PenBase, the shrimp antimicrobial peptide penaeidin database: Sequence-based classification and recommended nomenclature

Yannick Gueguen1*, Julien Garnier1, Lorenne Robert1, Marie-Paule Lefranc2, Isabelle Mougno3, Julien de Lorgeril1, Michael Janc3, Paul S. Gross4, Gregory W. War5, Brandon Cuthbertson5, Margherita A. Barracco6, Philippe Bulet6, André Aumelas7, Yinshan Yang7, Dong Bo8, Jianhai Xiang8, Anchalee Tassanakajon9, David Piquemal10, Evelyne Bachère1

1Ifremer-CNRS-University of Montpellier 2, UMR 5171 Génome Population Interaction Adaptation, 2 Place E. Bataillon, CC80, F-34095 Montpellier cedex 5, France.
2Laboratoire d’ImmunoGénétique Moléculaire, LIGM, Université Montpellier II, Institut Universitaire de France, UPR CNRS 1142, IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France
3Laboratoire d’Informatique, de Robotique et de Microélectronique de Montpellier, LIRMM, UMR 5506 CNRS-UM2, 161 rue Ada 34392 Montpellier cedex 5, France.
4Department of Biochemistry and Molecular Biology, Medical University of South Carolina, Box 250509, 173 Ashley Avenue, Charleston, SC 29425, USA.
5Departmentamento de Biologia Celular, Embriologia e Genética, Universidade Federal de Santa Catarina, Florianópolis, Brasil.
6Atheris Laboratories, Case postale 314, CH-1233 Bernex-Geneva, Switzerland.
7Centre de Biochimie Structurale, Université Montpellier I, 15 avenue Charles Flahaut, F-34060 Montpellier cedex 2, France.
8Institute of Oceanology, Chinese Academy of Science, 7 Nanhai Road, Qingdao 266071 P.R. China.
9Shrimp Molecular Biology and Genomics Laboratory, Department of Biochemistry, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand.
10Skuld-Tech Company, Université Montpellier II, Place E. Bataillon, CC091, F-34095 Montpellier cedex 5, France.

*Corresponding author: ygueguen@ifremer.fr

Abstract

Antimicrobial peptides play a major role in innate immunity. The penaeidins, initially characterized from the shrimp Litopenaeus vannamei, are a family of antimicrobial peptides that appear to be expressed in all penaeid shrimps. As of recent, a large number of penaeid nucleotide sequences have been identified from a variety of penaeid shrimp species and these sequences currently reside in several databases under unique identifiers with no nomenclatural continuity. To facilitate research in this field and avoid potential confusion due to a diverse number of nomenclatural designations, we have made a systematic effort to collect, analyse, and classify all the penaeidin sequences available in every database. We have identified a common penaeidin signature and subsequently established a classification based on amino acid sequences. In order to clarify the naming process, we have introduced a “penaeidin nomenclature” that can be applied to all extant and future penaeidins. A specialized database, PenBase, which is freely available at http://penbase.immunagua.com, has been developed for the penaeidin family of antimicrobial peptides, to provide comprehensive information about their properties, diversity and nomenclature.

Keywords: penaeidin, penaeid shrimp, antimicrobial peptide, crustacea, sequence database, nomenclature.
Introduction

Antimicrobial peptides (AMP) play a major role in innate immunity, conserved in evolution, and present in all phyla of the living kingdom. They are mostly cationic and amphipathic molecules although they present a great diversity in terms of structural features as well as biological properties and functions. More than 850 antimicrobial peptides have now been discovered in plants, vertebrates, and invertebrates (Antimicrobial Sequences Database, AMSDb [1]) and very recently, two databases dedicated to antimicrobial peptides, the Antimicrobial Peptide Database (APD) [2] and ANTIMIC [3], were created. To date, in penaeid shrimps, two kinds of antimicrobial peptides have been fully characterized, namely the penaeidins from hemocytes [4] and anionic hemocyanin-derived peptides isolated from shrimp plasma [5]. In addition, recent studies utilizing a genomic approach led to the characterization of other antimicrobial effectors in shrimp, i.e. anti-LPS factor (ALF) and crustin [6-8].

Penaeidins were first characterized from Litopenaeus vannamei using a biochemical approach and molecular cloning techniques. Three peptides (initially named penaeidin 1, 2 and 3) were isolated in their active and mature forms (5.48 – 6.62 kDa) from the hemocytes of shrimp [4]. However, subsequent phylogenetic analysis indicated that penaeidin 1 and -2 could be classified in the same class [9]. Penaeidins are composed of an N-terminal proline-rich domain, followed by a C-terminal domain containing 6 cysteine residues organized in two doublets. This overall structure is quite unique among the AMP families [4]. The antimicrobial activity spectrum of penaeidin 2 and 3 from L. vannamei has been established through the production and analysis of recombinant peptides [10]. More recently, a new subgroup of penaeidins, named penaeidin 4 by their authors, has been identified in L. vannamei using a genomic approach [9], and a synthetic chemical peptide was produced to investigate and characterize the functional properties and spectrum of activity [11]. Penaeidins possess antibacterial activity predominantly directed against Gram-positive bacteria and antifungal activity against filamentous fungi. Very recently, the solution structure of the recombinant penaeidin 3 from L. vannamei and of the synthetic penaeidin 4 from L. setiferus have been determined, revealing the overall organization of the two domains and the arrangement of the disulfide bonds [12-13].

Recent studies have revealed the presence of penaeidin mRNAs in different penaeid shrimp species [6, 8, 14-15]. Moreover, in both L. vannamei and L. setiferus, the penaeidin subgroups (penaeidin 2, 3 and 4) were shown to be expressed in a single individual [6]. Expressed sequence tag (EST) analysis from hemocyte cDNA libraries has shown that penaeidin transcripts are very abundant. In fact, penaeidins appear to represent 10.7 and 20% of all the sequences isolated from hemocytes of L. vannamei and L. setiferus, respectively [6].

Most groups working on shrimp from different parts of the world have now isolated numerous penaeidin sequences mostly by genomic approaches, and this family appears to be ubiquitous among penaeid shrimp species. At the moment more than 200 sequences can be found in the EMB/GenBank/DDBJ databases, some of which have yet to be published. The current nomenclature for penaeidins is very confusing, with several different names given to the same gene product and distinct names (i.e. penaeidin number) given to variants of the same molecule. Therefore, we suggest that the “penaeidin field” adopts a common nomenclature based on amino acid sequence similarity. Herein, we introduce a new database, PenBase, to provide comprehensive information about penaeidin properties, diversity and nomenclature. PenBase has been implemented according to the IMGT Scientific chart rules that are based on the IMGT-ONTOLOGY concepts [16]. The description labels, the subgroup and penaeidin nomenclature, the amino acid unique numbering, and the penaeidin origin are based on the DESCRIPTION, CLASSIFICATION, NUMEROTATION and OBTENTION concepts, respectively [17-19]. Owing to that standardized approach, PenBase provides for the first time a coherent management of data related to penaeidins.
**Penaeidin signature determination**

We have made a systematic effort to collect and analyse all publicly accessible penaeidin sequence data. Penaeidin sequences were analysed from 8 shrimp species: *Litopenaeus vannamei* (Pacific White Shrimp or White leg Shrimp), *L. setiferus* (Atlantic White Shrimp or Northern White Shrimp), *L. stylirostris* (Blue Shrimp), *Farfantepenaeus paulensis* (Sao Paulo Shrimp), *L. schmitti* (Southern white shrimp), *Penaeus semisulcatus* (Green tiger Prawn), *Fenneropenaeus chinensis* (Fleshy Prawn) and *Penaeus monodon* (Giant or Black tiger prawn). Three different categories of data were collected: cDNA sequences, peptidic sequences with no corresponding nucleotide sequence and nucleotide sequences originating from EST projects. The determination of the family signature is based on common amino acid sequences shared by penaeidins, and that include 2 regions, the signal peptide (L-REGION) and the mature peptide (PEPTIDE-REGION). In our analysis, 104 sequences coming from EST project were not used because they contained uncertain nucleotides sequences in the PEPTIDE-REGION. Multiple alignments using ClustalX [20] were constructed with 110 sequences and a penaeidin family signature was designated based on the coding amino acid sequences. Many sequences appear 100% identical at the amino acid level while only 34 penaeidin sequences differ at the level of at least one amino acid. The PEPTIDE-REGION from the penaeidin family is characterized by the presence of 2 domains, a proline-rich N-terminal domain and a cysteine-rich C-terminal domain presenting two doublets of cysteine residues. NMR solution structure studies revealed that the N-terminal domain rich in proline residues is unconstrained. In contrast, the structure of the C-terminal domain stabilized by three conserved disulfide bonds in the 1-3, 2-5 and 4-6 patterns is well defined [12]. This domain consists in an amphipathic helix linked to upstream and downstream coils by the 1-3 and 2-5 disulfide bonds and these two coils are in turn linked together by the 4-6 disulfide bond. This disulphide arrangement would be characteristic of the penaeidin family according to recent studies on solution structure of penaeidin 4 for which a similar pattern was determined [13]. From the alignment and structural studies, the signature of the penaeidin family has been defined as shown in figure 1 by 10 conserved amino acids in the signal peptide sequence; a threonine and two proline residues conserved in the N-terminal domain; and the conserved cysteine array of the C-terminal structured domain.

**Penaeidin nomenclature and subgroup determination**

A detailed analysis of the sequence alignments led to the classification of all the existing penaeidin amino acid sequences in 3 distinct subgroups: **Penaeidin 2 (PEN2)**, **Penaeidin 3 (PEN3)**, **Penaeidin 4 (PEN4)**. Based on these analyses, we introduced the following nomenclature that will be now used for the penaeidins. The name of a penaeidin is composed of (i) 6 letters in italics (3 for the genus and 3 for the species) followed by a space, (ii) the penaeidin subgroup followed by a hyphen and (iii) a number for the identification inside the subgroup (Example: **Litvan PEN3-1**: sequence from *Litopenaeus vannamei*, PEN3 subgroup, first sequence identified. The determination of the penaeidin subgroups (PEN2, PEN3 and PEN4) is based on the amino acid sequences of the mature peptide (PEPTIDE-REGION). Following amino acid alignment analyses, conserved key residues appear to be characteristic for each of the three subgroups (Figure 2). These key residues are highlighted in yellow in figure 2. For example, all penaeidins of the “subgroup PEN3” possess the following signature: Gln\(^1\), Gly\(^5\), Arg\(^{-}\), Val\(^{18}\) or Gly\(^{18}\), Ser\(^{35}\), Arg\(^{37}\) or His\(^{37}\) or Pro\(^{37}\), Gln\(^{43}\), Ser\(^{46}\) or Ala\(^{46}\).

One penaeidin peptide name (example **Litvan PEN3-1**) can correspond to several related sequences which may differ at the nucleotide level without amino acid changes in the PEPTIDE-REGION. The different cDNA sequences that correspond to nucleotide variations are identified by an asterisk followed by a number, *01 being assigned to the first cDNA sequence (example **Litvan PEN3-1*01). In addition, one penaeidin peptide name and sequence (example **Litvan PEN3-1**), can also correspond to one or several EST sequences. If an EST sequence is identical to an already known cDNA PEPTIDE-REGION nucleotide sequence, the EST is classified under the same name (example **Litvan PEN3-1*01**) in the EST sequence section. If not, the EST is classified as an EST sequence (example **Litvan PEN3-1*EST1**) until the sequence is confirmed by sequencing and classified as a cDNA.
Organisation of the database

PenBase was built on the Debian Sarge GNU/Linux operating system (http://www.debian.org/) using the freeware Apache 2 web server (http://www.apache.org/), PHP script language (http://www.php.net/) and PostgreSQL relational database system (http://www.postgresql.org/). The PenBase home page contains an introduction to the penaeidin family and lists the accessible sections in the database:

1. Properties
2. Signatures and subgroups
3. Nomenclature
4. Penaeidin Database
5. Primer library
6. Bibliography
7. Forum
8. Penaeidin links

The database section lists all known penaeidins sorted by shrimp species or subgroups. Individual penaeidin records contain detailed information about each penaeidin including peptide name, nucleotide and peptide sequence data, biochemical properties (including pI, molecular mass, signal peptide length, proline-rich and cysteine-rich domain lengths), contact information, references (including authors, article titles, journal, page, volume and year) and links to PubMed and EMBL/GenBank. In this section, Blast and alignment tools are available. The “Signature and subgroup” section describes the criteria for penaeidin identification and classification. The forum section is a place where anyone can exchange information or ask questions regarding penaeidins. References throughout the database are linked to the Bibliography section. The bibliography includes a list of all the papers related to penaeidins.

Availability and future developments

PenBase can be accessed directly at http://penbase.immunaqua.com. Users of PenBase are requested to cite this article when referencing the database. Researchers in this field are invited to make suggestions and comments by using the forum section. PenBase contains more than 200 entries of penaeidins and is expected to quickly grow up with the rapid development of genomic and proteomic projects. As more information about penaeidins becomes available, the database will be expanded and improved accordingly.

Acknowledgements

This study was part of a collaborative project supported by the European Commission, DG XII, in the program International Co-operation with Developing Countries, INCO-DC, Contract n° ICA4-CT-2001-10023 (IMMUNAQUA, http://immunaqua.com). It is also supported by IFREMER, CNRS and the University of Montpellier II. M. Janech was supported by the National Science Foundation, International Research Fellowship Program. M.A. Barracco was supported by a post-doctoral scholarship given by CAPES (Brazil) and J. de Lorgeril was supported by a doctoral funding from the New Caledonian Territories and the IFREMER. This work was partially supported by the National Science Foundation (NSF grant #MCB0110576 to PS Gross & MCB0315393 to GW Warr), the National Marine Fisheries Service (NA03NMF4720362 to PS Gross & GW Warr), the South Carolina Sea Grant Consortium (R/MT-6 to PS Gross), United States Department of Agriculture (USDA NRICGP-CSREES/AREA Grant# 2002-35201-11620 to PS Gross) and the South Carolina Department of Natural Resources. Any opinions, findings, or conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the supporting bodies mentioned herein.
References


Figure 1: Penaeidin signature.
In bold are represented the amino acids conserved in all the penaeidins. X is for any residue. Numbers 1, 2, 3, 4, 5, and 6 above the sequence are related to the cysteine rank in the mature sequence and used in the text to describe the disulfide bond pattern. Numbers below the sequence correspond to the amino acid position in the mature peptide.
Figure 2: Determination of the penaeidin subgroup signatures

In grey are represented the key amino acids that are used for the subgroup signatures. In bold, amino acids which are specific of only one subgroup. The double arrow indicates the peptide signal cleavage site. **Litopenaeus vannamei**: Litvan, **Litopenaeus setiferus**: Litsset, **Litopenaeus stylirostris**: Litsch, **Penaeus semisulcatus**: Pensem, **Fenneropenaeus chinensis**: Fenchi, **Penaeus monodon**: Penmon, **Farfantepenaeus paulensis**: Farpau.