

# HyPep: A New Strategy to Accelerate Peptide Discovery with A Combination of *De Novo* Sequencing and Homology Database Search

Weifeng Cao<sup>1</sup>, Mingming Ma<sup>2</sup>, Qiang Fu<sup>1</sup>, Lingjun Li<sup>1,2</sup>

Department of Chemistry<sup>1</sup> & School of Pharmacy<sup>2</sup>, University of Wisconsin, Madison, WI 53705



## Overview

### Purpose

To improve the efficiency and accuracy of peptide discovery for the species without genomic information.

### Methods

A combination of automatic *de novo* sequencing and homology database search is employed to process tandem MS data. An efficient homology algorithm is developed to search query sequences against a home-built peptide database. A list of possible peptide sequences is obtained and verified by manual *de novo* sequencing.

### Results

A local neuropeptide database containing 5825 entries was constructed. A PO extract of *Cancer borealis* was analyzed to identify 17 putative peptides within several minutes, in which 4 peptides were verified by manual *de novo* sequencing.

## Introduction

The discovery of new peptides from organisms without genomic information remains a challenge to date. *De novo* peptide sequencing has been the standard method to solve this problem. However, due to the very low efficiency with manual *de novo* sequencing, often times people turn to automatic *de novo* sequencing algorithms to perform peptide sequencing from tandem mass spectrometry data. Although most of the peptide sequencing software offers fast speed, many problems still exist such as poor accuracy and more redundant information. In this study, we aim to develop a new strategy that combines automatic *de novo* sequencing algorithm and in-house database searching for accelerated neuropeptide discovery from decapod crustacean model organisms without sequenced genome.

## Methods

- ❖ The *Cancer borealis* pericardial organs (PO) was fractionated firstly by HPLC and then introduced into nanoLC-ESI-QTOF.
- ❖ The raw spectra were converted to .pkl formatted files with ProteinLynx.

## Methods

- ❖ The .pkl files were interpreted by Peaks.
- ❖ A local neuropeptide database was constructed specifically for crustacean species.
- ❖ An in-house program HyPep was developed and employed to do homology search for any query sequence against the local database.
- ❖ The query sequences would be evaluated by HyPep to determine if they are possible peptides.
- ❖ The possible peptide sequences would be verified by manual *de novo* sequencing

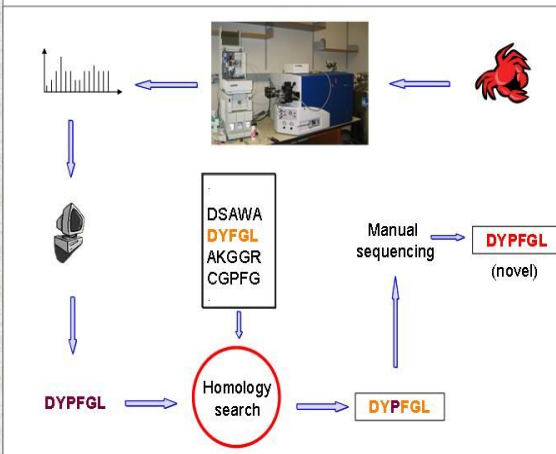


Figure 1. General framework for peptide discovery

## Results

### Local neuropeptide database

- ❖ Sources: 1) public databases like NCBI etc.  
2) publications  
3) the work from our group
- ❖ Peptides are grouped based on neuropeptide families or species.
- ❖ A total of 5825 neuropeptides
- ❖ FASTA formatted and Access formatted.

### HyPep algorithm

- ❖ Model description:
 
$$\begin{matrix} \forall & a_1 a_2 \dots a_n & (\text{query sequence}) \\ & b_1 b_2 \dots b_m & (\text{database sequence}) \end{matrix}$$
- Find maximum match

## Results

### Homology search algorithm:

- Connect to peptide database, which contains three tables: peptide, neuro and candidate
- Delete all sequences from candidate table
- For each sequence mySeq in the neuro table
  - For each record dbSeq in the peptide table
    - score = 0
    - weight = length(mySeq) + |length(mySeq) - length(dbSeq)|
    - score1 = forwardVariableSearch(mySeq, dbSeq)/weight
    - score2 = backwardVariableSearch(mySeq, dbSeq)/weight
    - score3 = forwardFixedSearch(mySeq, dbSeq)/weight
    - score4 = backwardFixedSearch(mySeq, dbSeq)/weight
    - score = score1 + score2 + score3 + score4
    - If (score > threshold) Insert dbSeq into the candidate table
- Endfor
- Endfor
- Sort the candidate table by scores

Figure 2. HyPep algorithm

### Performance test

- ❖ Conditions:
  - Instrument: nanoLC-ESI-QTOF (Waters)
  - Pkl converter: ProteinLynx (Waters)
  - Automatic *de novo*: Peaks 4.5 (BSI)
  - Database: in-house neuropeptide database
  - Homology search: HyPep, BLAST
  - Score threshold: 50% (HyPep)
- ❖ Standard sample: 1) Peptide mixture (6 peptides)  
2) BSA digest (tryptic)

	Peptide Mixture	BSA digest
No. of ID's	6	40

- ❖ Biological Sample: PO extract (one HPLC fraction)

