EVALUATION OF VARIOUS AUTODOCK MOLECULAR DOCKING PROGRAMMES IN PREDICTING THE NEURAMINIDASE INHIBITORY ACTIVITY

NURUL NADIA BINTI RAIME

UNIVERSITI SAINS MALAYSIA

2015

EVALUATION OF VARIOUS AUTODOCK MOLECULAR DOCKING PROGRAMMES IN PREDICTING THE NEURAMINIDASE INHIBITORY ACTIVITY

by

NURUL NADIA BINTI RAIME

Thesis submitted in fulfilment of the requirements
for the degree of Master of Science

OCTOBER 2015

ACKNOWLEDGEMENT

Alhamdulillah. In the name of Allah, the Entirely Merciful, the Especially Merciful. All praise is due to Allah, Lord of the worlds. With His guidance, I'm finally completing this dissertation.

My utmost gratitude and deepest appreciation goes to my supervisor; Professor Dr. Habibah A Wahab, who has shown the attitude and the substance of a genius: she continually and persuasively conveyed a spirit of adventure in regard to research and scholarship, and an excitement in regard to explore new experience. Without her supervision and constant help, this dissertation would not have been possible. I would also like to express my gratitude to Associate Dr. Susumu Date and Dr Kohei Ichikawa- sensei of Cyber Media Centre, Osaka University for helpful and thoughtful discussion during my exchange study in Japan. I would also like to thank Dr Mariyana and Dr Vannajan who gave me constructive comments and warm encouragement along the way.

I would like to show my greatest appreciation to PHDs lab members especially Dr Choi Sy Bing, Dr. Belal Al Najjar, Mohd Yusuf, Maywan and Saira for insightful comments and suggestions along my Master journey. To my colleagues in the lab; Teh Ban Hong, Lee Guan Seng, Lim Kok Keong, Hanim, Adliah and Fatehah who colour up my lab day and keep me to strive towards the end. My deepest thought goes to my beasties; Nurul Izzati and Rina Nuwarda who always been there for me in shine or in rain. An appreciation also granted to my former lab member of Shimojo Laboratory in Osaka University especially Tada Taiki-san, Morita Taiki-san and other lab members who always offer their handful hand during my exchange study.

I would also like to express my gratitude to Institute of Postgraduate Studies (IPS) and Ministry of Science, Technology, and Innovation, Malaysia (MOSTI) for granting me full-handed scholarship of Post Graduate Scheme (PGS) during my study. Thank you too to Universiti Sains Malaysia for granting our lab with Research University grant.

Last but not least, I would like to thank my husband, Mohd Azerulazree, for his love, kindness and support he has shown in the moments when there was no one to answer my queries. Special thanks also go to my family. Words cannot express how grateful I am to my beloved mother, Surinah Muhamad, father, Raime Mohd Zain, my mother-in law, Noorzanah Lee and father-in-law, Jamilan Md Noor, for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far.

Nurul Nadia Raime

TABLE OF CONTENTS

Page

CHAPTER TWO: MATERIALS AND METHODS

CHAPTER THREE: RESULTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD3	AutoDock 3
AD4	AutoDock 4
AD4_flex	AutoDock 4 of flexible recceptor
ADVina	AutoDock Vina
ADVina_flex	AutoDock Vina of flexible recceptor
ALA	Alanine
ARG	Arginine
ASN	Asparigine
ASP	Aspartic acid
BFGS	Broyden-Fletcher-Goldfarb-Shanno
C	Carbon
$CaCl_2$	Calcium chloride
CADD	Computer aided drug design study
Conc.	Concentration
ddH_2O	Deionized water
DNA	Deoxyribonucleic acid
DPF	Docking parameter file
DSII	NCI Diversity Set II
DUD	Directory of Universal Decoy
FEB	Free energy of binding
FS	Fitness score
GA	Genetic algorithm
GLU	Glutamic acid

GOLD	Genetic Optimization Ligand Docking
GPF	Grid parameter file
H	Hydrogen
HA	Haemagglutinin
HBond	Hydrogen bond
HPAI	Highly Pathogenic Avian Influenza
HTS	High-throughput screening
IC_{50}	Fifty percent inhibitory concentration
ILE	Isoleucine
ILI	Influenza- Like Illness
LGA	Lamarckian Genetic Algorithm
LYS	Lysine
$M_{1,2}$	Molarity
MUNANA	2'-(4-methylumbelliferyl)-a-D-*N*-acetylneuraminic acid sodium salt hydrate
NA	Neuraminidase, sialidase
NAI	Neuraminidase inhibitor
NaOH	Sodium hydroxide
NCI	National Cancer Institute
Neu5Ac	Sialic acid, N-acetyl neuraminic acid
Neu5Ac2en	DANA, 2-deoxy-2,3-dehydro-N-acetyl-neuraminic acid
NP	Nucleoprotein
NS	Non-structural
OST	Oseltamivir
PDB	Protein Data Bank

PRO	Proline
RMSD	Root Mean Square Deviation
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
SBDD	Structure based drug design
SBDD	Structure based drug design
SER	Serine
SW	Solis and Wets
THR	Threonine
TRP	Tryptophan
TYR	Tyrosine
$V_{1,2}$	Volume
VS	virtual screening
WHO	World Health Organization
ZMR	Zanamivir

LIST OF SYMBOLS

%	Percentage
°C	Celcius
Å	Angstrom
µg/mL	microgram/millilitre
µL	microlitre
mL	millilitre
µM	micromolar
mM	millimolar
U	Unit
U/mL	Unit/millilitre
π	pi

LIST OF PUBLICATIONS

RAIME, Nurul Nadia (2011). Principle study to identify and improve ranking methods of virtual screening between tools; Autodock 3.0.5 and Autodock Vina; based on neuraminidase system. The 2nd International Seminar on Chemistry 2011, 24/25 November 2011, Universitas Padjadjaran, Indonesia.

RAIME, Nurul Nadia (2011). Autodock 3, Autodock 4, Autodock Vina and GOLD; The principle study to identify ranking methods between tools for virtual screening based on Neuraminidase system. Computational Aided Drug Design 2011. 9-12 Dicember 2011, Universiti Sains Malaysia, Penang.

PENILAIAN TERHADAP PELBAGAI PROGRAM PENDOKKAN MOLEKUL DALAM MERAMAL AKTIVITI PERENCAT NEURAMINIDASE

ABSTRAK

Kajian ini dijalankan untuk menilai skor dan kedudukan program pendokkan molekul; AutoDock 3.0.5 (AD3), AutoDock 4.2 (AD4) dan AutoDock Vina (ADVina) dalam meramal aktiviti perencat neuraminidase dari NCI diversiti set II (DSII) dan untuk membandingkan skor kedudukan ligan dengan program Genetic Optimization Ligand Docking (GOLD) yang berfungsi sebagai ukuran kedudukan. Kemudian, ramalan ini dibandingkan dengan NA enzim assay. Terdapat dua jenis protein yang digunakan dalam kajian ini iaitu; set pertama adalah terhadap protein kekal dan set kedua adalah terhadap pilihan asid amino yang fleksibel pada protein. Pemeriksaan maya protein NA kekal telah dijalankan terhadap AD3, AD4 dan ADVina, manakala NA dengan asid amino yang fleksibel telah dijalankan hanya terhadap AD4 dan ADVina. Ini disebabkan oleh ketiadaan fungsi asid amino yang fleksibel dalam program AD3. Seterusnya, 15 kompaun teratas dari setiap set disiasat dengan aktiviti enzim assay NA untuk mengetahui tingkah laku aktivitinya. Pada masa yang sama, pendokkan molekul dengan program GOLD terhadap 15 sebatian teratas dari setiap set telah dijalankan untuk menganalisis status hubungannya. Hasil menunjukkan protein NA kekal menggunakan program AD4 menunjukkan keputusan aktiviti assay 4 daripada 15 sebatian yang mempunyai lebih daripada 50% perencatan (kompaun 37, 45, 49 dan 51). Protein NA kekal bagi ADVina dan AD3 memberikan dua (kompaun 30, dan 31) dan tiga kompaun (kompaun 5, 8, 16)

masing-masing lebih daripada 50% perencatan. Pendokkan molekul dengan asid amino NA yang fleksibel oleh program ADVina menghasilkan 2 kompaun lebih daripada perencatan 50% (kompaun 20, dan 31) manakala program AD4 tidak berjaya menghasilkan kompaun lebih 50% tahap perencatan. Ligan yang menghasilkan lebih 50% aktiviti perencatan didapati dari set program pendokkan molekul yang berbeza. Penemuan ini melambangkan bahawa setiap program pendokkan molekul menangkap ligan secara esklusif. Walau bagaimanapun, berdasarkan keputusan kajian, tiada bukti menunjukkan skafold yang eksklusif ditangkap oleh program dok yang berbeza. Oleh itu, tiada signifikasi yang jelas ke atas ligan kedudukan teratas dengan perencatan aktiviti lebih daripada 50%.

EVALUATION OF VARIOUS AUTODOCK MOLECULAR DOCKING PROGRAMMES IN PREDICTING THE NEURAMINIDASE INHIBITORY ACTIVITY

ABSTRACT

This study was done to evaluate the scoring and ranking of molecular docking programmes; AutoDock 3.0.5 (AD3), AutoDock 4.2 (AD4) and AutoDock Vina (ADVina) in predicting the inhibitory activity of neuraminidase inhibitor taken from NCI Diversity set II (DSII) and to compare those ranking and scoring with that of Genetic Optimization Ligand Docking (GOLD), as the gold standard for ranking. Then, validate this prediction with NA enzyme assay. There were two types of protein used; first set was against rigid protein and second set was against specified flexible residues of protein. Virtual screening of rigid NA protein was run against AD3, AD4 and ADVina, whilst flexible side chain of NA was run only against AD4 and ADVina. This was due to unavailable flexible side chain function in AD3 manual programme. Top 15 compounds from each set were further investigated with NA enzyme activity assays for its activity behaviour. Simultaneously, docking simulations with GOLD against top 15 compounds from each set were done to analyse its correlation status. The result showed screening of rigid NA protein using AD4 gave promising assay activity of 4 out of 15 compounds which have more than 50% inhibition (compound 37, 45, 49 and 51). Rigid NA protein against ADVina and AD3 gave two (compound 30, and 31) and three compounds (compound 5, 8, 16) with over 50% inhibition respectively. Docking with flexible residues of NA with ADVina resulted in 2 compounds with more than 50% inhibition (compound 20, and 31) and AD4 with no compound over 50% inhibition level. Compounds with more

than 50% inhibition activity were captured from different docking sets. Findings symbolized that each docking set capture ligand exclusively. However, based on the results, it showed no exclusive scaffold captured by distinguished docking software. Hence, there was no significant difference on top 15 selected compound constituents with more than 50% activity.

CHAPTER ONE

INTRODUCTION

1.0 Problem of statement

Molecular docking is one of the most popular molecular modelling techniques. It starts with a simple theory of 'lock and key' model, and yet the ideas are expanding paralleled with computer technologies. Thousands of docking softwares can be found on the internet; either for free or with licence required (Kitchen et al., 2004). Each docking program has its own advantage over the others (Cole et al., 2005, Perola et al., 2004, Bursulaya et al., 2003, Paul and Rognan, 2002, Kellenberger et al., 2004). The programs have their own specialties and advantages on how to dock compound into the protein with different manipulated searching algorithm and force field.

Predicting how a small molecule binds at the binding site is not an easy task although there are many simple docking tools available online. There are many problems associated with protein-ligand docking simulation (Cole et al., 2005). These include reliability of conformation prediction, the ability of scoring function and the capability of searching algorithm to find optimise ligand conformations and orientations (Cole et al., 2005, Perola et al., 2004, Leach et al., 2006, Huang et al., 2006, Warren et al., 2006).

Explosion of researches in biomolecular x-ray crystallography provides continuously the addition of protein and nucleic acid structures in various databases. These structures might be the targets for bioactive agents in controlling diseases. Using molecular docking, these targets can be ascertained to interact with the

potential therapeutic agents. With the aid of virtual screening and good computational power, the docking process can be accelerated and applied to thousand compounds in search of potential therapeutic agents (Verdonk et al., 2004, Seifert and Lang, 2008). In addition, by screening compound libraries, it may increase hit rates in lead discovery in a fast time track. Hence, without doubts, molecular docking can help researchers to save money, raw materials as well as time consumption (Verdonk et al., 2004).

As a case study, searching for potential therapeutic agent(s) for Neuraminidase (NA) of influenza virus was selected. Previously, successful NA inhibitors; zanamivir (ZMR) [Ralenza^{TM}] and oseltamivir (OST) [Tamiflu^{TM}] (von Itzstein et al., 1993, Kim et al., 1997) were discovered from structure based drug design (SBDD). However, due to the ability of NA to undergo mutation; novel inhibitors are needed continuously to fight against the virus (D'Ursi et al., 2009, García-Sosa et al., 2008).

The aim of this research is thus to evaluate the scoring and ranking of molecular docking programmes; AutoDock 3 (AD3), AutoDock 4 (AD4), AutoDock Vina (ADVina) in predicting the inhibitory activity of neuraminidase inhibitor and to compare those ranking and scoring with that of Genetic Optimization Ligand Docking (GOLD), the gold standard for ranking. Subsequently, these predictions were validated with neuraminidase enzyme assay. The test compounds for docking were taken from NCI Diversity Set II (DSII) library which comprises 1364 compounds. The ranking was based according to their free energy of binding (FEB) between ligand and receptor.

1.1 Molecular Modelling

Molecular modelling briefly can be interpreted as idealised description of a system or process, usually by using mathematical language devised to facilitate calculations and predictions of respective molecule state (Leach and Schomburg, 1996). The collaboration of chemistry, biology, and physics with the aid of mathematical interpretation has brought up a new dimension toward human achievement in explaining the 'unseen' properties and mechanisms of molecular stage reaction. Previously, molecular modelling research is restricted to only a small number of scientists due to lack of computer hardware and software access. Nowadays, computer workstations are growing and can be purchased in relatively cheaper price than before. Plus, molecular modelling can now be performed in any laboratory and classroom sake.

Before the twentieth century, human only obtained medicines in the form of herbs and potions (Silverman, 2004). The isolation of active compounds in herbs was only done in the late nineteenths century which lead to the revolutionary of medicinal chemistry (Patrick, 2009, Silverman, 2004). Explosion of drug discovery has resulted in increasing billion of dollar pharmaceutical companies (Coupez and Lewis, 2006, Macarron et al., 2011). However, searching for suitable drug against specific disease is such a lengthy process without the aid from Computer Aided Drug Design (CADD) (Murray et al., 1999, Klebe, 2006). Traditional way of drug design was costly with time consuming and wasted in used of raw materials. Alternatively, the introduction of virtual screening (VS) before experimental high-throughput screening (HTS) (Macarron et al., 2011, Hertzberg and Pope, 2000) helps to filter for leads from library of thousand compounds virtually and consequently reduced the assay

volume (Ghosh et al., 2006, Patrick, 2009). In addition, the combination between structure based drug design and high speed clusters computer has allowed larger screening of libraries of compounds against pharmaceutical-relevant targets (Cosconati et al., 2010b, Beuscher Iv and Olson, 2007). Nowadays, virtual screening is a routine method when it comes to screen biologically active compounds in the nick of time. Screening can easily be done against a library of compounds such as from NCI repository and also large database like Directory of Universal Decoy (DUD) database (Huang et al., 2006).

Various operations carried out in molecular modelling involve the use of programs or algorithm which are able to calculate the structure and property of investigation's molecule. For example, it is possible to calculate the energy of a particular arrangement of atoms (conformations), modification of structure to create a minimum energy and calculation of varieties of properties (heat of formation, charge, dipole moment. Basically, there are two categories for computational method; molecular mechanics (MM) and quantum mechanics (QM) which are used to calculate structure and property data of molecule in question. QM considers the interaction between electron and nuclei to calculate the properties of the molecule. While, MM treats atom as a whole, without consideration of electron and nuclei interaction. Methods of calculation chosen depends on the type of calculation needs to be done as well as molecule size. QM calculation is limited to calculate molecule containing ten to hundreds of atoms while MM calculation can calculate molecule with thousands of atoms. Hence, at the present time, only MM approach is suitable for calculation of protein interaction with consideration of computer capability.

1.1.1 Molecular Mechanics (MM)

In MM, a molecule is treated as a series of atoms (spheres) connected by bonds (springs). Equations are used which follow the laws of classical physics and apply them to nuclei without consideration of electrons. Then, equations derived from classical mechanics are used to calculate the different interactions and energies (force filed) resulting from bond stretching, angle bending, torsional energies and non-bonded interactions. Such calculations require parameters or data's being stored in tables within the programme and which describe interactions between different sets of atoms. MM is useful for the calculations of energy minimization, identifying stable conformations, generating various conformations, energy calculations for specific conformations and studying of molecular motions (Patrick, 2009).

One of the vital calculations by MM is energy minimization of 3D structures. Once after the structure is built or downloaded directly from any source, energy minimizations must be carried out to avoid unfavourable bond lengths, bond angles or torsion angles (Kitchen et al., 2004, Leach and Schomburg, 1996). The energy of the starting structure is calculated to see whether it is stable or not. If the structure is unstable, even a slight alteration in bond length or bond angle will have a large effect on the overall energy of the molecule which will result in large energy difference (ΔE). Hence, modification of bond lengths, bond angles, torsion angles and non-bonded interactions are necessary until a stable conformation is obtained. There are various software packages available to calculate the energy minimizations such as Hyperchem, Discovery Studio, Chem 3D and etc. (Patrick, 2009). In this study, Hyperchem package 6.01 for windows was used for minimization purposes.

1.1.2 Structure Based Drug Design (SBDD)

Structure based drug design was initiated based on the availability of reliable drug target structures from experimental sources such as X-ray crystallography and NMR spectroscopy (Leach and Schomburg, 1996, Hetényi and van der Spoel, 2002). Screening of ligands by structure based method is proven to increase the likelihood of finding compatible ligand from among many structures in the database (Shoichet, 2004).

There are many successful screening of genuinely novel compounds such as inhibitors of protein kinases, HIV-protease, carbonic anhydrase, neuraminidase (NA), and thrombin (Bortolato et al., 2013, Rizzo et al., 2000). SBDD has been applied to NA system and resulted in the discovery of active inhibitors towards influenza virus A. NA has a conserved catalytic site with its natural inhibitor (sialic acid (Neu5Ac)). The structural information's of NA has been utilized to mimic the transition state of Neu5Ac which in the form of DANA (Neu5Ac2en) to produce new inhibitors (OST and ZMR) with higher binding affinity towards receptor (von Itzstein et al., 1993).

1.1.3 Molecular Docking

There are abundance numbers of docking programmes available on the internet and the number is increasing over the years. Examples of frequently used docking programmes are AutoDock (Morris et al., 2009), GOLD (Verdonk et al., 2003), DOCK (Makino and Kuntz, 1997), FlexX (Rarey et al., 1996) , Glide (Friesner et al., 2004) and ICM (Abagyan et al., 1994).

The rule of thumbs when applying molecular docking is to balance between the desire for a robust and accurate procedure (Morris et al., 1998). Secondly is to balance the desire to keep the computational demands at a reasonable level (Morris et al., 1998, Morris et al., 2009). Historically, early approaches to ligand docking is by assuming both ligand and protein to be rigid as applied in DOCK program (Kuntz et al., 1982). This assumption encounters many limitations in many studied systems (Jones et al., 1997, Hetényi and van der Spoel, 2002). Then, the theory evolved with significant modification such as enabling the ligand conformational degrees of freedom to be explored (Leach et al., 2006). Nowadays, most of docking programs treat the ligand as flexible with rigid or slightly flexible receptors (Leach et al., 2006, Morris et al., 2009). However, by increasing degrees of flexibility of certain receptor might increase the computational complexity and also the cost (Carlson, 2002).

1.1.3.1 Evolution of Molecular Docking's Software

Original approach of molecular docking comes from the concept of 'lock and key' of rational drug design. The theory is evolving with time to mimic the actual situation of docking hence optimise the results. Currently, most researchers roughly class the docking mode into four models (Figure 1.1); 1) lock and key, 2) conformation isomerism, 3) induced-fit and 4) conformation selection (Chen, 2015, Mukhopadhyay, 2014). However, the precise algorithms used to fit the 'key' (ligand) into the 'lock' (receptor protein) vary across programmes. A user may choose different docking modes based on the ability of their computational hardware and the character of target protein. Basically, rigid docking refers to the protein and ligand being fixed so that bond angles or lengths are not changeable. While, flexible docking allows for conformational shifts, changeable bond angles and lengths of either receptor or ligand or both at a time.

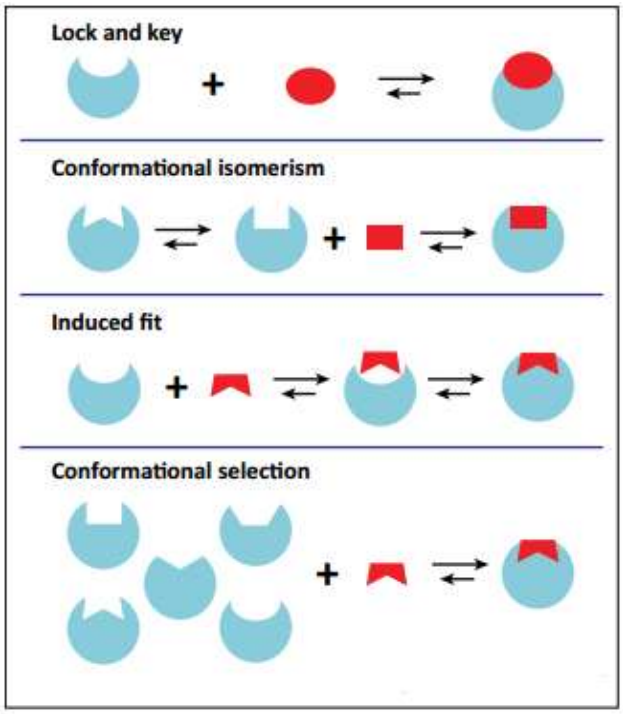


Figure 1.1: The docking mode roughly divided into four types of docking mode (Chen, 2015).

Efficient protein-ligand docking can be obtained with aid from two essential components, which are ‘sampling’ and ‘scoring’ (Huang and Zou, 2010, Schulz-Gasch and Stahl, 2004, Kroemer, 2007). Sampling can be divided to two categories which are ligand sampling and protein flexibility. Sampling refers to a probability of finding a correct orientation of flexible ligand and the proper conformation of the receptor molecule (Kroemer, 2007). Ligand sampling indicates orientations of ligand around specified binding site while protein flexibility occur when ligand bind to the binding site and induce protein conformational changes. More significant changes will occur when the side chains of a macromolecule are assigned as flexible (Huang and Zou, 2010).

Figure 1.2 shows basic graphic simulation of sampling and scoring in docking with large compound collections. One way of achieving this is via *in silico* or virtual screening (VS) of compound libraries against receptor’s binding site or commonly known as SBDD. Screening of compounds against specific binding pocket in a variety of positions, conformations, and orientations will results in multiple ‘pose’ predictions (Figure 1.2). In order to identify the energetically most favourable ‘pose predictions’, each pose is evaluated or ‘scored’ based on its complementarity to the target in terms of shape and properties such as electrostatics. A good score for a given compound indicates that it is potentially a good binder. This process is repeated for all compounds in the libraries which are subsequently rank-ordered by their scores or predicted affinities (for example binding energy) (Kroemer, 2007).

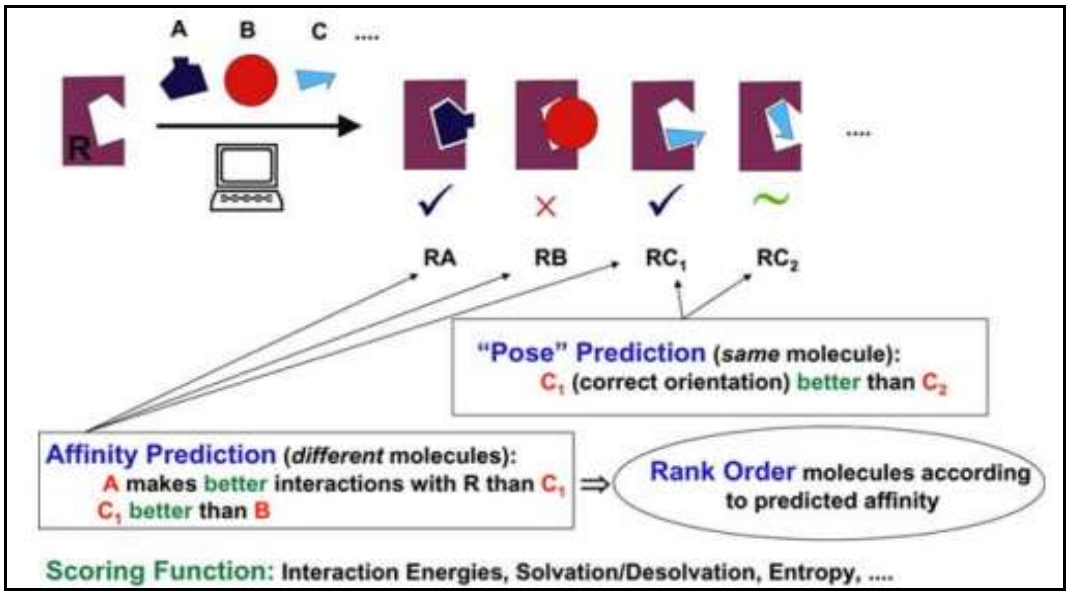


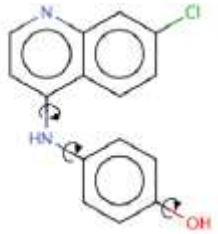
Figure 1.2: R symbolizes a receptor structure whiles A, B, and C represents small molecules to be docked into the receptor (Kroemer, 2007).

Scoring function will determine each molecule's ranking. Scoring is meant to evaluate affinity between a ligand and a protein based on the energy function and subsequently to rank the ligands according to the relative binding affinity. Numerous scoring functions have been developed and they can be divided into three groups: 1) force field based; 2) empirical; and 3) knowledge-based scoring functions (Schulz-Gasch and Stahl, 2004, Huang and Zou, 2010).

In this study, four types of docking programmes were used. The softwares were chosen are those common tools used among researchers (Chen, 2015) and availability in the laboratory: three AutoDock (AD) programmes from The Scripps Research Institute group while GOLD was subscribed from CCDC software Ltd. Company (Jones et al., 1997). AD programmes are docking tools released under open source licences (GNU General Public Licence and Apache Open Source Licence) (Chang et al., 2010, Goodsell et al., 1996, Morris et al., 1998, Morris et al., 2009). The first AD version was first distributed in the year of 1990 (Goodsell and

Olson, 1990), followed by the second version in the year of 1996 (Goodsell et al., 1996), while the third version was released in 1998 (Morris et al., 1998) and the number of user has grown with the latest version of AD4 programme released in 2007 (Morris et al., 2009, Chen, 2015). **Table 1.1** below briefly describes each docking programmes used in this experiment.

Table 1.1: Comparison between the docking programmes used in the study

	AD3	AD4	Vina	GOLD
Molecular representation				
Main program	consists of two main programs: • AD performs the docking of the ligand to a set of grids describing the target protein; • Autogrid pre-calculates these grids.	consists of two main programs: • AD performs the docking of the ligand to a set of grids describing the target protein; • Autogrid pre-calculates these grids.	Does not require choosing atom types and pre-calculating grid maps for them. • Calculates the grids internally, for the atom types that are needed, and it does this virtually instantly.	Using Gold Suite program which include: • Hermes visualizer for structure and binding site preparation, result's viewing and descriptor's calculation. • GoldMine for analysis of post results.
Macromolecule flexibility (side chain)	No	Yes	Yes	Yes

Continued:- **Table 1.1**: Comparison between the docking programmes used in the study

Atom types	• C • N • O • S • H • X or M	Expanded set of atom types as AD3 included: • A (aromatic carbon) • C (aliphatic carbon) • HD (hydrogen donor) • NA (nitrogen acceptor) • OA (oxygen acceptor) • SA (sulphur acceptor)	Not required to choose • Used automatically according to X-score (Wang et al., 2002) atom typing scheme	SYBYL atom type (Ash et al., 1997) Example: • sp2 oxygen= O.2 • sp3 oxygen= O.3 • tetrahedral nitrogen= N.3/N.4 • planar nitrogen= N.pl3/N.am • carboxylate oxygen= O.co2 • etc.
Initial structure	• MOL2 • PDB	• MOL2 • PDB	• MOL2 • PDB	• MOL2 • PDB
Input files	• Ligand –pdbq • Macromolecule –pdbqs	• Ligand –pdbqt • Macromolecule –pdbqt	• Ligand –pdbqt • Macromolecule –pdbqt	• N/A
Scoring function	• Van der Waals • Electrostatic • Hydrogen bond • Torsional penalty • Desolvation	• Van der Waals • Electrostatic • Hydrogen bond • Torsional penalty • Desolvation • Update in charge based-desolvation term • Improvement in directionality of H bond • Unbound state model	• Mostly inspired by X-score (Wang et al., 2002) • Hydrophobic interaction (van der Waals) • Hydrogen bond • Torsional penalty	• Protein-ligand hydrogen bond energy (external H-bond) • Protein-ligand van der Waals (vdw) energy (external vdw) • Ligand internal vdw energy (internal vdw) • Ligand torsional strain energy (internal torsional)

Continued:- **Table 1.1**: Comparison between the docking programmes used in the study

Searching algorithm	• SA (Simulate annealing) • LGA (Lamarckian GA)	• SA • LGA	• Inspired by X-score (Wang et al., 2002) with some modification	• GoldScore • ChemScore • ASP (Astex Statistical Potential) • CHEMPLP (Piecewise Linear Potential (PLP)) • User Defined Score

1.2 Influenza Neuraminidase

1.2.1 The Influenza virus

One of the darkest moments over mankind history was written in the year of 1918-1919, where mankind were afflicted with influenza virus disease or better known as 'Spanish' A(H1N1) influenza virus. The virus swept across the globe and infected approximately 25% of the entire population and sacrificed at least 50 million human's lives worldwide (Reid et al., 2003, Lin et al., 2000, Potter, 2001). In the following years, the A (H1N1) virus has continued to circulate among humans and caused a number of severe outbreaks in the year between 1920s-1950s (Viboud et al., 2006b, Viboud et al., 2006a). Particularly, an outbreak in the year of 1957, which also known as 'Asian flu', showed that A(H1N1) had been replaced by a reassorted A(H2N2) influenza virus and caused almost 1 million deaths (Shen et al., 2009). In 1968, the virus had evolved again in the form of H3N2 and it caused what was better known as 'Hong Kong Flu' pandemic (Viboud et al., 2006b, Viboud et al., 2006a). The outbreak caused approximately 1 million deaths. In 1977, an identical A(H1N1) virus from 1950s caused the Russian influenza pandemic (Nakajima et al., 1978, Horimoto and Kawaoka, 2001).

Around 1997, H5N1 virus was first isolated from human (Claas et al., 1998, Chan, 2002). Highly Pathogenic Avian Influenza (HPAI) H5N1 virus was virulent and believed to occur among birds and able to cause high mortality especially in domestic poultry. However, transmission of virus from bird to human were recorded but human to human transmission is quite rare (Yuen and Wong, 2005). Research showed that if transmission between human 'efficiently' occur, the effect might be exceeding the total death caused by 1918 Spanish flu (Grienke et al., 2012, Santos-

Preciado et al., 2009). Thus, this fact had raised concern over the possibility of a major pandemic among world populations (Horimoto and Kawaoka, 2001, Snacken et al., 1999).

In the year of 2009, World Health Organization (WHO), due to the rapid global spread of a strain influenza A (H1N1) virus, has warranted moving the global pandemic alert level to phase 6 (WHO, 2010a) as 18366 deaths recorded around the world in such short time span. Phase 6 indicated that a global human pandemic of this virus isolate was under way (WHO, 2010b, Schnitzler and Schnitzler, 2009, CDC, 2010). Malaysia was also not spared as evidence from the confirmation by Ministry of Health Malaysia, where 91 death cases reported due to A (H1N1) and about 15380 cases for patients with Influenza-like Illness (ILI) symptom as of 17th July 2010 (MOH, 2010).

Since then, unpredictable sporadic spreading of various influenza viruses has been recorded up to 2013 (WHO, 2013b). Starting from 2011 to 2012, about 309 cases of human infected H3N2 virus were reported in the United State of America alone. Meanwhile, in 2013, an outbreak caused by H7N9 virus occurred in China with 135 people infected (CDC, 2013b, WHO, 2013a). Fortunately, both strains show no sustainability in human to human transmission yet and most of the cases were due to long exposure towards pigs (CDC, 2013a, Ke et al., 2013). The unforeseen effect of the influenza virus transmission is felt globally each year. Therefore, extreme effort is needed to control the spreading of the virus.

1.2.2 Structure and characteristics of influenza virus

Influenza virus, which belongs to *orthomyxoviridae* family, consists of three different strains, types A, B and C. However, most major outbreaks of influenza are due to influenza type A. Type B usually caused milder infections among human compared to type A as it mutates slowly over time (Kamali and Holodniy, 2013). Meanwhile, type C is the most stable strain compared to both of the strains and not relatively common to cause any harm to human being (Whitford, 2005).

Generally, viral classification (into the three different types) is done based upon the matrix and nucleoprotein (NP) antigens (Webster et al., 1992). Influenza A virus can be further classified into subtype based on the antigenic properties of its surface glycoproteins of haemagglutinin (HA) and neuraminidase (NA), where both HA and NA play essential roles for infection (Udommaneethanakit et al., 2009, Horimoto and Kawaoka, 2001, Webster et al., 1992, Gamblin and Skehel, 2010, Kamali and Holodniy, 2013) to proceed. Currently, they are about 16 serotypes of HA (H1-H16) and 9 serotypes of NA (N1-N9) found (Kamali and Holodniy, 2013, Shi et al., 2010) giving 144 probabilities of the arrangement combinatorial diversity of the possible A/HN subtypes (Metzgar et al., 2010). For HA, the majority of human infections were caused by H1 and H3 and in occasional cases caused by H5, H7 and H9. NA with types N1, N2 and N9 are currently known circulating in human population (Kamali and Holodniy, 2013).

Typically, virus can be described as protein packages which delivered some genetic material to desire host cells. Influenza A virus is a member of orthomyxovirus family with negative strand RNA composed of eight unique segments (**Figure 1.3).** The virus is in the form of pleomorphic object, where it exists as irregular round bodies. The diameter of the virus is about 80 to 120nm. Eight

RNA segments bind to a nucleoprotein (NP) and associated with a polymerase complex and are located inside a shell. The shell is formed by the matrix protein and surrounded by a lipid membrane (derived from the host cell) embedded with two important glycoproteins, haemagglutinin (HA) and neuraminidase (NA) (Whitford, 2005, Patrick, 2009).

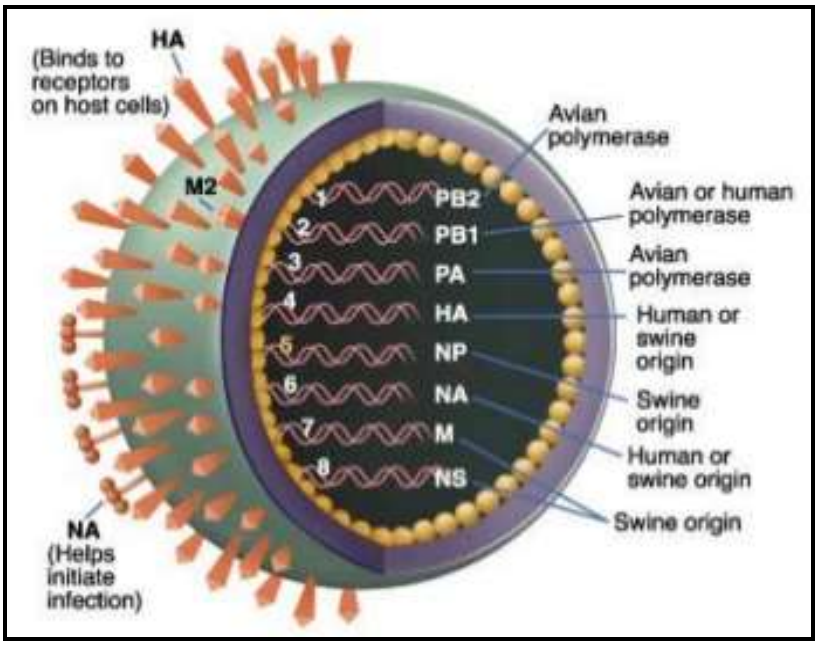


Figure 1.3: Schematic representation of influenza virus virion.

For cells to be infected, a single virus particle must possess eight unique RNA segments. The segments then encode 10 recognised gene products which are two viral polymerase basic subunits (PB1, and PB2), Polymerase A (PA), HA, NP, NA, M1 and M2 proteins, NS1 (non-structural 1) and NS2 (non-structural 2) proteins (Tan et al., 2015). Basic function for each gene is mentioned as below. HA is responsible for virions binding and fusion towards the host cell. While, NA plays a role to cleave terminal sialic acid from the host cell and release new virions for continuing the spread of the virus. PB1, PB2 and PA are three polymerase proteins

associated with RNA polymerase and are responsible for replication and translation process (Rodriguez-Frandsen et al., 2015). M1 protein on the other hand, is the most abundant protein in the virus virion (Nayak et al., 2004). M2 protein acts as an ion channel which serves to control the pH of the virion's interior (Webster et al., 1992). Non-structural protein; NS1 and NS2 functions are postulated to kill the host cells responded to the virus infection (Gong et al., 2007). However, specific function of NS2 is still unknown (Whitford, 2005, Webster et al., 1992, Nayak et al., 2004).

1.2.3 Neuraminidase: The structure and function

Neuraminidase (NA) also known as sialidase (EC 3.2.1.18) is one of the major surface glycoprotein despite HA of the influenza virus with approximately molecular weight of 240kD. The viral enzyme was discovered by Alfred Gottschalk and co-workers (Gong et al., 2007, Gubareva et al., 2000). It is known for its role to cleave terminal α-ketosidic linkage sialic acid residues (**Figure 1.4**) which linked to glycoproteins, glycolipids and oligosaccharides (terminal carbohydrate unit of upper respiratory tract and lung) (Xu et al., 2008, Grienke et al., 2012, Webster et al., 1992, Nayak et al., 2004, Colman, 1994) and continuously released final infected progeny to infect more cells. This receptor destroying enzyme also facilitates in promoting penetrations of virus through human's epithelium airway (Xu et al., 2008, Grienke et al., 2012).

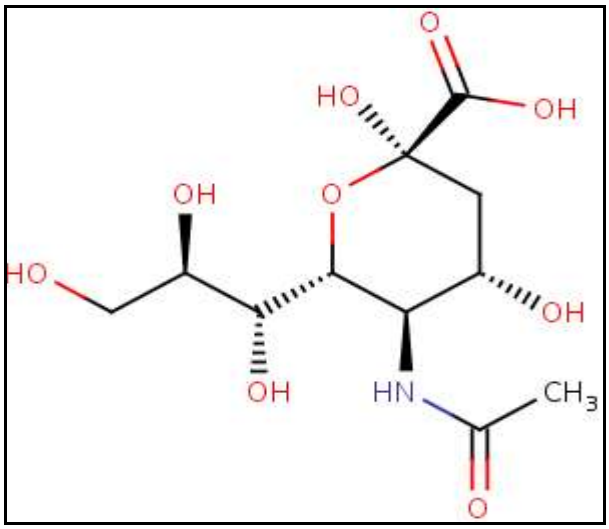


Figure 1.4: Sialic acid, *N*-acetylneuraminic acid, Neu5Ac

NA can be grouped based on antigenicity into 9 different serotypes N1 to N9 (Kamali and Holodniy, 2013, Colman, 1994). Within the different of NA serotype, they are further distinguished by two types of groups; 1) group 1: N1, N4, N5, N8 and 2) group 2: N2, N3, N6, N7, and N9 (Kamali and Holodniy, 2013, Webster et al., 1992, Vavricka et al., 2011). Generally, most researchers suggested that the group was categorised based on the so-called flexible '150- loop' on NA amino acid where group-1 indicates an open form of NA protein while group-2 is in a closed position of sialidase *apo* form structure (Rudrawar et al., 2010, Russell et al., 2006, von Itzstein, 2007). The open form causes the prominent formation of pocket located at 150-loop sialidase structure (Grienke et al., 2012).

General NA structure is composed of four identical subunits forming homotetramer of tetramic glycoprotein (Patrick, 2009, Colman, 1994, Grienke et al., 2012). The tetramer possessed mushroom-like shape which anchored to the viral membrane by a single hydrophobic sequence of some 29 amino acids near the stalk of N-terminus (Patrick, 2009, Colman, 1994, Grienke et al., 2012), which result in the capability of enzyme to be split enzymatically from the surface without loss of any antigenic activity (Patrick, 2009). The terminal hydrophobic sequence acts as

signal a for transport to the cell membrane and as transmembrane domain (Webster et al., 1992). The morphology of NA is basically composed of cytoplasmic domain, N-terminus domain and stalk region (Gubareva et al., 2000). The active site is placed at the center of each subunit with highly conserved residues (Varghese et al., 1983, Gong et al., 2007, Patrick, 2009).

NA structure fits its function. Briefly, other than cleaving terminal α-ketosidic linkage, NA also corresponds to the process of cell infection, cell replication, catabolism, signal transduction, and delivery of viruses into target cell (Gong et al., 2007, Magesh et al., 2006). Amino acids in NA catalytic site are highly conserved and less mutation occur inside the catalytic site compared to other sites in the enzyme. In contrast, chances of mutation in HA to occur is higher than NA which prompted most researchers to select NA as a drug target (Webster et al., 1992, von Itzstein, 2007) for further manipulation against mutated influenza virus.

1.2.4 Neuraminidase active site

Most of NA possessed common feature of catalytic site with similar interacting amino acids (Xu et al., 2008, Gong et al., 2007) (**Figure 1.5**) (Stoll et al., 2003) defines the common active site in influenza virus. Basically, the catalytic site is divided into 5 sites and denominated subsites as S1-S5. The labelled is in counter clockwise model by using the crystal structure of the substrate-based inhibitor DANA (Neu5Ac2en) embedded to the active site. DANA (2-Deoxy-2,3-Dehydro-N-Acetyl-Neuraminic Acid) is a transition-state analog formed of Neu5Ac.

Site S1 involved three residues in example of ARG118, ARG292 and ARG371. The residues provide a positively charged electrostatic and hydrogen bonding environment for anionic substituents from the inhibitor, such as carboxylate. Site S2 on the other hand consists of GLU119 and GLU227 which provide a negatively charged site of the active site. Site S3 contains a small hydrophobic region established from the residues of TRP178 and ILE222 near to the polar region of ARG152. Meanwhile, site S4 consists of hydrophobic region derived from the residues of ILE222, ALA246 and the hydrophobic face of ARG224. This site is not occupied by any portion of the Neu5Ac2en inhibitor. Site S5 is a mixed polarity region and comprised of the carboxylate of GLU276 (trans conformation) and the methyl of ALA246. GLU276 can exist as an alternative gauche conformation with its carboxylate ion-paired together with ARG224. When GLU276 is in this pose, its methylenes join with ALA246 to create a hydrophobic pocket in S5. Additionally, one more important residue in the catalytic site is ASP151 and it is not formally defined part in the subsites. While the residue's carboxylate does not make a direct contact with Neu5Ac2en, ASP151 is believed to play an important role in catalysis by polarizing the scissile glycosidic linkage (Stoll et al., 2003, Gong et al., 2007).

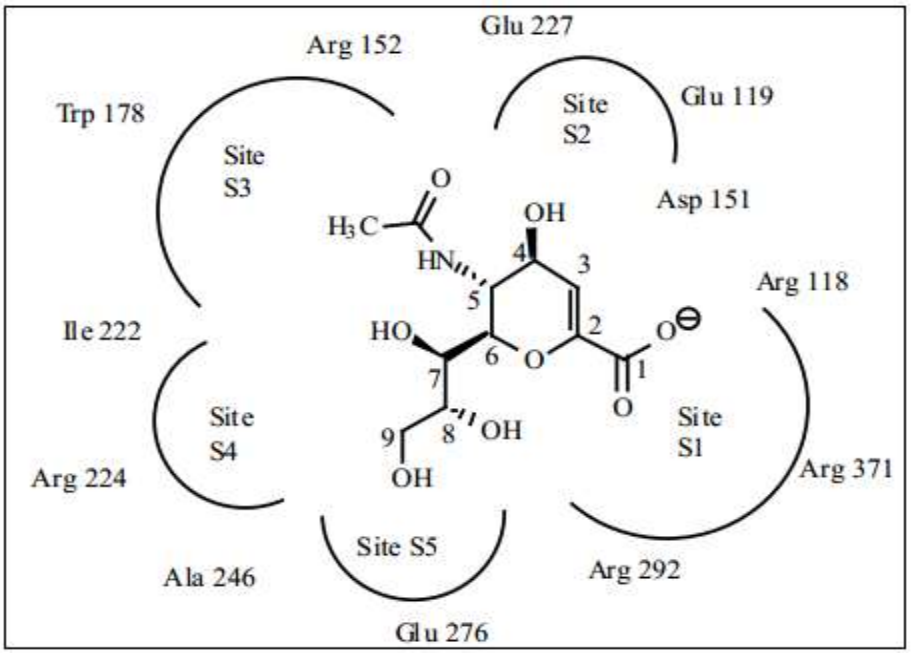


Figure 1.5: Diagram of NA active site with Neu5Ac2en (DANA) inhibitor (PDB: 1NNB) showing the general location of sites S1-S5 and important nearby residues (Stoll et al., 2003).

1.2.5 Mechanism of viral infection

Haemagglutinin (HA) initiates the viral infection via interactions with cell surface receptors which contain sialic- acid group (N-acetyl neuraminic acid (Neu5Ac)) (Melidou et al., 2010). The virus binds to α-2,3-linked and α-2,6-linked sialic acid receptors which are found on the surface of cells located deep in the epithelium airway (Maldonado et al., 2006, Matrosovich et al., 2004). Once HA is attached to the glycans containing the terminal sialic acid on the surface of the epithelial cells (at respiratory tract), it then allows the penetration of the virus into the host cell, by endocytosis where fusion of the viral and endosomal membranes allows the release of viral RNA. The RNA then will be transferred to the nucleus, where the viral genome is transcribed and amplified before translation of mRNA in the cytoplasm. Then, the new viral genome strains will be packaged with structural proteins in ribonucleoprotein particle (RNP). Afterwards, NA will cleave the terminal linkage of sialic acid receptors which result in the release of the progeny virions (viral genome) from the infected host cells (Whitford, 2005, Wang et al., 2010, Hu, 2010). **Figure 1.6** shows the life cycle of the influenza A virus and the pathway of viral entry and exit from host cells.