Biotechnology Advances 35 (2017) 337–349

Contents lists available at ScienceDirect

Biotechnology Advances

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

Research review paper

Computational tools for exploring sequence databases as a resource for antimicrobial peptides

W.F. Porto [a,c], A.S. Pires [a], O.L. Franco [a,b,*]

[a] Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia Universidade Católica de Brasília, Brasília, DF, Brazil
[b] S-Inova Biotech, Pós-graduação em Biotecnologia, Universidade Católica Dom Bosco, Campo Grande, MS, Brazil
[c] Porto Reports, Brasília, DF, Brazil

ARTICLE INFO

Article history:
Received 7 November 2016
Received in revised form 12 January 2017
Accepted 8 February 2017
Available online 12 February 2017

Keywords:
Data mining
Structural genomics
Local alignments
Profile-HMM
Regular expression
Antimicrobial activity prediction
Molecular modelling

ABSTRACT

Data mining has been recognized by many researchers as a hot topic in different areas. In the post-genomic era, the growing number of sequences deposited in databases has been the reason why these databases have become a resource for novel biological information. In recent years, the identification of antimicrobial peptides (AMPs) in databases has gained attention. The identification of unannotated AMPs has shed some light on the distribution and evolution of AMPs and, in some cases, indicated suitable candidates for developing novel antimicrobial agents. The data mining process has been performed mainly by local alignments and/or regular expressions. Nevertheless, for the identification of distant homologous sequences, other techniques such as antimicrobial activity prediction and molecular modelling are required. In this context, this review addresses the tools and techniques, and also their limitations, for mining AMPs from databases. These methods could be helpful not only for the development of novel AMPs, but also for other kinds of proteins, at a higher level of structural genomics. Moreover, solving the problem of unannotated proteins could bring immeasurable benefits to society, especially in the case of AMPs, which could be helpful for developing novel antimicrobial agents and combating resistant bacteria.

Contents

1. Introduction . . . 337
2. AMPs in databases . . . 338
3. Sequence search: the simplest way . . . 338
   3.1. Local alignments and pattern matching . . . 338
   3.2. All classes against one or more species . . . 341
   3.3. One class against all species . . . 341
   3.4. Additional applications for pattern matching . . . 341
4. Antimicrobial activity prediction: sequence order *versus* sequence size variation . . . 342
5. What can structure prediction reveal? . . . 344
6. Conclusions . . . 346
References . . . 347

1. Introduction

The explosive growth in database data requires techniques and tools to transform the amount of data into useful information and knowledge. Data mining, which is also referred to as knowledge discovery in databases, has clearly increased in importance. Mining information from databases has been recognized by many researchers as a hot spot in different areas (Chen et al., 1996; Porto et al., 2014b).

* Corresponding author at: Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia Universidade Católica de Brasília, Brasília, DF, Brazil.
E-mail address: ocfranco@pos.ucb.br (O.L. Franco).
URL: http://www.capb.com.br (O.L. Franco).

http://dx.doi.org/10.1016/j.biotechadv.2017.02.001
0734-9750/

In the post-genomic era, the growing number of sequences deposited in databases has been the key reason why databases are becoming a resource of novel biological information. Nowadays, the Universal Protein Resource (UniProt) (The UniProt Consortium, 2014) has 71,555,392 sequences deposited, split into 553,231 (0.8%) manually annotated from Swiss-Prot and 71,002,161 (99.2%) automatically annotated and not reviewed from TrEMBL (accessed in January, 2017). Therefore, the number of proteins from Swiss-Prot does not track the number of TrEMBL's, and a number of sequences are still annotated as hypothetical, unnamed or unknown sequences. In this context, the availability of complete genome sequences has given a new aim to modern biology: cataloguing all proteins that are responsible for every essential cellular function (Galperin and Koonin, 2012). Among the most diverse functions, the defence function against infectious microorganisms has gained attention. With the availability of the first genomes and transcript data (*e.g.* expression sequence tags, EST), the first reports of small cysteine-rich sequences resembling antimicrobial peptides (AMPs) were performed (Fedorova et al., 2002; Graham et al., 2004; Mergaert et al., 2003).

AMPs are evolutionarily ancient molecules that have been identified from diverse sources, such as microorganisms, plants and animals. They play an important role in the innate immune system and are the first line of defence to protect internal and external surfaces of the host - reviewed in Silva et al. (Silva et al., 2011). Overall, AMPs have an amphipathic and cationic structure, with a positive net charge between +3 and +9, and their size can range from 12 to 100 amino acid residues. Nowadays, there are two major classifications for AMPs, the first one being based on the structure (Silva et al., 2011) and the second one on the presence or absence of disulphide bonds (Brogden, 2005) (Fig. 1).

In addition to these classical classifications, in this review, AMPs are classified according to the origins of their amino acid sequence, being split into natural, encrypted[1] and designed AMPs. According to this, the natural AMPs are a product of specific genes (Dürr et al., 2006; Nguyen et al., 2013; Vlasak et al., 1983; Zasloff, 1987; Zhu et al., 2012); the encrypted AMPs are those that are encrypted in larger proteins, which are released after proteolytic cleavage (Brand et al., 2012; Okubo et al., 2012; Papareddy et al., 2010; Sigurdardottir et al., 2006; Wang et al., 2012); and the designed ones are artificial AMPs, developed through rational design techniques (Cardoso et al., 2016; Landon et al., 2008; Loose et al., 2006; Wieczorek et al., 2010) (Table 1).

In the last decade, the identification of AMPs from databases has gained attention as a branch of structural genomics and bioinformatics. Several approaches have been applied for the identification of AMPs from databases, including local alignments, regular expressions (REGEX), activity prediction by machine learning methods and also three-dimensional structure predictions (Fig. 2). These approaches have been useful to improve the annotation of such sequences and provide feedback from the databases (Fig. 2). For biotechnology, these approaches are a valuable tool, since they have the advantages of fast sequence identification and low costs, if compared to the peptide purification process, allowing the rapid discovery of promising novel antimicrobial agents. Therefore, this review is focused on approaches and new techniques for the identification of natural and encrypted AMPs, and also their advantages and limitations.

2. AMPs in databases

Before searching for a peptide in a database, it is important to know what to seek. Although the characteristics listed above are useful, they are extremely generic. Considering the more than 60 million sequences in TrEMBL, how many sequences will have a net charge between +3 and +9, 12 to 100 amino acid residues, stabilized or not by disulphide bridges? Therefore, some additional refinements are needed. These refinements can come from previous knowledge on AMP classes, arising from the literature itself, or from a set of sequences from an AMP database.

There are two main classes of AMP database: (i) the generalist, which holds any kind of AMP and (ii) the specific, which could be divided into two subclasses: (a) databases that hold a specific class of AMP (*e.g.* only defensins or cyclotides) and (b) databases that hold a supergroup of AMPs (*e.g.* only plant peptides or only cyclic peptides).

Unfortunately, a universal database with all AMP data does not exist yet, and therefore the information is split into several databases (Table 2). Recently, it was demonstrated that there is an overlap between AMP databases; however, each one has a degree of uniqueness, containing some exclusive sequences (Aguilera-Mendoza et al., 2015).

3. Sequence search: the simplest way

3.1. Local alignments and pattern matching

Sequence alignment is the main method for comparing biological sequences (Polyanovsky et al., 2011). Since the proteins are deposited in databases as their primary sequences, the most convenient way to search for similar sequences is to use local sequence alignments. For this purpose, the main tools are BLAST (Altschul et al., 1997) and FASTA (Pearson, 1990), with BLAST being the most popular for identifying AMPs.

The search for AMPs using local alignments has been performed through a number of iterations of local alignments: the first iteration uses a seed sequence for performing the initial search, and then in the second one, the sequences retrieved from the previous iteration are used as new seed sequences. This process is repeated until no new sequences are found. Additional filters could be used, such as the presence of signal peptide and/or some amino acid pattern.

Despite the efficacy of the local alignment strategy for finding novel AMPs in databases (Mulvenna et al., 2006; Zhu, 2008), this approach could fail to identify some sequences, if compared to pattern-matching approaches (Mulvenna et al., 2006; Porto et al., 2012c). There are two main strategies for searching for sequences by patterns: the use of profile Hidden Markov Models (profile-HMM) (Eddy, 1998; Silverstein et al., 2005, 2007) or regular expressions (REGEX) (Thompson, 1968; Mulvenna et al., 2006; Porto et al., 2012c). The two methods are quite similar. A REGEX is a precise and succinct description of a search pattern in text string format, providing a method for locating specific character strings embedded in character text (Thompson, 1968), where each REGEX position may be fixed, ambiguous or wild card; and when a protein sequence is described by a REGEX, it means that the REGEX matches the protein sequence (*i.e.* the REGEX "[WU]IL{1,2}IA[MN]" virtually matches any variation of the name "WILLIAM", starting with "W" or "U", indicated by "[WU]", with one or two L's, "L{1,2}" and finishing with "M" or "N", indicated by "[MN]", without variation in the other characters). A profile-HMM is a probabilistic model, which describes a multiple sequence alignment profile (Eddy, 1998), giving the probability distribution over a potentially infinite number of sequences. Since REGEX has no probabilities associated with its positions, it is less restrictive than profile-HMM.

The methodology of REGEX or profile-HMM for the identification of novel AMPs is similar. Overall, a set of homologous sequences must be aligned, and the alignment must be submitted to a specific program such as Pratt (Jonassen, 1997) or HMMER (Finn et al., 2011) for the identification of REGEX or profile-HMM, respectively. In addition, instead of building the pattern search, it could be retrieved from a database, such as Prosite for REGEX (Sigrist et al., 2013) or Pfam for profile-HMM (Finn et al., 2014). It is important to highlight that in the case of REGEX, it could be constructed and/or edited by hand, by using the amino acid physico-chemical properties, for example. Then, after building, selecting and/or editing a pattern, the search against the database is performed, and as well as through local alignments, the presence

[1] The term "encrypted" was used here instead of "cryptic" in order to keep the context of the manuscript by Brand and co-workers (Brand et al., 2012).