Data-Driven Design of Diagnostic Kits and Therapeutic Peptides

Emmanuel O Salawu[1,2,¶], Cheng-Yu Tsai[1,3,¶], Hongchun Li[1,5,¶], Guan-Yu Lin[4,¶], Ting-Yu Kuo[4,¶], Adarsh Sharma[1], Kai-Di Hu[1], Liyin Voon[1], Yu-Lin Lu[1], Ximai Ke[1],Chung-Yu Lan[4,6]*, Hua-Wen Fu[4,6]*, Lee-Wei Yang[1,2,6,7,]*

[1]Institute of Bioinformatics and Structural Biology, National Tsing-Hua University, Hsinchu 30013, Taiwan, [2]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Sciences, Academia Sinica, Taipei 11529, Taiwan, [3]Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, Taipei 10055, Taiwan, [4]Institute of Molecular and Cellular Biology, Department of Life Science, National Tsing Hua University, Hsinchu 30013, Taiwan. [5]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, USA, [6]Department of Life Science, National Tsing Hua University, Hsinchu 30013, Taiwan. [7]Physics Division, National Center for Theoretical Sciences, Hsinchu 30013, Taiwan

¶ Authors who have equal contributions

* To whom correspondence should be addressed: Lee-Wei Yang. Tel: +88635742467; Fax +88635715934; Email:lwyang@life.nthu.edu.tw; Hua-Wen Fu. Tel: +88635742485. Fax: +88635715934 Email: hwfu@life.nthu.edu.tw, and Chung-Yu Lan. Tel: +88635742473; Email: cylan@life.nthu.edu.tw

ABSTRACT

Learning from experimentally determined interacting secondary structural motifs, we compiled a database to facilitate a data-driven design of therapeutic peptides (TPs). 1.7 million helical peptides (HPs) in >130 thousand proteins are extracted along with their interacting partners from the protein data bank (PDB). The sequences of the HPs are developed into a searchable database (TP-DB) by creating indices that map specific peptide patterns to locations of matched HPs in the TP-DB. Leveraging TP-DB to search for a potent membrane-insertion pattern WXXWXXW, established by our microsecond-long MD simulations, we found a positively charged HP that matches the pattern has a commensurate minimal inhibitory concentration (MIC) against *Candida albicans* (fungus) as compared to previously characterized homologs. With identifying peptides containing the affinity determinant motifs DYKXX[DE] of FLAG-tag within pathogenic proteins, which PHI-BLAST failed to find, we successfully discovered *Helicobacter pylori* neutrophil-activating protein (HP-NAP), a virulence factor of *H. pylori*, to contain a stretch of sequence DYKYLE that can be recognized by the anti-FLAG M2 antibody. By doing so, we repurposed a purification-tag-specific antibody into a diagnostic kit for pathogen's virulence factors. Taken together, we believe that TP-DB and its pattern-based search engine provide a new opportunity for a (secondary-)structure-based design of peptide drugs and diagnostic kits for pathogens without inferring evolutionary homology between sequences sharing the same pattern. TP-DB is made available at http://dyn.life.nthu.edu.tw/design/

1. Introduction

“Nature is a tinkerer, not an inventor" [1]. Similar sequences and structural motifs are used across kingdoms for important functional purposes that are subject to physical laws and to evolved chemical dependency for life. One enticing application after learning so much from the now >140,000 PDB structures and the functions they encode is to reassemble them into therapeutics for medicinal purposes.

A number of efforts have been laid to establish fragment-based design of drug leads [2,3] and associated databases/software [4]. However, there has not been a systematic approach to assemble structurally resolved protein fragments for the design of therapeutics. Preferably, given a known functional motif, say, a motif that can be recognized by an antibody or a sequence pattern essential to bactericidal activity, whether there is a proper search engine to locate all the matched structural fragments reporting relevant analytics in a timely fashion? To address such a need and showcase protein-fragment-based therapeutic design, we introduce the Therapeutic Peptide Design dataBase (TP-DB).

In this seminal work, we extracted ~1.7 million (1,676,119) structurally resolved helices and their contacting/interacting helical partners from the protein data bank (PDB), from which helical propensity and coordination (contact) number are derived (**Table 1**). We then establish a pattern-specific search engine to find sequences bearing particular spatio-chemical properties instead of finding evolutionary homologs that is the purpose of Pattern Hit Initiated BLAST (PHI-BLAST) [5]. PHI-BLAST requires the input of both pattern and the template sequence. As a result, the PHI-BLAST does leave out the sequences that meet the queried pattern but are not indicative of evolutionary homology, which differs from the functionality of TP-DB. Here, we are able to use TP-DB to develop a new series of anti-fungal/antimicrobial peptides, and convert the antibody against a purification tag into a diagnostic reporter to detect *Helicobacter pylori* infection. Methods and results are reported below.

2. Methods

2.1. Extraction of Helical Peptide Sequences from the Protein Data Bank (PDB)

To obtain the amino acid sequences that fold into helices in nature, we processed the PDB [6] files of 130,000+ experimentally determined protein structures and extracted secondary structure information from the header part of the PDB files. This allowed us to obtain the peptide sequences corresponding to the helices in each of the proteins. In addition, for a given helix, its contacting neighbors in 3D space, the adjacent helical peptides that are <4 Å away (per heavy atoms), are also documented.

2.2. Creation of TP-DB

The collected peptides sequences are carefully developed into a searchable database by creating indexes that map peptide patterns (such as Y***G**K, which is equivalent to “Y 3 G 2 K”) into where they could be found in structurally solved proteins. "Where" a given peptide pattern can be found is defined by the PDB ID, the chain ID, and the index of the position where the pattern begins in that chain of the PDB. For easy computability and high flexibility of expression without losing accuracy, a given pattern (Y***G**K) is represented by a key (such as "Y 3 G 2 K") which is a combination of alphabets and numbers. The keys serve as the indexes for the database. For example, the key "Y 3 G 2 K" points to all the locations where a "Y" could be found such that the fourth (i.e. 3 + 1) amino acid downstream to it is a "G" and the third (i.e. 2 + 1) amino acid after "G" is a "K".

For each of the $peptide_p$ in the TP-DB (**Fig. 1A**), we scan the sequence of $peptide_p$ and generate its keys containing three amino acids (which are henceforth referred to as anchors) at a time (**Fig. 1B**). The three anchors do not have to be consecutive amino acids in the sequence. By design, we allow zero to four amino acids in between the first anchor and the second anchor, and between the second anchor and the third anchor. We represent each key obtained from the peptide as “CmDnE” such that C, D, and E are the one letter codes of amino acids and m, n are numbers of spacings. Therefore, $peptide_p$ with sequence ADEKKFWGKYLYEVA has keys that range from A0D0E, D0E0K, …, E0V0A, A0D1K, …, K4Y4A as shown in **Fig. 1B**. While scanning the sequence of $peptide_p$ for its keys, we extract the start position of each key at the same time. The keys and their start positions (values) are indexed to make up the database (**Fig. 1C**). For a given key-value pair in the database, the value is an array of peptide identifiers and all the start positions of that key in each of the peptides (**Fig. 1C**). The building of this database index (made up by the key-value pairs) is the database creation itself. Find TP-DB at http://dyn.life.nthu.edu.tw/design/

A

```
Peptide P:    ADEKKFWGKYLYEVA        Peptide Z:    TAFEGGILKKGHHCSYTKH
Positions:    012345678901234        Positions:    012345678901234567 8
```

B

```
Keys in Peptide P                    |   Start Positions in Peptide P
A0D0E, D0E0K, ..., Y0E0V, E0V0A      |   0, 1, ..., 11, 12
A0D1K, D0E1K, ..., L0Y1V, Y0E1A      |   0, 1, ..., 10, 11
...                                  |   ...
A0D4W, D0E4G, ..., G0K4V, K0Y4A      |   0, 1, ..., 7, 8
...                                  |   ...
A4F0W, D4W0G, ..., G4E0V, K4V0A      |   0, 1, ..., 7, 8
...                                  |   ...
A4F4L, D4W4Y, ..., K4K4V, K4Y4A      |   0, 1, ..., 3, 4
```

C

```
DB_INDEX = { A0D0E: [ [Peptide P , 0], [Peptide P+1 , 17, 21, ...], ...],
             D0E0K: [ [Peptide P , 1], [Peptide P+9 , 13, 25, ...], ...],
             ...
             K4V0A: [ [Peptide P , 8], [Peptide P+78 , 1, 64, ...], ...],
             ...
             K4Y4A: [ [Peptide P , 4], [Peptide P+2 , 23, 41, ...], ...],
             ... }
```

D

Sample patterns of interest	Keys in the DB	Examples of peptides that the patterns can match
ADE	A0D0E	ADEKKFWGKYLYEVA
DEK	D0E0K	ADEKKFWGKYLYEVA
K----VA	K4V0A	ADEKKFWGKYLYEVA
K----Y----A	K4Y4A	ADEKKFWGKYLYEVA

E

```
Simple Query: K 4 Y 4 A, which is equivalent to K4Y4A or K----Y----A
Direct Fetch: K4Y4A from DB_INDEX in the RAM: →
                [ [Peptide P , 4], [Peptide P+2 , 23, 41, ...], ...],
```

F

[1] Complex Query: Process into a Combination of Simple Queries

Example of Complex Queries	Matching Patterns	Constituent Simple Queries
A 3 G 2 K 3 H 4 K	A 3 G 2 K 3 H 4 K	A 3 G 2 K K 3 H 4 K
A/Y 3 G 2 K 3 H 4 K	A 3 G 2 K 3 H 4 K	A 3 G 2 K K 3 H 4 K
	Y 3 G 2 K 3 H 4 K	Y 3 G 2 K K 3 H 4 K
A 2,3 G 2 K 3,4 H 4 K	A 2 G 2 K 3 H 4 K	A 2 G 2 K K 3 H 4 K
	A 2 G 2 K 4 H 4 K	A 2 G 2 K K 4 H 4 K
	A 3 G 2 K 3 H 4 K	A 3 G 2 K K 3 H 4 K
	A 3 G 2 K 4 H 4 K	A 3 G 2 K K 4 H 4 K

[2] Simple Queries' results are aggregated into the required results

Fig. 1. Creation and Querying of TP-DB. (**A**) Depiction of $peptide_P$ and $peptide_Z$, and the positions of their amino acids. (**B**) Examples of the possible keys in $peptide_P$. (**C**) The Database's index contains keys which point to the information regarding where each of the keys could be found. The information includes peptide identifiers, and the start positions of each key in each of the peptides that contains that key. (**D**) The

patterns of interest can be translated into formats that are similar to those of the keys in the database. (**E**) Results for simple queries can be directly fetched from the database without extended preprocessing of the query. (**F**) A complex query can be systematically broken down into joint results of simple queries.

2.3. Querying the TP-DB

The design of the database and its indexes make it easy to query the database even when the patterns of interest are not simple. To query the database, the user specifies at least three anchor amino acids and the number of amino acids between the anchors, such that "ADE", "K----VA", and "K----Y----A" could be queried using "A 0 D 0 E", "K 4 V 0 A", and "K 4 Y 4 A" respectively (**Fig. 1D**). For such simple queries (such as "K 4 Y 4 A"), the results are fetched directly from the database's index that resides in the server's random-access memory (**Fig. 1E**). Furthermore, the design of the indexes of the database makes possible its efficient querying even when one needs to search for non-trivial patterns, which is discussed next.

We define a non-trivial pattern as a pattern that is not directly a key in the TP-DB but which could be pre-processed into a combination of simple patterns and subsequently into simple queries that could be directly found in the database. Therefore, when a non-trivial pattern is queried, we break it down until we reach its components that correspond to keys that could possibly be found in the database as illustrated with the examples in **Fig. 1F**. For instance, "A/Y 3 G 2 K 3 H 4 K" is broken down into a combination of two sub-queries/machining patterns "A 3 G 2 K 3 H 4 K" and "Y 3 G 2 K 3 H 4 K" and the five anchors in each of the machining patterns are treated as a combination for two keys each with three anchors such that the third anchor of the first key is the same as the first anchor of the second key as shown in **Fig. 1F**. We then query the database for the keys. A systematic combination of the results from all the keys (while taking into account the regions where the keys overlap) makes it possible to construct the needed results for the non-trivial query.

2.4. The TP-DB

To make the developed database easily accessible to protein scientists, chemists and biochemists, we have developed a web server interface on top of the developed TP-DB. The web server accepts queries in the format describe above and the results of the queries are processed in real-time. A simple html interface to the web server is available at http://dyn.life.nthu.edu.tw/design/ for the pattern search, and at https://dyn.life.nthu.edu.tw/design#helixblast for the BLAST search (parameterized for short sequences). The use of the latter is beyond the scope of this paper.

2.5. Physiochemical Properties and Ranking of the Helical Peptides in TP-DB

The helical peptides returned by a query is ranked by their helical propensity scores and then by their contact scores. The helical propensity score for a given peptide sequence matched is computed by summing up the "HPnNA" of the amino acids that make up the sequence, where "HPnNA" is the "Helical Propensity score of an amino acid normalized by the Natural Abundance of that amino acid" as shown in **Table 1**. The contact score is the average number of amino acids contacting each of the amino acids in the helix within a 15Å range. Sample outputs of the web server (with the helical peptides ranked accordingly) are shown in the **Table 2**.

2.6. General MD Simulations Protocol

All MD simulations were performed using NAMD package 2.9 [7] in NPT ensemble with periodic boundary conditions. Latest CHARMM forcefields [8,9] were employed for the peptide, water and lipids, respectively. All simulations were performed at time step of 2 fs and trajectories were recorded every 5000 steps (i.e. at every 10 ps). RATTLE and SETTLE algorithms are applied to constrain hydrogen atoms in peptides and waters. Cutoff of 12 Å with switch distance 10 Å and pair list distance 14 Å are applied when calculating non-bonded interactions. With periodic boundary conditions, the Particle Mesh Ewald method [10] was employed for calculations of long-range electrostatic interactions. Temperature was maintained at 310 K using Langevin dynamics [11] and pressure was controlled at 1 ATM using Nosé-Hoover Langevin piston [12].

2.7. Lipid Preparation and Equilibrium *in silico*

Zwitterionic membrane with 78 POPC (termed as “PC” hereafter) lipids was used to mimic eukaryotic membrane [13,14]. The AMPs and lipid bilayers were solvated in explicit TIP3 water molecules [15]. Na^+ and Cl^- ions were added to neutralize the system. The initial area per lipid of PC was prepared to match their experimental values measured at 303 K, which is 68.3 ± 1.5 $Å^2$ [16]. The CHARMM-GUI server was employed to build these lipid-water-ion systems [17].

The lipid-water-ion systems were equilibrated at 310 K using an NPT ensemble with gradually reduced restraints for 10 ns, followed by a 100 ns equilibration without restraints. The thickness of the membranes

was the difference in the z direction between the center of mass (COM) of the P atoms of the upper and lower leaflets of the lipid bilayers. In the last 50 ns, snapshots whose thicknesses were the same as the average thickness were selected as the equilibrated lipid-water-ion system to which the AMPs were loaded. The AMP-lipid-water system was neutralized with Na^+/Cl^- ions and equilibrated in NVT and NPT ensembles for a total of 3.6 ns. After equilibration, the AMP-lipid-water-ion systems were simulated for 200 ns at 310 K using an NPT ensemble.

2.8. Free diffusion simulations for AMPs in the presence or absence of lipids and soft boundary condition

NMR spectra-solved α-helical conformations and fully extended (random-coil) AMPs were simulated together with PC membrane in a water box containing 0.02 M NaCl. For the fully extended AMPs, they were first simulated with an NPT ensemble for 50 ns in the absence of lipids. The resulting snapshots of the AMPs were clustered into 5 groups based on their structural similarity with a cutoff RMSD of 0.1 Å using the "clustering" plug-in of the VMD software [18]. A representative conformation from the biggest cluster was selected as the initial conformation for an AMP before it was loaded with the equilibrated membrane for the freely diffused insertion simulation.

2.9. Production of recombinant *Helicobactor* pylori neutrophil-activating protein (HP-NAP) and maltose-binding protein (MBP)

Recombinant *H. pylori* neutrophil-activating protein (HP-NAP) was expressed in *E. coli* BL21(DE3) cells harboring the expression plasmid pET42a-NAP and purified by either two consecutive gel-filtration chromatography as previously described [19] or a small-scale DEAE Sephadex negative mode batch chromatography as previously described [20]. Maltose-binding protein (MBP) was prepared the same as the procedure for production of MBP fused with the polypeptide containing residues Arg77 to Glu116 of HP-NAP as described below except that *E. coli* BL21(DE3) cells harboring the pMALc2g expression vector was used for expression.

2.10. Cloning of HP-NAP$_{R77\text{-}E116}$ into a MBP fusion protein expression vector

The plasmid DNA pET42a-NAP encoding a *napA* gene from *H. pylori* strain 26695 [GenBank:AE000543.1, Gene: HP0243] was prepared as previously described [19]. The DNA fragment coding for polypeptide containing residues Arg77 to Glu116 of HP-NAP (HP-NAP$_{R77-E116}$), which contains the D-Y-K-x-x-[DE] motif, was amplified by PCR from the plasmid pET42a-NAP using the forward and reversed primers containing BamHI and HindIII site, respectively. The forward primer is 5'-ATAAGGATCCCGTGTTAAAGAAGAAACTAAAAC-3' and the reversed primer is 5'-TTAATAAGCTTTAATTCTTTTTCAGCGGTGTTAGAG-3'. The PCR reaction was carried out with 10 ng plasmid DNA pET42a-NAP as a template and KAPA HiFi PCR Kit (Kapa Biosystems, Inc.) in a Mastercycler Gradient 5331 (Eppendorf, Germany). An initial denaturing phase of 95 °C for 5 min was followed by 39 cycles of 98 °C for 20 sec, 67 °C for 15 sec, and 72 °C for 15 sec. A final elongation phase of 72 °C for 2 min was also included. The amplified DNA fragments encoding HP-NAP$_{R77-E116}$ were then cloned into pJET1.2/blunt vectors using the CloneJET PCR Cloning Kit (Thermo Fisher Scientific Inc.). The resulting plasmid was designated as pJET1.2/blunt-HP-NAP$_{R77-E116}$. The insert was sequenced to confirm the correct DNA sequence. The correct insert was digested from pJET1.2/blunt-HP-NAP$_{R77-E116}$ with BamHI and HindIII and then cloned into the pMALc2g expression vector [21]. The resulting plasmid was designated as pMALc2g-HP-NAP$_{R77-E116}$.

2.11. Production of MBP-tagged HP-NAP$_{R77-E116}$

E. coli BL21(DE3) cells harboring pMALc2g-HP-NAP$_{R77-E116}$ were streaked on a lysogeny broth (LB) agar plate containing 100 μg/ml ampicillin and incubated at 37 °C for 16 hr. A single colony was picked and inoculated into 4 ml of LB containing 100 μg/ml ampicillin and the culture was incubated at 37 °C with shaking at 170 rpm for 16 hr. A volume of 2 ml of the overnight culture was inoculated into 200 ml LB containing 100 μg/ml ampicillin and the inoculated culture was incubated at 37 °C with shaking at 170 rpm for 2 hr until the OD_{600} reached 0.5. The expression of MBP-tagged HP-NA$_{PR77-E116}$ was induced by the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.3 mM and the culture was incubated at 37 °C with shaking at 180 rpm for 3 h until the OD_{600} reached 1.7. Then, the cells were centrifuged at 6,000 x g at 4 °C for 15 minutes to remove the supernatant and the cell pellets were stored at -70 °C.

The cell pellet from a 200 ml culture of *E. coli* expressing recombinant MBP-tagged HP-NAP$_{R77-E116}$ were re-suspended in 20 ml of ice-cold buffer containing 20 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM ethylenediaminetetraacetic acid (EDTA), and 1 mM dithiothreitol (DTT), plus 0.1% (v/v) protease inhibitor mixture (PI mix). The PI mix contained 0.13 M phenylmethylsulfonyl fluoride (PMSF), 0.03 M N-alpha-

tosyl-L-lysyl-chloromethyl ketone (TLCK), and 0.03 M N-tosyl-L-phenylalanyl-chloromethyl ketone (TPCK). The bacterial suspensions were disrupted by Emulsiflex C3 high-pressure homogenizer (Avestin) operated at a range of 15,000-20,000 psi for 7 times at 4 °C. The lysates were centrifuged at 30,000 x g at 4 °C for 1 hr to separate insoluble and soluble proteins by using a Hitachi himac CP80WX ultracentrifuge (Hitachi Koki Co. Ltd., Tokyo, Japan). Then, 5 mL supernatant containing the soluble proteins were loaded onto a 1-ml MBPTrap HP column (GE Healthcare Bio-Sciences), which was pre-equilibrated with 20 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM EDTA, and 1 mM DTT, at a flow rate of 0.5 ml/min at 4 °C by ÄKTA Purifier. The column was eluted with 20 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM EDTA, 1 mM DTT, and 10 mM maltose at a flow rate of 1 ml/min at 4 °C by ÄKTA Purifier. The flow-through and elution fractions were analyzed by SDS-PAGE on a 12% gel. The elution fractions containing recombinant MBP-tagged HP-NAP$_{R77\text{-}E116}$ were collected and concentrated to a concentration higher than 1 mg/ml.

2.12. Recombinant HP-NAP-based enzyme linked immunosorbent assay (ELISA)

Nunc MaxiSorp ninety-six-well enzyme linked immunosorbent assay (ELISA) plates (Nunc, Rochester, NY, USA) were coated with 0.3 μg of recombinant HP-NAP in 100 μl of bicarbonate buffer (pH 9.0) for each well at room temperature for 16 hr. Each well was washed with 300 μl of phosphate buffered saline (PBS), pH 7.4, containing 20 mM Na_2HPO_4, 1.47 mM KH_2PO_4, 137 mM NaCl, and 2.7 mM KCl, with the addition of 0.1% tween-20 (PBS-T) three times for 10 min each time. The wells were blocked with 250 μl PBS with 1% bovine serum albumin (BSA) for 2 hr and then washed with 300 μl of PBS-T buffer three times for 10 min each time. The anti-FLAG M2 antibody (Sigma-Aldrich, Cat# F-3165) and its corresponding mouse IgG antibody (Sigma-Aldrich, Cat# I5381) at a concentration of 600 ng/ml and the hybridoma culture supernatant containing mouse monoclonal antibody MAb 16F4 [22] against HP-NAP at a dilution of 1:5000 in 100 μl of PBS-T buffer containing 1% BSA were added into each well. The plate was incubated at room temperature for 1 hr and then the wells were washed with 300 μl of PBS-T buffer three times for 10 min each time. The horseradish peroxidase-conjugated goat anti-mouse secondary antibody (Jackson ImmunoResearch) at a dilution of 1:10000 in 100 μl of PBS-T buffer containing 1% BSA was loaded into each well. The plate was incubated at room temperature for 1 hr and then the wells were washed with 300 μl of PBS-T buffer three times for 10 min each time. The color was developed using 3,3',5,5'-tetramethylbenzidine (TMB) peroxidase substrate (Thermo Scientific). The reaction was terminated by the addition of 2 N H_2SO_4, and the absorbance at 450 nm was measured by Bio-rad iMark microplate absorbance reader (Hercules, CA).

2.13. Western blot analysis

Western blotting was performed essentially the same as previously described [23]. The membrane was probed with either anti-FLAG M2 antibody (Sigma-Aldrich) at a concentration of 1 μg/ml or the hybridoma culture supernatant containing mouse monoclonal antibody MAb 16F4 [22] against HP-NAP at a dilution of 1:2000.

2.14. Determination of the Minimum Inhibitory Concentration (MIC)

To measure antimicrobial activity of peptides, minimum inhibitory concentrations (MIC) are determined by broth-dilution assay as described in Clinical and Laboratory Standards Institute (CLSI) documents M27-A3 for yeasts [24] and M07-A9 for bacteria [25] with some modifications. Briefly, *E. coli* ATCC 25922 and *C. albicans* SC5314 are recovered from frozen stock by growing overnight at 37°C on LB (Luria-Bertani) and YPD (Yeast Extract-Peptone-Dextrose) agar plates, separately. A single colony of *E. coli* is inoculated into 5 ml LB medium and grown overnight at 37°C with shaking at 220 rpm. Cells are then diluted into MHB (Mueller–Hinton broth) at 1:100, and grown until the optical density at 600 nm (OD_{600}) achieves 0.4 to 0.8. For *C. albicans* SC5314, a single colony was inoculated in YPD broth and grown overnight at 30℃ with shaking at 180 rpm. Cells from the overnight culture are subcultured into fresh YPD and grown to mid-log phase. Finally, *E. coli* and *C. albicans* cells obtained from above-described are diluted into MHB and LYM, a modified RPMI 1640 medium [26] to a concentration of ~5×10^5 cells/ml and ~4×10^4 cells/ml, respectively. To determine the MIC values, one hundred microliters of cells was placed in each well of a clear bottom 96-well microplate, followed by independently adding 100 μl peptides with different concentrations (in MHB or LYM). The final concentrations in the wells are 120, 60, 30, 15, 7.5, 3.75 and 1.88 μg/ml for peptides, ~ 2.5×10^5 cells/ml for *E. coli* and ~2×10^4 cells/ml for *C. albicans*. The plate was incubated at 37°C with shaking at 220 rpm for 18 h (*E. coli*) and at 180 rpm for 24 h (*C. albicans*). The OD_{600} values are measured using a iMark™ Microplate Absorbance Reader (Bio-Rad). The MIC_{90} value is determined by 90% reduction in growth compared with that of the peptide-free growth control. Experiments are performed in pentaplicate, and MIC_{90} is determined as the majority value out of the 5 repeats.

3. Results and Discussions

3.1. Database Features and Statistics

Overall, we extracted 1,676,119 helical peptides from the experimentally solved protein structures in the protein data bank (PDB). The helical peptides contain 24,301,682 amino acids in total. The probability distribution of different amino acid type in the helices are shown in the second column (HP: Helical Propensity) of **Table 1**. Similarly, relative natural abundance (NA) of the amino acids are shown in the third column of **Table 1**. The relative natural abundance of the amino acids was obtained from Expasy (http://web.expasy.org/docs/relnotes/relstat.html). The helical propensity normalized by the natural abundance (HPnNA) and its natural logarithm are shown in the fourth and the fifth column of **Table 1**. The log (HPnNA) is used as one of the metrics for ranking the helical peptides as described in the methods section above.

Table 1. Amino acids sorted by their helical propensity

AA*	HP	NA	HPnNA	log (HPnNA)
A	0.109	0.083	1.320	0.277
W	0.014	0.011	1.284	0.250
E	0.084	0.067	1.246	0.220
L	0.115	0.097	1.192	0.175
Q	0.045	0.039	1.145	0.135
Y	0.033	0.029	1.130	0.122
K	0.064	0.058	1.100	0.095
M	0.026	0.024	1.079	0.076
R	0.059	0.055	1.067	0.065
F	0.038	0.039	0.984	-0.016
I	0.058	0.059	0.978	-0.022
H	0.022	0.023	0.969	-0.031
D	0.052	0.055	0.952	-0.049
N	0.038	0.041	0.936	-0.066
T	0.049	0.053	0.918	-0.086
V	0.062	0.069	0.902	-0.103
C	0.012	0.014	0.876	-0.132
S	0.053	0.066	0.804	-0.218
G	0.046	0.071	0.650	-0.431
P	0.021	0.047	0.445	-0.810

* AA = Amino Acid; HP = Helical Propensity; NA = Natural Abundance; HPnNA = Helical Propensity normalize by Natural Abundance

3.2. Anti-FLAG M2 antibody that recognizes a purification-tag is repurposed into a potential diagnostic tool for human pathogens, facilitated by TP-DB but not PHI-BLAST

To find specific pattern in a biological sequence while referring homology, Pattern Hit Initiated (PHI)-BLAST has been the main (if not only) bioinformatics tool to perform so [5]. PHI-BLAST was designed to address evolutionary relevance of two proteins carrying the same or similar motifs, so it renders results considering both the pattern match and local sequence similarity with statistically significance. To showcase the prowess and disparate utility of the TP-DB, we demonstrate below how a purification tag (herein FLAG-tag) and its antibody (herein M2 antibody) can be repurposed into a diagnostic kit for detecting human pathogens. FLAG-tag is known to have a sequence DYKDDDDK [27] with the pattern D-Y-K-x-x-[DE] being experimentally confirmed as the main affinity determinant motif [28]. "DYK" here is the strongest determinant while the last D (or E) is of secondary importance. We would like to search TP-DB for proteins that (i) are not included in the PHI-BLAST search results and (ii) belong to proteins in human pathogens but not in non-pathogenic bacteria.

3.2.1. Confirmed by ELISA and Western Blot, HP-1 is Recognized by anti-FLAG M2 Antibody Specifically at a stretch containing the D-Y-K-x-x-[DE] motif

When querying the pattern "D 0 Y 0 K 2 D/E" against the TP-DB, we can find totally 93 sequences with unique 19 sequences. While using PHI-BLAST to search NCBI's non-redundant protein sequences (nr) with the pattern D-Y-K-x-x-[DE] (in PROSITE format) and FLAG sequence "DYKDDDDK", we could not find any sequence with an E-value less than 100. On the other hand, among the 19 TP-DB-identified sequences, HP-NAP, containing the sequence "**DYK**YL**E**", can be found in *Helicobacter pylori* strain 26695 (*accession no.* AE000543, ATCC) but not in regular non-pathogenic bacteria. If the anti-FLAG M2 antibody can recognize this randomly selected pathogenic protein for its containing a FLAG-determinant motif (herein D-Y-K-x-x-[DE]), we can potentially use the same concept to repurpose the original use (say, purification) of any known pair of antibody and antigen recognition motif (say, M2 antibody and D-Y-K-x-x-[DE]) into a new use, herein the potential diagnostic kit for human pathogens. To our heartfelt delight, the anti-FLAG M2 antibody was indeed found to be capable of recognizing HP-NAP as analyzed by ELISA (**Fig. 2A** *left*) and Western blot (**Fig. 2A** *right*).

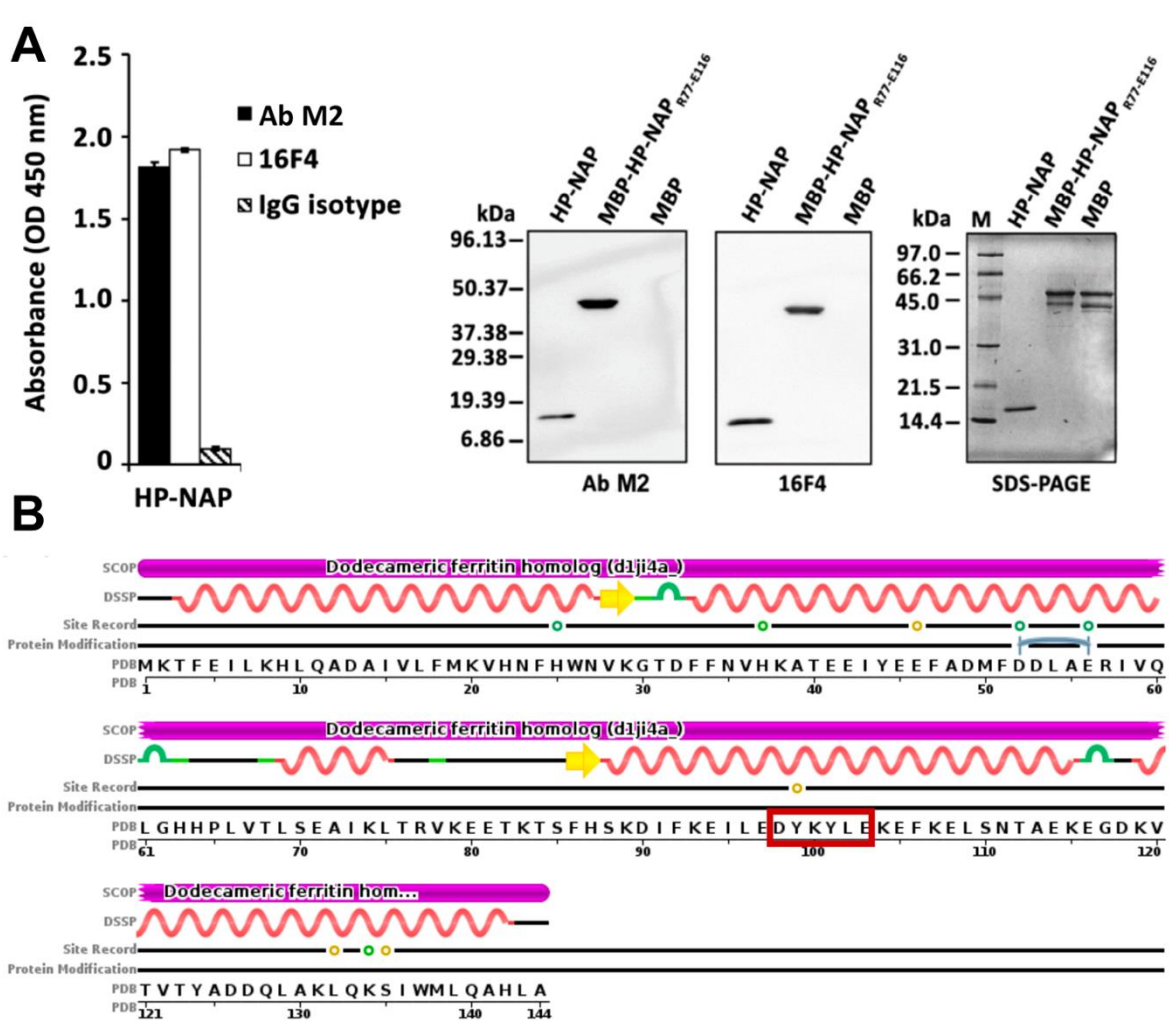

Fig. 2. Detection of HP-NAP by anti-FLAG M2 antibody. (**A**) Detection of HP-NAP by anti-FLAG M2 antibody using recombinant HP-NAP-based ELISA. Recombinant HP-NAP was purified by a small-scale DEAE Sephadex negative mode batch chromatography. Purified HP-NAP as an antigen was coated on an ELISA plate as described in Methods. The anti-FLAG M2 antibody, its corresponding mouse IgG antibody, and the antibody 16F4 against HP-NAP, as positive control, were subjected to recombinant HP-NAP-based ELISA. The result is presented as absorbance at $OD_{450\ nm}$. Data were presented as the mean ± S.D. of one experiment in duplicate. Similar results were obtained in three independent experiments. (**B**) Detection of recombinant HP-NAP by anti-FLAG M2 antibody using Western blot analysis. Recombinant HP-NAP was purified by two consecutive gel-filtration chromatography. MBP-tagged HP-$NAP_{R77\text{-}E116}$ and MBP were partially purified by the MBPTrap HP column using ÄKTA Purifier. These purified proteins, 1 μg each, were subjected to SDS-PAGE on a 15 % gel and Western blot analysis with anti-FLAG M2 antibody and the antibody 16F4. Molecular masses (M) in kDa are indicated in the left of the blots and the gel. Similar results were obtained in two independent experiments. **(B)** The sequence of HP-NAP. Note that "DYKYLE", highlighted in the red box, matches the pattern D-Y-K-x-x-[DE] and is included in the R77-E116 stretch. The image is download from PDB (code:1JI4; [6]) with minor modifications.

As a result, the antibody (herein M2 antibody) of a purification kit can be *repurposed into a diagnostic kit* for detecting human pathogens, herein *H. pylori*, through recognizing native HP-NAP by ELISA or HP-NAP monomer as a 17 kDa band by Western blot.

3.3. α-Helical Antimicrobial Peptide (AMP) Design

To showcase the good use of TP-DB, we first examined in details the insertion processes of two previously reported helical antimicrobial peptides (AMPs), W3_p1 (sequence: KK**W**RK**W**LK**W**LAKK; [29]) and W3_p2 (sequence: KK**W**LK**W**LK**W**LKKK; [30]) by MD simulations. The two amphiphilic AMPs have their hydrophobic side containing 3 equally spaced tryptophans, while positively charged lysines/arginines face the other side, which is a typical feature and requirement for helical AMPs. The goal of the simulations is to understand each residue's physicochemical role in the membrane insertion process while earlier reports [29,31,32] have suggested the deeper a peptide penetrates the lipid membrane, the higher bactericidal potency it can have. A collective insertion events by multiple AMPs can initiate a membrane thining and eventual erupting process.

Microsecond long simulations for both peptides are performed and the dynamic interaction between AMPs and lipid bilayers before the final insertion was recorded and analyzed (see **Movie S1**).

The detailed insertion processes are described in **Supplementary Material 1**. In general, as analyzing the snapshots in **Movie S1**, we found the three tryptophan residues can interact with the choline groups (through cation–π interaction), phosphate groups (hydrogen bond) and aliphatic tails either from the same or different POPC lipids (**Fig. 3A**). The N-terminus rather than the C-terminus is the anchor to initiate a long insertion, while the 3 tryptophans can contact each of the aforementioned three lipid moieties in timely order, especially the one closest to the N-terminus (**Fig. 3B**). The exact snapshots summarizing such a process for W3_p2 are given in **Fig. 3A** where minimally two lipids need to be involved in the process. On the contrary, W3_p1 could interact with a single PC lipid molecule via all of its tryptophan residues in sequential order (first W3, then W6 and lastly W9) and therefore inserted into the membrane faster. In other words, AMP insertion process utilizing three tryptophan anchors to interact with the choline, phosphate and tail motifs of one lipid molecule rather than those of two lipid molecules simultaneously secure a higher chance for an early and possibly deeper insertion into the membrane. Knowing the importance of the pattern WXXWXXW (or "W 2 W 2 W" in TP-DB's expression) in a helical structure, we are interested in finding other helices in TP-DB, which (1) contain the W2W2W motif, (2) are positively charged to grant the amphiphilicity and (3) can maintain helical structure in isolation (suggested by relatively fewer tertiary contacts from neighboring parts of the protein).

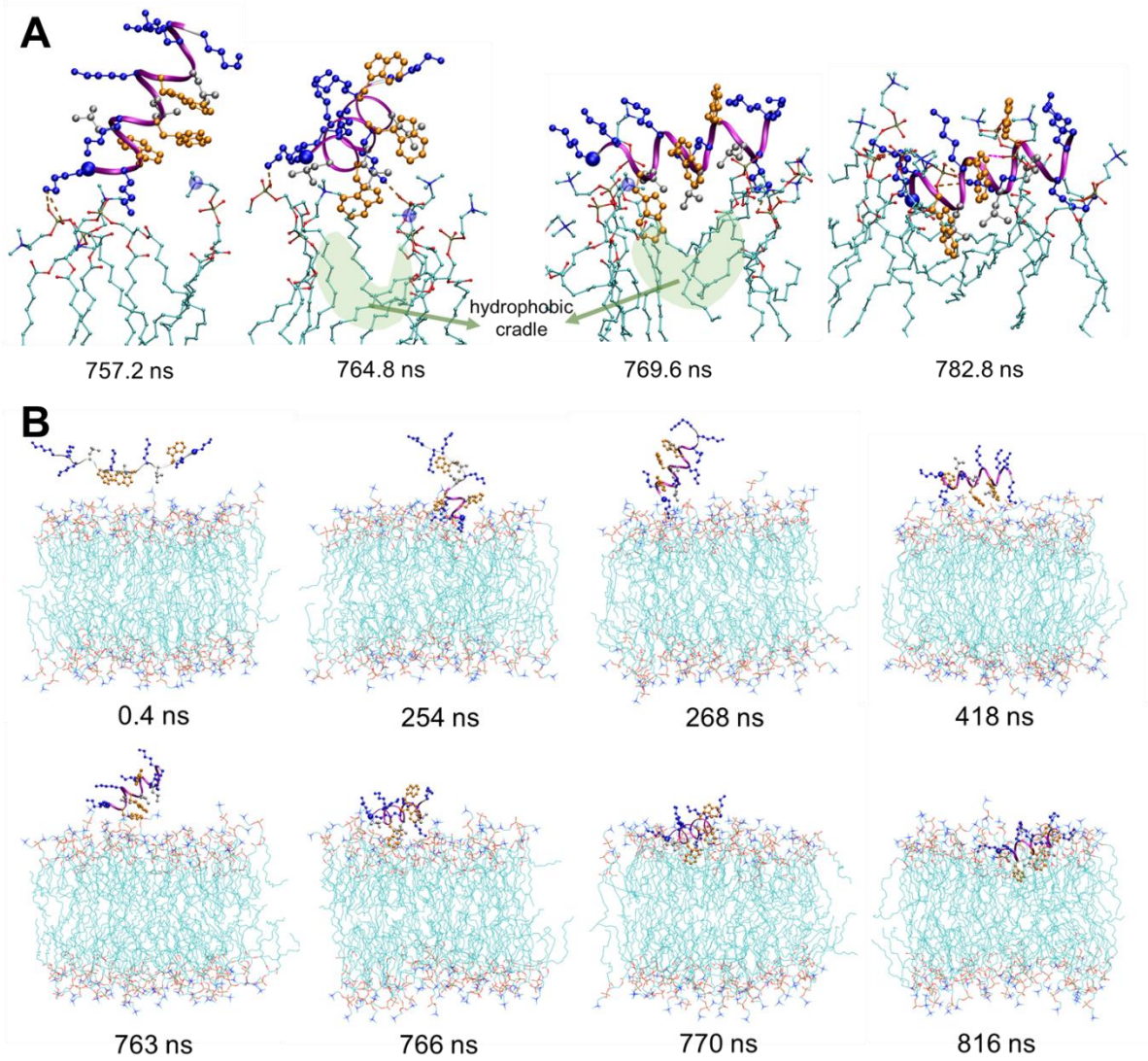

Fig. 3. Mechanistic details of the insertion of W3_p2 into PC. (**A**) The major four stages of the insertion steps, detailed in SI. The salt bridges (SBs) and hydrogen bonds (HBs) are represented using ochre and pink dashed lines, respectively. Lipids with any atom within 3 Å of the AMPs are shown with part of the aliphatic tails removed for clarity. The heavy atoms of the lipids are represented in thin CPK; O, P and C atoms are colored in red, brown and cyan, respectively. The N atoms of the PC lipids that have cation-π interactions with Trps are highlighted in the blue transparent VDW ball, while the golden ball-and-sticks are the three tryptophans facing the same side of the helical AMP. The heavy atoms of the positively charged and hydrophobic residues of the AMPs are represented in blue and silver CPK, respectively, whereas the aromatic residues are highlighted in orange. The Cα atoms of the N-terminal residues are highlighted using the blue VDW ball. (**B**) Representative snapshots of W3_p2 inserted into the model PC membranes. Reproduced and modified from a publicized patent [30].

The search result is shown in **Table 2**, we pick the most positively charged W2W2W motifs "WKCWARWRL" from the mycobacterium tuberculosis Zinc-dependent metalloprotease-1 (Zmp1) (PDB: 3zuk) for its relatively high positive charge and helical propensity. This motif is used to replace the original motif WLKWLKW in the AMP W3_p2 (sequence: KK **WLK** **WLK** **WLK** KK) and to create a new (potential) AMP.

Table 2. Results[#] obtained from the TP-DB for the query "W 2 W 2 W"

#	U#	Matched Sequence (MS)	Matched Pattern	Full Helix (FH)	PDB ID: Chain	Positions in PDB (MS), (FH)	Helical Propensity	Contact	Interacting Partners
1	1	WLEWIRW	W2W2W	DVQQLLIWLEWIRWESD	4eba:F	(287, 293), (280, 296)	1.189	8.857	4eba:F (324, 338)
10	2	WQQWLNW	W2W2W	DLATFWQQWLNW	3tsu:A	(651, 657), (646, 657)	1.131	7.857	
11	3	WQRWENW	W2W2W	VNSYLWQRWENWFWNVTLR	2lzq:A	(10, 16), (5, 23)	1.105	8.000	
12	4	WDAWLNW	W2W2W	WDAWLNWFR	1z8h:D	(198, 204), (198, 206)	1.089	7.286	
13	5	WKCWARW	W2W2W	NVEDWKCWARWRLIRARA	3zuk:B	(283, 289), (279, 296)	1.055	9.143	3zuk:B (415, 434)
15	6	WAWWISW	W2W2W	AGEWLGSWTIFYWAWWISWSPFVGMFLAR	4llh:A	(371, 377), (359, 387)	1.039	10.571	4llh:A (449, 479) 4llh:A (511, 525) 4llh:A (527, 545)
25	7	WVDWWQW	W2W2W	QWVDWWQWWVK	4n7k:L	(259, 265), (258, 268)	0.985	7.286	4n7k:M (81, 88) 4n7k:M (178, 192)
86	8	WKDWESW	W2W2W	SVLRKALHDSLHDCSHWFYTRWKDWESWYSQSF	3zrh:A	(481, 487), (460, 492)	0.799	8.857	3zrh:A (522, 533)
87	9	WTAWSTW	W2W2W	WTAWSTWKYC	3cb7:A	(106, 112), (106, 115)	0.638	7.714	
88	10	WPEWWNW	W2W2W	GWPEWWNWWLE	3wmm:L	(268, 274), (267, 277)	0.345	8.857	3wmm:M (82, 90) 3wmm:M (180, 193)
89	11	WPEWWGW	W2W2W	WPEWWGWWL	7prc:L	(259, 265), (259, 267)	-0.020	8.429	

Only the unique 11 sequences are shown here and the red block highlights the new motif we used for the AMP design.

[#] The result page is available online at http://dyn.life.nthu.edu.tw/design/helixdb/result/5a0b43e3a.

To further validate the use of TP-DB for W2W2W, we measured the minimal inhibitory concentration (MIC_{90}) at which growth of 90% of the pathogenic bacterium *E. coli* ATCC 25922 and fungus *C. albicans* SC5314 can be inhibited. A series of concentration of 120, 60, 30, 15, 7.5, 3.75 and 1.88 μg/ml of peptides are tested. Experiments are performed in pentaplicate, and MIC_{90} is determined as the majority results of the 5 replicates. The result (**Fig. 4**) shows that the newly designed AMP W3_db5 (KK **W**KC**W**AR**W**RL KK) has an anti-fungal MIC of 7.5 μg/ml, which is commensurate with that of W3_p1 (7.5 μg/ml) and W3_p2 (7.5 μg/ml), while it is a level lower at anti-bacterial MIC of 30 μg/ml than that of W3_p1 (15 μg/ml) and W3_p2 (15 μg/ml).

It should be noted that the original protein (mycobacterium tuberculosis Zinc-dependent

metalloprotease-1; PDB ID: 3ZUK) that contains this potent stretch "**W**KC**W**AR**W**RL" is not a transmembrane or membrane-bound protein; the helical stretch is situated close but not fully exposed to the water-protein interface. Therefore, TP-DB brings an interesting opportunity for researchers to put together

structurally resolved elements, matching a desired pattern, for the rational design of new therapeutics, herein showcased by a new class of AMP that is 6 point mutations away from the reported W3_p1 and W3_p2.

From the design point of view, these 6 point mutations do not seem to change much the required insertion thermodynamics [33] for the New AMP still being able to penetrate the peptidoglycans, reach the lipid-water interface, stay amphiphilic and structurally helical, assured by TP-DB, and eventually insert into the cell membrane.

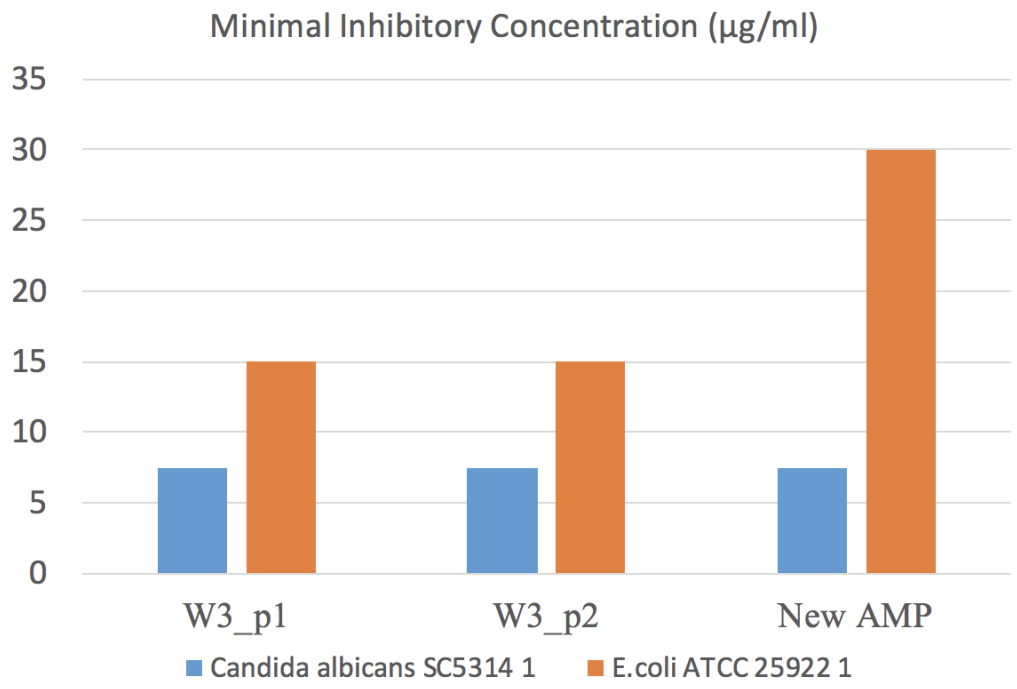

Fig. 4. The anti-fungal and anti-microbial activity of TP-DB-designed AMP (W3_db5) versus those of W3_p1 and W3_p2 that share the same "W 2 W 2 W" motif. MIC values are derived from the majority of pentaplicate measurements.

4. Conclusions

The pattern-based search engine without relying on sequence homology, realized in the newly introduced TP-DB, has been shown to provide a new avenue to the re-use or re-assembly of structured peptide fragments into new therapeutics or diagnostic kits. New therapeutics can therefore be designed not only with known functional sequence motifs but also with reliable structural templates. We foresee its future

extension to incorporate search engines for beta-sheet binders as well as for finding helices/β-sheets with the same Pfam motifs.

Acknowledgements

E.O.S. acknowledges financial supports from MOST and Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan. We thank vast computational resources for this project provided by High Performance Computing Infrastructure (HPCI), Japan and National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) of Taiwan. This work was supported by the Ministry of Science and Technology, Taiwan (103-2627-M-007-001 and 104-2113-M-007-019 to L.W.Y).

References

1. Jacob F. (1977) Evolution and tinkering. Science 196, 1161-1166
2. Jencks WP. (1981) On the attribution and additivity of binding energies. Proc Natl Acad Sci USA 78, 4046-4050
3. Erlanson DA. (2012) Introduction to Fragment-Based Drug Discovery. in Fragment-Based Drug Discovery and X-Ray Crystallography (Davies, T. G., and Hyvönen, M. eds.), Springer Berlin Heidelberg, Berlin, Heidelberg. pp 1-32
4. Kutchukian PS, Lou D, and Shakhnovich EI. (2011) *In Silico* Fragment-Based Generation of Drug-Like Compounds. in Library Design, Search Methods, and Applications of Fragment-Based Drug Design, American Chemical Society. pp 151-177
5. Zhang Z, Miller W, Schäffer AA, Madden TL, Lipman DJ, et al. (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 26, 3986-3990
6. Gilliland G, Berman HM, Weissig H, Shindyalov IN, Westbrook J, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28, 235-242
7. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26, 1781-1802
8. MacKerell AD, Bashford D, Bellott, Dunbrack RL, Evanseck JD, et al. (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J Phys Chem B 102, 3586-3616
9. Klauda JB, Venable RM, Freites JA, O'Connor JW, Tobias DJ, et al. (2010) Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. J Phys Chem B 114, 7830-7843
10. Darden T, York D, and Pedersen L. (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. J Chem Phys 98, 10089-10092

11. McQuarrie DA. (2000) Statistical mechanics, II ed., University Science Books, Sausalito, Calif.
12. Feller SE, Zhang YH, Pastor RW, and Brooks BR. (1995) Constant-pressure molecular-dynamics simulation - the Langevin piston method. J Chem Phys 103, 4613-4621
13. Tsai C-W, Hsu N-Y, Wang C-H, Lu C-Y, Chang Y, et al. (2009) Coupling molecular dynamics simulations with experiments for the rational design of indolicidin-analogous antimicrobial peptides. J Mol Biol 392, 837-854
14. Wang Y, Schlamadinger DE, Kim JE, and McCammon JA. (2012) Comparative molecular dynamics simulations of the antimicrobial peptide CM15 in model lipid bilayers. Biochim Biophys Acta Biomembr 1818, 1402-1409
15. Mahoney MW, and Jorgensen WL. (2000) A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. J Chem Phys 112, 8910-8922
16. Kucerka N, Tristram-Nagle S, and Nagle JF. (2005) Structure of fully hydrated fluid phase lipid bilayers with monounsaturated chains. J Membr Biol 208, 193-202
17. Jo S, Lim JB, Klauda JB, and Im W. (2009) CHARMM-GUI membrane builder for mixed bilayers and its application to yeast membranes. Biophys J 97, 50-58
18. Humphrey W, Dalke A, and Schulten K. (1996) VMD: Visual molecular dynamics. J Mol Graph Model 14, 33-38
19. Wang C-A, Liu Y-C, Du S-Y, Lin C-W, and Fu H-W. (2008) Helicobacter pylori neutrophil-activating protein promotes myeloperoxidase release from human neutrophils. Biochem Biophys Res Commun 377, 52-56
20. Yang Y-C, Kuo T-Y, Hong Z-W, Chang H-W, Chen C-C, et al. (2015) High yield purification of Helicobacter pylori neutrophil-activating protein overexpressed in Escherichia coli. BMC Biotechnol 15, 23
21. Walker IH, Hsieh P-c, and Riggs PD. (2010) Mutations in maltose-binding protein that alter affinity and solubility properties. Appl Microbiol Biotechnol 88, 187-197
22. Iankov ID, Haralambieva IH, and Galanis E. (2011) Immunogenicity of attenuated measles virus engineered to express Helicobacter pylori neutrophil-activating protein. Vaccine 29, 1710-1720
23. Hong Z-W, Yang Y-C, Pan T, Tzeng H-F, and Fu H-W. (2017) Differential effects of DEAE negative mode chromatography and gel-filtration chromatography on the charge status of Helicobacter pylori neutrophil-activating protein. PLoS One 12, e0173632
24. Clinical Laboratory Standards Institute. (2008) Reference Method for Broth Dilution Antifungal Susceptibility Testing of Yeasts; Approved Standard—Third Edition, M27-A3.
25. Clinical Laboratory Standards Institute. (2012) Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically; Approved Standard—Ninth Edition, M07-A9. Clinical and Laboratory Standards Institute, Wayne, PA.
26. Rothstein DM, Spacciapoli P, Tran LT, Xu T, Roberts FD, et al. (2001) Anticandida activity is retained in P-113, a 12-amino-acid fragment of histatin 5. Antimicrob Agents Chemother 45, 1367-1373
27. Hopp TP, Prickett KS, Price VL, Libby RT, March CJ, et al. (1988) A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification. Biotechnology (N Y) 6, 1204-1210

28. Slootstra JW, Kuperus D, Plückthun A, and Meloen RH. (1997) Identification of new tag sequences with differential and selective recognition properties for the anti-FLAG monoclonal antibodies M1, M2 and M5. Mol Divers 2, 156-164
29. Yu H-Y, Huang K-C, Yip B-S, Tu C-H, Chen H-L, et al. (2010) Rational Design of Tryptophan-Rich Antimicrobial Peptides with Enhanced Antimicrobial Activities and Specificities. ChemBioChem 11, 2273-2282
30. Yang L-W, Cheng J-W, Li H-C, Tsai C-Y, and YU H-Y. (2017) Evaluation system for the efficacy of antimicrobial peptides and the use thereof. United States
31. De Santis E, Alkassem H, Lamarre B, Faruqui N, Bella A, et al. (2017) Antimicrobial peptide capsids of de novo design. Nat Commun 8, 2263
32. Pyne A, Pfeil M-P, Bennett I, Ravi J, Iavicoli P, et al. (2017) Engineering monolayer poration for rapid exfoliation of microbial membranes. Chem Sci 8, 1105-1115
33. Almeida PF, and Pokorny A. (2009) Mechanisms of Antimicrobial, Cytolytic, and Cell-Penetrating Peptides: From Kinetics to Thermodynamics. Biochemistry 48, 8083-8093