BioSystems 113 (2013) 1–8

Contents lists available at SciVerse ScienceDirect

BioSystems

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

# Stepwise identification of potent antimicrobial peptides from human genome

Li Yan[a,*], Yuxian Yan[b], Hongqi Liu[a], Qi Lv[b]

[a] Department of Burn and Plastic Surgery, Affiliated Hospital of Logistics University of Chinese People's Armed Police Forces, Tianjin 300162, China
[b] The Central Laboratory of Logistics University of Chinese People's Armed Police Forces, Tianjin 300162, China

ARTICLE INFO

*Article history:*
Received 26 February 2013
Received in revised form 18 March 2013
Accepted 31 March 2013

*Keywords:*
Antibiotics
Infection
Antibacterial agent
Bioinformatics
Rational drug design

ABSTRACT

The increasing incidence of hospital acquired infections caused by antibiotic resistant pathogens has led to an increase in morbidity and mortality, finding alternative antibiotics unaffected by resistance mechanisms is fundamentally important for treating this problem. Naturally occurring proteins usually carry short peptide fragments that exhibit noticeable biological activity against a wide variety of microorganisms such as bacteria, fungi and protozoa. Traditional discovery of such antimicrobially active fragments (*i.e.* antimicrobial peptides, AMPs) from protein repertoire is either random or led by chance. Here, we report the use of a rational protocol that combines *in silico* prediction and *in vitro* assay to identify potential AMPs with high activity and low toxicity from the entire human genome. In the procedure, a three-step inference strategy is first proposed to perform genome-wide analysis to infer AMPs in a high-throughput manner. By employing this strategy we are able to screen more than one million peptide candidates generated from various human proteins, from which we identify four highly promising samples, and subsequently their antibacterial activity on five strains as well as cytotoxicity on human myoblasts are tested experimentally. As a consequence, two high-activity, low-toxicity peptides are discovered, which could be used as the structural basis to further develop new antibiotics. In addition, from 1491 known AMPs we also derive a quantitative measure called antibacterial propensity index (API) for 20 naturally occurring amino acids, which shows a significant allometric correlation with the theoretical minimal inhibitory concentration of putative peptides against Gram-positive and Gram-negative bacteria. This study may provide a proof-of-concept paradigm for the genome-wide discovery of novel antimicrobial peptides by using a combination of *in silico* and *in vitro* analyses.



## 1. Introduction

In past centuries, one of the major achievements of medicine was the development of antibiotics, which can kill a broad spectrum of microorganisms. Unfortunately, the widespread use of antibiotics has promoted the emergence of antibiotic-resistant strains that often pose serious health problems. For example, nosocomial bacterial infections that are resistant to therapy cost more than $2 billion and account for more than 80,000 direct deaths in the United States and 5000 in England every year (Martín-Madrazo et al., 2009); the magnitude places it among the 10 leading causes of death in the North America. Discovery of new classes of antibiotics has long been great attraction in the medicine community. However, the past 40 years have seen only three new classes of antibiotics enter market (lipopeptides, oxazolidinones, and streptogramins), all geared toward Gram-positive bacterial infections (Marr et al., 2006).

Many efforts have been directed toward finding alternative antibiotics unaffected by resistance mechanisms. Antimicrobial peptides (AMPs) are regarded as one of the most promising alternatives in this regard (Jenssen et al., 2006). These peptides are typically 12–50 amino acids in length with 2–9 excess basic residues (Arg or Lys) and up to 50% hydrophobic amino acids; they fall into four major structural categories based on their amphiphilic conformations that are preformed or occur after membrane interaction, namely, amphipathic α-helices, β-sheet with 2–4 strands, loop coils, and extended structures (Cherkasov et al., 2008). To date, more than 800 natural peptides have been described that not only kill pathogenic microorganisms, including Gram-positive and Gram-negative bacteria, viruses, protozoa and fungi, but also play a central role in recruiting and promoting elements of the innate immune system (Peters et al., 2010). Nowadays, AMPs are used not only to prevent infections, but also in a variety of applications such as controlling symbiotic microbial communities (Mergaert and Kondorosi, 2012), killing cancer cells (Schweizer, 2009), and

* Corresponding author. Tel.: +86 2223701458.
*E-mail addresses:* lli_yan@126.com, yyyanli1234@126.com (L. Yan).

0303-2647/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.biosystems.2013.03.021

interfering virus replications (Kovalchuk et al., 2007). With rapid increase in the number of naturally occurring and chemically synthesized AMPs, some severe problems associated with the peptidic therapeutics are emerged, such as cytotoxicity (Pacor et al., 2002) and hypersensitivity (Alberola et al., 2004). For example, Pacor et al. found that some highly cationic, artificial α-helical AMPs show marked toxicity on blood cells (lymphocytes and erythrocytes) (Levy, 2000). Reducing the helical propensity of peptides would result in the substantial decrease in both toxicity and antibacterial potency.

In recent years, it has been found that peptide fragments derived from self-proteins are the ideal agents used to treat human diseases with high safety, low toxicity, and immune tolerance (Sato et al., 2006), suggesting that the human proteins could be exploited to generating new and safe AMPs. For example, the peptide RQA21 (RQAREHSERKKRRRESECKAA), which corresponds to the C-terminal region of human extracellular superoxide dismutase (SOD), has recently been found to exert potent antimicrobial effects against *Escherichia coli*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Bacillus subtilis*, and *Candida albicans* (Pasupuleti et al., 2009). In addition, although a wide variety of human proteins and peptides have antimicrobial activity and play important roles in innate immunity, discovery of AMPs by comprehensively screening the whole human genome still remains largely unexploited. Very recently, Amaral and co-workers have reported the successful identification of AMPs by theoretical survey of eukaryotic genome databases (Amaral et al., 2012). The work promoted us to perform a genome-wide analysis to identify potent AMPs from human proteins by using a new combination approach that consists of two distinct but complementary aspects, *i.e.* computational prediction and experimental assay.

In recent years, bioinformatic techniques have been widely applied to analyze and design AMPs (Loose et al., 2006). Here, we first proposed a three-step inference strategy to extract potential AMP candidates from ~1,000,000 human protein fragments of 9-amino acid-long; this method can be regarded as a good compromise between computational efficiency and accuracy. Subsequently, few highly promising peptides are synthesized, purified, and assayed to determine their antibacterial potency on three Gram-negative bacteria and two Gram-positive bacteria, and to test their cytotoxicity on normal human cells. We also investigate the molecular mechanism and biological implication underlying the sequence and structure of newly confirmed AMPs and their interactions with bacterial phospholipid mimics.

## 2. Materials and methods

### 2.1. Datasets

Four datasets were compiled from distinct sources to serve for this study, namely, ARA (**A**MPs **R**etrieved from **A**PD2), AMA (**A**MPs with **M**IC **A**ctivity), ALA (**A**MPs with **L**uminescence **A**ctivity), and HPFS (human protein fragment set).

The ARA curates experimentally confirmed cationic AMPs deposited in the APD2 database (Wang et al., 2009). Here, only the AMPs composed by conventional amino acid types, possessing excess positive charges, and against both Gram-positive and Gram-negative bacteria were considered. As a result, 1491 AMPs were extracted from the APD2 to define the ARA.

AMA and ALA are the collections of cationic AMPs with experimentally measured activity. The AMA consists of 101 so-called CAMEL-s peptides; these peptides were created by the respective fusion of the C- and N-terminal sequences of natural peptides Cecropin and Melittin (Mee et al., 1997; Edlund et al., 1998; Oh et al., 2000), and their antibacterial activity was quantified as the average of minimal inhibitory concentrations (MICs) on 24 bacterial strains (Oren and Shai, 1999). Previously, the CAMEL-s peptide dataset have already been successfully applied to develop several AMP prediction models (Cherkasov and Jankovic, 2004; Shu et al., 2009). The ALA compiles the complete single-mutation profile of a 12-mer peptide Bac2A (RLARIVVIRVAR-$NH_2$), the linear variant of the smallest naturally occurring AMP––bovine neutrophil cationic peptide bactenecin (RLCRIVVIRVCR) (Wu and Hancock, 1999). The total 240 ($12 \times 20$) mutants were synthesized on cellulose sheets with high-throughput SPOT technique and screened using a loss of energy-dependent luminescence activity in a *luxCDABE*-expressing *P. aeruginosa* isolate (Hilpert et al., 2005).

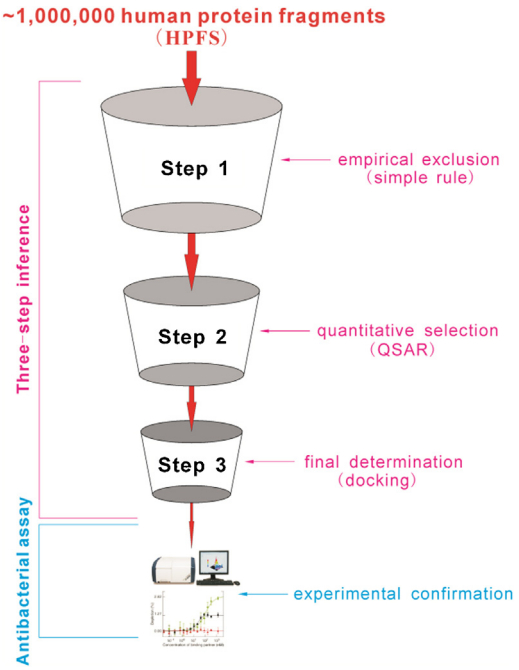


**Fig. 1.** Schematic representation of the three-step inference and antibacterial assay procedure to identify potent AMPs from the human genome containing about one million valid protein fragments.

The HPFS is an assembly of valid human protein fragments with 12-residue-long, the shortest length of AMPs found in natural proteins (Hilpert et al., 2006). The current UniProt database includes 131,076 human protein entries (August, 2012) (Uniprot, 2010), from which only 19,464 manually reviewed, nonredundant sequences (identity ≤ 90%) were extracted. These sequences were then virtually broken into 1,127,081 overlapping 12-mer peptide fragments to constitute the HPFS. Subsequent analyses will be addressed on this set to derive desired AMPs (vide post).

### 2.2. Three-step inference

#### 2.2.1. Overview of the method

Schematic representation of the three-step inference method proposed here to computationally identify potential AMPs from the human genome is shown in Fig. 1. In this study, only very few highly promising peptides will be selected from the vast HPFS pool consisting of ~1,000,000 human protein fragments and ultimately put forward antibacterial assay procedure. Apparently, we require to developing a reliable but also efficient strategy to perform such selection.

Because there would be a large proportion of inactive fragments existed in the HPFS, we first carried out a coarse-grained empirical screening to fast exclude those of typical non-AMPs from the crude pool (Step 1). Subsequently, several quantitative structure–activity relationship (QSAR) models were built using different machine learning methods and peptide characterization techniques; these models were then employed to blindly vote for left peptide candidates and only the peptides with unanimous vote were further considered (Step 2). Finally, we performed a long-scale atomistic molecular docking to carry out intensive sampling in the interaction spaces of the few voted peptides with the model molecule that mimic the outer membrane of bacteria; the docking-resultant binding energies were used to accurately evaluate the affinity of peptides to bacterial outer membranes (Step 3).

#### 2.2.2. Step 1: empirical exclusion of typical non-AMPs

The goal of the first step was to exclude those peptides that do not follow the typical feature of AMPs, rather than to select effective AMPs. Therefore, we herein suggested a very empirical rule to filter out typical non-AMPs from the crude HPFS pool, that is, a peptide should be discarded if it matches at least one of the two criteria. (i) The peptide is neutral or negatively charged. (ii) The average value of antibacterial propensity indices (API) of all amino acid residues over the peptide is less than 1.2, a safer threshold for excluding those non-AMPs. Here, API values of 20 natural amino acids were derived using the 1491 experimentally confirmed AMPs in ARA set (Table 1), which are simply defined as follows:

$$\mathrm{API}_{\mathrm{Xaa}} = \frac{p_{\mathrm{Xaa}}}{1/20} \quad (1)$$

where Xaa indicates any one of the 20 amino acids, and $p_{\mathrm{Xaa}}$ is the probability of Xaa occurring in AMPs, which can be characterized as the ratio of the number of