor of ovine origin: Mass spectrometry demonstration of the PCA-His-Pro-NH_2 sequence. *C. R. Hebd. Seances Acad. Sci.* **269:** 1870–1873.

Chartrel, N., Dujardin, C., Youssef Anouar, J.L., Decker, A., Clerens, S., Do-Régo, J.-C., Vandesande, F., Llorens-Cortes, C., Costentin, J., Beauvillain, J.-C., et al. 2003. Identification of 26RFa, a hypothalamic neuropeptide of the RFamide peptide family with orexigenic activity. *Proc. Natl. Acad. Sci.* **100:** 15247–15252.

Conlon, J.M., Deacon, C.F., Grimelius, L., Cedermark, B., Murphy, R.F., Thim, L., and Creutzfeldt, W. 1988. Neuropeptide K-(1-24)-peptide:

Storage and release by carcinoid tumors. *Peptides* **9:** 859–866.
Connell, J.M.C. and Davies, E. 2005. The new biology of aldosterone. *J. Endocrinol.* **186:** 1–20.
Costa, M., Brookes, S.J.H., and Hennig, G.W. 2000. Anatomy and physiology of the enteric nervous system. *Gut* **47:** iv15–iv19.
Duckert, P., Brunak, S., and Blom, N. 2004. Prediction of proprotein convertase cleavage sites. *Protein Eng. Des. Sel.* **17:** 107–112.
Duguay, S.J., Lai-Zhang, J., and Steiner, D.F. 1995. Mutational analysis of the insulin-like growth factor 1 prohomone processing site. *J. Biol. Chem.* **270:** 17566–17574.
Efrat, S., Linde, S., Kofod, H., Spector, D., Delannoy, M., Grant, S., Hanahan, D., and Baekkeskov, S. 1988. β-Cell lines derived from transgenic mice expressing a hybrid insulin gene oncogene. *Proc. Natl. Acad. Sci.* **85:** 9037–9041.
Eipper, B.A., Stoffer, D.A., and Mains, R.E. 1992. The biosynthesis of neuropeptides: Peptide α-amidation. *Annu. Rev. Neurosci.* **15:** 57–85.
EUKSIG, Center for Biological Sequence Analysis. http://www.cbs.dtu.dk/ftp/signalp/euksig.red.
Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al. 2006. Pfam: Clans, web tools and services. *Nucleic Acids Res.* **34:** D247–D251.
Gazdar, A.F., Chick, W.L., Oie, H.K., Sims, H.L., King, D.L., Weir, G.C., and Lauris, V. 1980. β-Cell lines derived from transgenic mice expressing a hybrid insulin gene oncogene. *Proc. Natl. Acad. Sci.* **6:** 3519–3523.
Hinuma, S., Shintani, Y., Fukusumi, S., Iijima, N., Matsumoto, Y., Hosoya, M., Fujii, R., Watanabe, T., Kikuchi, K., Terao, Y., et al. 2000. New neuropeptides containing carboxy-terminal RFamide and their receptor in mammals. *Nat. Cell Biol.* **400:** 703–708.
Hökfelt, T. 1991. Neuropeptides in perspective: The last ten years. *Neuron* **7:** 867–879.
Hsu, S.Y. 1999. Cloning of two novel mammalian paralogs of relaxin/insulin family proteins and their expression in testis and kidney. *Mol. Endocrinol.* **13:** 2163–2174.
Jiang, Y., Luo, L., Gustafson, E.L., Yadav, D., Laverty, M., Murgolo, N., Vassileva, G., Zeng, M., Laz, T.M., Behan, J., et al. 2003. Identification and characterization of a novel RF-amide peptide ligand for orphan G-protein-coupled receptor SP9155. *J. Biol. Chem.* **278:** 27652–27657.
Kastin, A., ed. 2006. *Handbook of biologically active peptides.* Academic Press, New York.
Katafuchi, T., Kikumoto, K., Hamano, K., Kangawa, K., Matsuo, H., and Minamino, N. 2003. Calcitonin receptor-stimulating peptide, a new member of the calcitonin gene-related peptide family. *J. Biol. Chem.* **278:** 12046–12054.
Krogh, A., Brown, M., Mian, I.S., Sjölander, K., and Haussler, D. 1994. Hidden Markov models in computational biology. *J. Mol. Biol.* **235:** 1501–1531.
Livingstone, C. and Barton, G. 1993. Protein sequence alignments: A strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9:** 745–756.
Mains, R.E. and Eipper, B.A. 1979. Synthesis and secretion of corticotropins, melanotropins, and endorphins by rat intermediate pituitary cells. *J. Biol. Chem.* **16:** 7885–7894.
Neubuser, A., Koseki, H., and Balling, R. 1995. Characterization and developmental expression of Pax9, a paired-box-containing gene related to Pax1. *Proc. Natl. Acad. Sci.* **170:** 701–716.
Niederreither, K. and Dolle, P. 1998. In situ hybridization with ^{35}S-labeled probes for retinoid receptors. *Methods Mol. Biol.* **89:** 247–267.
Nielsen, H. and Krogh, A. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6:** 122–130.
Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6.
Ohtaki, T., Shintani, Y., Honda, S., Matsumoto, H., Hori, A., Kanehashi, K., Yasuko Terao, S.K., Takatsu, Y., Masuda, Y., Ishibashi, Y., et al. 2001. Metastasis suppressor gene KISS-1 encodes peptide ligand of a G-protein-coupled receptor. *Nature* **411:** 613–617.
Park, Y., Kim, Y.-J., and Adams, M.E. 2002. Identification of G protein-coupled receptors for *Drosophila* PRXamide peptides, CCAP, corazonin, and AKH supports a theory of ligand–receptor coevolution. *Proc. Natl. Acad. Sci.* **99:** 11423–11428.
Pedersen, J.S. and Hein, J. 2003. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* **19:** 219–227.
Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77:** 257–286.
Ramesh, P. and Wilpon, J.G. 1992. Modeling state durations in hidden Markov models for automatic speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-92*, Vol. 1, pp. 381–384. San Francisco, CA.
Rawlings, N.D., Morton, F.R., and Barrett, A.J. 2006. The peptidase database. *Nucleic Acids Res.* **34:** D270–D272.
Saland, L.C. 2001. The mammalian pituitary intermediate lobe: An update on innervation and regulation. *Brain Res. Bull.* **54:** 587–593.
Schmidt, W.E., Kratzin, H., Eckart, K., Drevs, D., Mundkowski, G., Clemens, A., Katsoulis, S., Schäfer, H., Gallwist, B., Kreutzfeldt, W., et al. 1991. Isolation and primary structure of pituitary human galanin, a 30-residue non-amidated neuropeptide. *Proc. Natl. Acad. Sci.* **88:** 11435–11439.
Severini, C., Salvadori, S., Guerrini, R., Falconieri-Erspamer, G., Mignogna, G., and Erspamer, V. 2000. Parallel bioassay of 39 tachykinins on 11 smooth muscle preparations. Structure and receptor selectivity/affinity relationship. *Peptides* **21:** 1587–1595.
Shichiri, M., Ishimaru, S., Ota, T., Nishikawa, T., Isogai, T., and Hirata, Y. 2003. Salusins: Newly identified bioactive peptides with hemodynamic and mitogenic activities. *Nat. Med.* **9:** 1166–1172.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Steiner, D.F. 1998. The preprotein convertases. *Curr. Opin. Chem. Biol.* **2:** 31–39.
Su, T., Liu, H., and Lu, S. 1998. Cloning and identification of cDNA fragments related to human esophageal cancer. *Zhonghua Zhong Liu Za Zhi* **20:** 254–257.
Thomas, G. 2002. Furin at the cutting edge: From protein traffic to embryogenesis and disease. *Nat. Rev. Mol. Cell Biol.* **3:** 753–766.
Thomas, G., Moffatt, P., Salois, P., Gaumond, M.-H., Gingras, R., Godin, E., Miao, D., Goltzman, D., and Lanctot, C. 2003. Osteocrin, a novel bone-specific secreted protein that modulates the osteoblast phenotype. *J. Biol. Chem.* **278:** 50563–50571.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Usdin, T.B., Hoare, S.R.J., Wang, T., Mezey, E., and Kowalak, J.A. 1999. TIP39: A new neuropeptide and PTH2-receptor agonist from hypothalamus. *Nat. Neurosci.* **2:** 941–943.
Vassilatis, D.K., Hohmann, J.G., Zeng, H., Li, F., Ranchalis, J.E., Marty, T., Mortrud, A.B., Rodriguez, S.S., Weller, J.R., Wright, A.C., et al. 2003. The G-protein-coupled receptor repertoires of human and mouse. *Proc. Natl. Acad. Sci.* **100:** 4903–4908.
Zhang, Z. and Wood, W.I. 2002. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **19:** 307–308.
Zhang, J.V., Ren, P.-G., Avsian-Kretchmer, O., Luo, C.-W., Rauch, R., Klein, C., and Hsueh, A.J.W. 2005. Obestatin, a peptide encoded by the ghrelin gene, opposes ghrelin's effects on food intake. *Science* **310:** 996–999.

Received July 13, 2006; accepted in revised form November 30, 2006.