CHAPTER I

INTRODUCTION

The theme of this thesis relates to the research and development of a peptide database. In order to fully understand what follows, a review of some key concepts in bioinformatics is needed (section 1.1). Then the CU Peptide Database is briefly described (section 1.2). Finally, the objectives of the research is summarized (section 1.3).

1.1 Review of Key Concepts

1.1.1 Peptide

A peptide is composed of amino acid residues joined covalently through peptide bonds. Many hundreds of different peptides have been isolated from hydrolyzates or synthesized by chemical procedures. Peptides are also formed in the gastrointestinal tract during the digestion of proteins by proteases, enzymes that hydrolyzed peptide bonds (Stryer, 1988).

1.1.1.1 Peptide Bonds

Peptide bond is the CO-NH linkage resulting from α-amino acids polymerized through the elimination of a water molecule. Peptides are linear polymers. Each amino acid residue is linked to its neighbors in a head-to-tail direction rather than forming branched chains.

1.1.1.2 Acid-Base Properties

Amino acids and peptides have acid-base properties. The α-amino acids have two or, for those with ionizable side groups, three acid-base groups (Voet, 1990). The relationship between pH and the ratio of acid to base, known as Henderson - Hasselbalch equation, is shown below:

$$pH = pK + \log \frac{[A^-]}{[HA]}$$

where pK = -log dissociation constant, [HA] is the concentration of acid and $[A^-]$ is the concentration of conjugate base

1.1.1.3 Peptide Stability

A peptide structure is the result of a delicate balance among powerful countervailing forces, which are:

✦ *Electrostatic Forces*

Molecules are collections of electrically charged particles, so their interactions are determined by the laws of classical electrostatics. The energy of association, U, of two electric charges, q_1 and q_2, that are separated by the distance r, is found by integrating the expression for Coulomb's law, $F = kq_1q_2/Dr^2$, to determine the work necessary to separate these charges by an infinite distance:

$$U = \frac{kq_1q_2}{Dr}$$

which $k = 9.0 \times 10^9$ j.m.c^{-2} and D is the dielectric constant of the medium in which the charges are immersed.

✦ *Van Der Waals Forces*

Van der Waals forces are the noncovalent associations between electrically neutral molecules. They arise from electrostatic interactions

among permanent and/or induced dipoles. These forces are responsible for numerous interactions of varying strengths between nonbonded neighboring atoms.

✦ *Hydrogen Bonding Forces*

Hydrogen bonds are electrostatic interaction between a weakly acidic donor group (D - H) and an acceptor atom (A) that accepts a lone pair of electrons. In biological systems, D and A can both be the highly electronegative N and O atoms and occasionally S atoms. Hydrogen bonds have association energies in the range -12 to -30 $kj.mol^{-1}$.

✦ *Hydrophobic Forces*

The hydrophobic effect is the influence that cause nonpolar substances to minimize their contacts with water, and amphipathic molecules to form micelles in aqueous solutions. Kauzmann (1950) pointed out that hydrophobic forces are a major influence in causing protein to fold into their native conformations.

✦ *Disulfide Bonds*

Disulfide bonds function to stabilize its three-dimensional structure since they form as a protein folds to its native conformation.

1.1.2 Database

It is widely known that databases have played a critical role in almost all area where computers are used, including business, engineering, medicine, law, education, and science. A database is an *organized* collection of information, usually with one central topic (Schwartz, 1994). Every database is composed of records. A record is composed of fields, which contain all the information that has been collected on one individual subject in the database.

In a computer database, the program that allows users to create and maintain a database is called a database management system (DBMS). Therefore, the DBMS is a general-purpose software system that facilitates the processes of defining, constructing, and manipulating databases for various applications (Elmasri and Navathe, 1994). *Defining* a database involves specifying the data types, structures, and constraints for the data to be stored in the database. *Constructing* the database is the process of storing the data itself on some storage medium that is controlled by the DBMS. *Manipulating* a database includes such functions as querying the database to retrieve specific data, updating the database to reflect changes, and generating reports from the data.

A database can be roughly classified as flat-file database and relational database, according to the program's relational capabilities; that is, its ability to simultaneously draw information from more than one database on the basis of shared fields (Schwartz, 1994). A flat-file database always consists of a single file. All fields that are required have to be contained within that data file. A relational database consists of two or more interrelated data files that have one or more key fields in common.

1.1.3 Biochemical Database

In biochemistry, databases have been long used to collect and present biochemical information in systematic ways, as exemplified by GenBank and SWISS-PROT. Genbank, developed by U.S. Department of Health and Human Services, contains the collection of all DNA and RNA sequences (Burks et al., 1990). SWISS-PROT is a protein database developed by Department of Medical Biochemistry, University of Geneva, Switzerland (Barker, George, and Hunt, 1990).

These databases are developed not by biochemists who are involved with biopolymer sequencing but by people who collected the data from several other researches and make them available to the general public. On the other hand, there are special databases developed for specific purposes. Analyses performed with the help of these databases would yield useful information which is not possible to obtain otherwise. Such databases are often developed by experimental biochemists. For example, Stockwell's "A large Database DNA Sequence Handling Program with Generalized Searching Specifications" can find the overlap of DNA fragments and assemble the complete DNA sequence (Stockwell, 1982). Garrels and Franza's "The REF52 Protein Database", another example which is created for analyzing two-dimensional gel patterns of protein, contains a large amount of protein gel patterns and allows one to find the pattern of any particular protein (Garrels and Franza, 1989).

In general, a protein database always include sequence, references, taxonomic data, annotations, keywords, and cross-references to other biochemical database field. Annotations usually consist of biological functions, post-translational modifications, domains, quaternary structure, similarities to other proteins, diseases associated with deficiencies, results of site-directed mutagenesis studies, conflicts and variants. In addition, The database may have calculation tools for calculating some properties of the protein such as molecular weight, net charge and pI.

1.1.4 The Internet

1.1.4.1 Definition

The Internet is the world's largest computer network. It is comprised of thousands of individual networks in countries around the world, all

communicating via the Internet Protocol (IP) to exchange information (Wiederspan and Shotton, 1995). Each network operates independently for the most part, with independent sources of funding and sometimes even independent decisions on how best to implement the IP suit.

1.1.4.2 Internet Protocols

The Internet supports a wide variety of software *protocols*, or communication rules and structures, to exchange information between host and desktop (Savola, Westenbroek and Heck, 1995). These protocols often work together to complete the job. For instance, one protocol may manage the transfer of data between two points while another handles the format of the data and its use. These protocols are:

✦ *Telnet*

Telnet is a protocol for defining communications between computers. Using a Telnet connection, a user can issue commands on a remote computer and view the results as if logged directly into that computer (RFC 854, 1995).

✦ *Simple Mail transfer protocol (SMTP)*

SMTP is a protocol for defining methods for transmitting electronic mail between computers, via mail servers (RFC 821, 1995).

✦ *UseNet News (NNTP)*

NNTP is a protocol for defining the transmission of network news articles, which are much like e-mail messages but are directly at group server rather than at individual clients (Engst, 1994).

✦ *File transfer protocol (FTP)*

FTP is a protocol for transferring binary and ASCII text file to and from Internet-connected host and desktop computer systems.

✦ *Gopher*

Gopher is a protocol that uses a menu hierarchy to organize and access information from any Internet host running Gopher servers.

✦ *hyperText Transfer Protocol (HTTP)*

HTTP is a protocol spoken by WWW servers (HTTP servers). The advantage of HTTP is connections are made only at the instant that new information is needed and are completely dropped in between. Once a connection is dropped, the server has no memory that it ever existed (no state information is saved). This allows an HTTP server to handle many more connections than an FTP or Telnet server, where the connection is held open by the client until the user chooses to close it (Savola et al., 1995).

1.1.4.3 World Wide Web (WWW)

The World Wide Web is one of the fastest growing media for the transfer of multimedia information between computer users across the world (Shobe and Ritchey, 1995). It's an information system that links data from many different internet services under one set of protocols. Web clients (also called browsers or viewers) interpret documents containing HyperText Markup Language (HTML) and JavaScript delivered from Web servers. These documents use hypertext links to connect different documents and information resources together. Once the user clicks a link, the client software will retrieve the linked document or jump to a specific position in the current document.

✦ HyperText Markup Language (HTML)

HyperText Markup Language is a system for marking up documents with tags that indicate how text in the documents should be presented and how the documents are linked together (Aronson, 1995). It is a versatile language, allowing authors to format text and create relationships, called hypertext links, between text and another document or another portion of the same document. For example, the text phrase "CU Peptide Database" in a web document can be an available hyperlink that retrieves additional information about CU Peptide Database and displays it on the user viewers, regardless of what the original document was about (see figure 1-1).

Another capability of HTML is Form, which allows server to obtain information from the users (Aronson, 1995). Filling in the form and clicking the submit button will allow users to receive information that match their search criteria.

✦ JavaScript

JavaScript changes the passive nature of the Internet and World Wide Web by allowing platform-independent code to be dynamically loaded and run on a heterogeneous net work of machines, such as the Internet. JavaScript has many advantages. It can run on any machine that has the Java interpreter ported to it. It can protect the client from unintentional attacks such as viruses. It is object-oriented and dynamic. In addition, because it could be considered a derivative of C and C++, it is familiar to developers who currently use those languages.

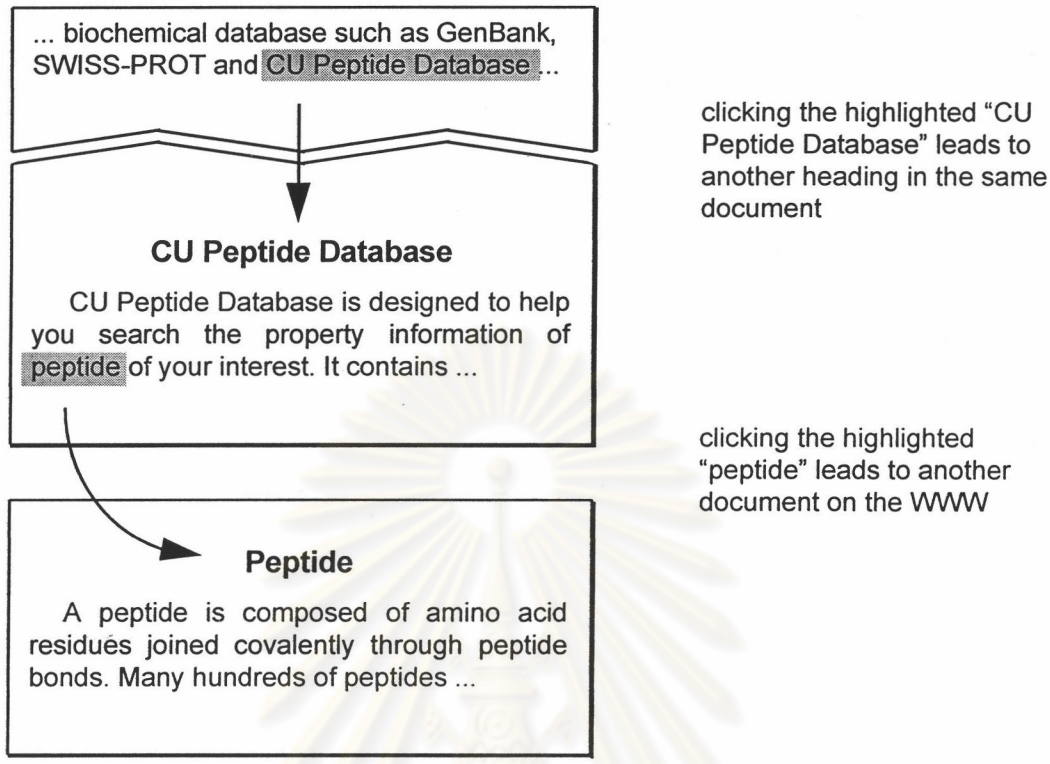

Figure 1-1 HyperText link

Hypertext link is a tool that links information together. It create relationship between topics in the same document or another document on the WWW. In a hyperlink, the text that provides the link should be related to the information it will access when the link is selected.

1.1.4.4 Common Gateway Interface (CGI) Applications

CGI applications are special applications that have an interface for communicating with HTTP servers (Wiederspan and Shotton, 1995). The purpose of these applications is to provide new capabilities for the HTTP servers. CGI application can be used as gateway between HTTP servers and other applications (such as database or text-searching software) or other services (such as on-line banking and purchasing systems).

1.2 CU Peptide Database Concept

In a research involving peptide fragments which non-naturally occurring but come from protein, enzyme, or hormone digestion, there are three methods to acquire the peptide. In the first method, the researcher extracts the peptide directly from such protein, enzyme or hormone (Graf and Li, 1974). In the second method, the peptide may be constructed by DNA recombinant techniques (Souza et al., 1984). Alternatively, it may also be constructed by peptide synthesis on a peptide synthesizer. The first two methods are complex and time consuming. The third method is expensive and difficult to local use, since there are only a few peptide synthesizers in Thailand. As such, peptide database should be created as a tool for giving researchers some data and predicting some property information of peptide before they do an experiment. In addition, peptide data are collected from chemical catalogues of several companies so that researcher can search the desired peptide and rapidly place an order from those companies.

According to the great development of information technology, people around the world can contact each other and can search data through the Internet. The Internet allows biochemists to access biochemical database such as GenBank or SWISS-PROT. So *CU Peptide Database* is also developed to be accessed via the Internet too, which involves 4 parts; a database program, CGI application, HTTP server and Internet browser. The four parts communicate to each other as shown in figure 1-2.

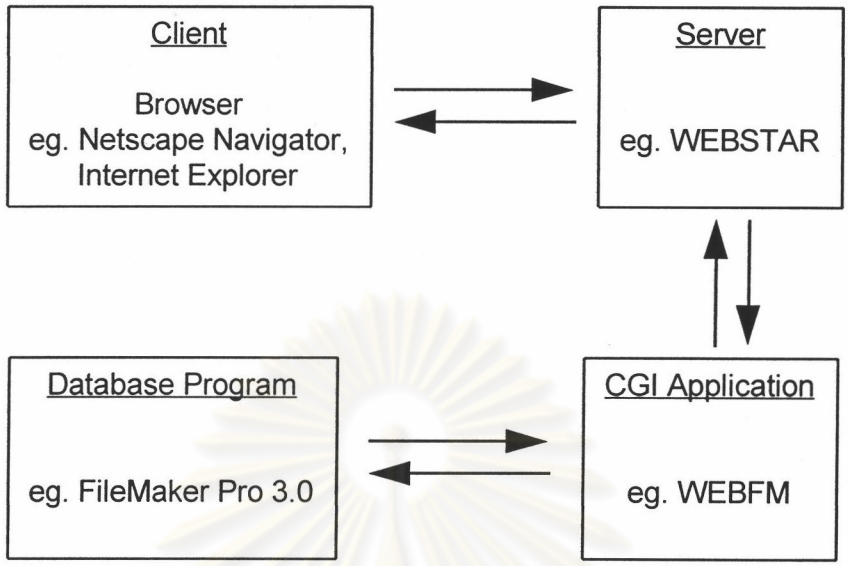

Figure 1-2 Communication among client, server, CGI application and database program

To make database accessible via the Internet, four pieces of software are required: (1) Internet browser such as Netscape Navigator or internet explorer, (2) HTTP server, (3) CGI application and (4) a database program. The four software communicate to each other as shown above.

1.3 Objectives of the Research

1. To complete the CU Peptide Database, with about 10 peptides in its original version (Tanasugarn et al., 1995)

2. To develop a search mechanism for properties of peptides such as molecular weight (MW), net charge at any pH, isoelectric point (pI), and hydrophobicity

3. To define cross references to other databases, for example, SWISS-PROT, PDB

4. To utilize the CU Peptide Database in testing a hypothesis by comparing calculated resulting parameter(s) from the CU Peptide Database with other data derived from experimental studies