Jurnal Teknologi, 34(C) 2001: 63–72

LITERATURE SURVEY OF PROTEIN SECONDARY STRUCTURE PREDICTION

SATYA NANDA VEL ARJUNAN[1], SAFAAI DERIS[2] & ROSLI MD ILLIAS[3]

Abstract. In the wake of large-scale DNA sequencing projects, accurate tools are needed to predict protein structures. The problem of predicting protein structure from DNA sequence remains fundamentally unsolved even after more than three decades of intensive research. In this paper, fundamental theory of the protein structure of the protein structure will be presented as a general guide to protein secondary structure prediction research. An overview of the state-of-the-art in sequence analysis and some principles of the methods invloved wil be described.

Key words: protein secondary structure prediction, neural networks.

Abstrak. Dengan wujudnya projek jujukan DNA secara besar-besaran, teknik yang tepat untuk meramalkan struktur protein diperlukan. Masalah meramalkan struktur protein daripada jujukan DNA pada dasarnya masih belum dapat diselesaikan walaupun kajian intensif telah dilakukan selama lebih daripada tiga dekad. Dalam kertas kerja ini, teori asas struktur protein akan dibincangkan sebagai panduan umum bagi kajian peramalan struktur protein sekunder. Analisis jujukan terkini serta prinsip yang digunakan dalam teknik-teknik tersebut akan diterangkan.

Kata kunci: peramalan stuktur sekunder protein, rangkaian neural.

1.0 INTRODUCTION

Proteins, the fundamental molecules of all organisms have three-dimensional structures that are fully specified by sequence of amino acids. The three-dimensional protein structure determines the functional properties of the protein. Proteins have many different biological functions; they may act as enzymes or as building blocks (muscle fibers) or may have transport function (for example, transport of oxygen). Determining protein structure from its amino acid sequence would greatly help understand the structure-function relationship. Hence, functions could be added or removed by changing their structure or synthesizing new proteins to obtain desired functions. For instance, by determining the structures of viral proteins it would enable researchers design drugs for specific viruses [1].

[1 & 2] Department of Software Engineering, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia, Phone: 60-7-556 6160 ext. 3752. Fax: 60-7-556 5044. Email: satyanandavel@hotmail.com

[3] Department of Bioprocess Engineering, Faculty of Chemical and Natural Resources Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia.

At present 100% accurate protein structures are determined experimentally using X-ray crystallograhic or Nuclear Magnetic Resonance (NMR) techniques. However, these methods are not feasible becasue they are tedious and time consuming, taking months or even years to complete [2]. In additions, large-scale sequencing projects (such as the Human Genome Project) produce protein sequences at very fast pace [3]. As a result, the gap between the number of known protein sequences (>150,000) [4] and the number of known structures (>4,000) [5] is getting larger. Protein structure prediction aims at reducing this sequence-structure gap. Until now however, the protein structure cannot be predicted 100% accurately theoretically. This is due to the fact that there are 20 different amino acids and thus there are too many ways in which similar structures can be generated in protein by different amino acid sequences [6].

2.0 PROTEIN STRUCTURE THEORY

There are four types of nucleotides (also called bases). They are Adenine (A). Thymine (T), Cytosine (C) and Guanine (G). Nucleotide sequences (also called DNA sequences) do not detemine the biological function of the the system. As mentioned previously, the functions are determined by the protein structure. There are three levels of protein structure: primary, secondary and tertiary structures which are one, two and three-dimensional respectively. The primary structure is the sequence of amino acids obtained from the nucleotide sequence. Table 1 lists the 20 animo acids

Table 1 The 20 amino acids and their corresponding nucleotide sequences

Amino Acid	**Tryptophan**	**Methionine**	**Tyrosine**	**Cystein**	**Phenylalanine**
Nucleotide Sequence	TGG	ATG	TAT TAC	TGT TGC	TTT TTC

Amino Acid	**Histidine**	**Glutamine**	**Asparagine**	**Lysine**	**Aspartic acid**
Nucleotide Sequence	CAT CAC	CAA CAG	AAT AAC	AAA AAG	GAT GAC

Amino Acid	**Glutamic acid**	**Isoleucine**	**Glysine**	**Alanine**	**Valine**
Nucleotide Sequence	GAA GAG	ATT ATC ATA	GGT GGC GGA GGG	GCT GCC GCA GCG	GTT GTC GTA GTG

Amino Acid	**Threonine**	**Proline**	**Serine**	**Leucine**	**Arginine**
Nucleotide Sequence	ACT ACC ACA ACG	CCT CCC CCA CCG	TCT TCC TCA TCG AGT AGC	TTA TTG CTT CTC CTA CTG	CGC CGC CGA CGG AGA AGG

Table 2 The three main classes of the amino acids

Class	Amino Acid
Hydrophobic (repels water)	Alanine, Valine, Phenylalanine, Proline, Methionine, Isoleucine, Leucine, Glycine
Charged Residues	Aspartic acid, Glutamic acid, Lysine, Arginine
Polar (Hydrophilic – attracted to water)	Serine, Threonine, Tyrosine, Histidine, Cysteine, Asparagine, Glutamine, Tryptophan

and their corresponding nucleotide sequences. Table 2 lists the three main classes of the amino acids. The secondary structure is the spatial relationship of amino acid residues that are close to one another in the primary structure. The tertiary structure is the spatial relationship of residues that are far apart in the primary structure.

2.1 PROTEIN SECONDARY STRUCTURE THEORY

The secondary structure has 3 regular forms: alpha (α) helices, beta (β) sheets (combination of beta strands) and loops (also called reverse turns or coils). In the problem of the protein secondary structure predictions, the inputs are the amino acid sequences while the output is the predicted structure (also called conformation, which is the combination of alpha helices, beta sheets and loops). A typical protein contains about 32% alpha helices, 21% beta sheets and 47% loops or non-regular structure.

Figure 1 shows a basic amino acid structure, which consists of an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom (H) and the side chain, all bonded to a carbon atom called alpha carbon (C_α). Each one of the 20 amino acids has the same structure except for its side chain. Peptide bonds join the carboxyl group of one amino acid to the amino group of another by eliminating water (H_2O). Figure 2 shows a peptide unit. A polypeptide is an unbranched structure of many amino acid sequence bonded with peptide bonds. An amino acid unit in the polypeptide chain is called a residue. The polypeptide chain starts at its amino terminus and ends at its carboxy1 terminus [6].

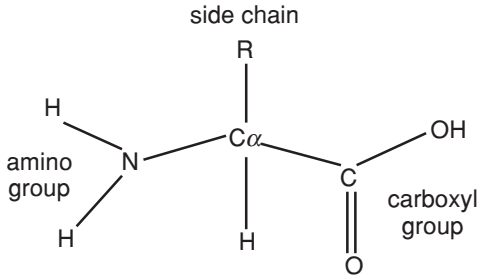

Figure 1 The basic amino acid structure

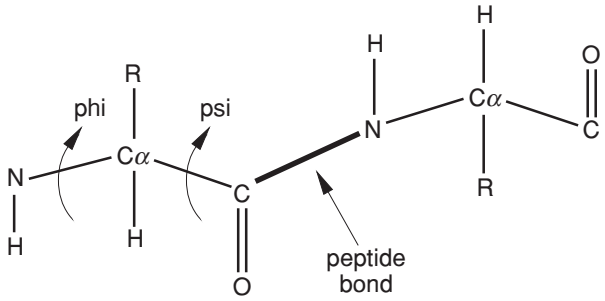

Figure 2 The peptide unit

Phi is the rotation angle around the N–C_α bond while psi is the rotation angle around the C_α–C bond. The rotations determine each protein's structure (i.e. alpha helix, beta sheet or loop). Amino acids in the interior of the protein molecule come from the hydrophobic class while amino acids from the polar class are at the surface of the molecule. Proteins evolved from a common ancestor are called homo-logous protein and they usually have similar amino acid sequences and conformations, and hence similar properties and functions. Researchers usually select nonhomologuos proteins from the protein data bank as working data for structure prediction research.

2.2 Characteristics of Alpha Helices

In alpha helices, all residues have phi and psi angle approximately –60° dan –50° respectively. There are 3.6 residues per turn (one turn/coil of helix). A hydrogen bond exists between the NH group of residue *n* and the CO group of residue *n*+3, which stabilizes the helix. Alpha helix has partial positive charge at the amino end and a partial negative charge at the carboxy1 end. This in turn causes both ends of alpha helices to be polar and therefore they are always at the surface of protein molecules. The lengths of alpha helices vary from 4 or 5 residues to over 40 residues. However, the average length is about 10 residues. The rise (height) per residue of alpha helix is 1.5Å along the helical axis. Side chains do not interfere with the alpha helix because they project out from it except for Proline where the last atom of the side chain bonds to the main chain N. Furthermore, Proline residue may cause a bend in an alpha helix. Alanine, Glutamic acid, Leucine and Methionine are good alpha helix formers while Proline, Glycine, Tyrosine and Serine are very poor alpha helix formers. In the protein molecule, alpha helices can be totally buried (all hydrophobic residues), partially buried (hydrophobic, polar and charged residues) and completely exposed (polar and charged residues). One problem is that short helices are difficult to predict [6].

2.3 Characteristic of Beta Sheets

Beta sheet is built from a combination of several polypeptide chains called beta strands. Beta strands are usually 5 to 10 residues long. They are aligned to each other such that hydrogen bonds can form between CO groups of one beta strand and NH groups on an adjacent beta strand and vice versa. Side chains point alternatively above and below the beta sheet. There are three ways to form beta sheet: parallel (all beta strands are in same direction), antiparallel (beta strands alternate in direction) and mixed (combination of parallel and antiparallel strands). However, mixed beta sheets occur rarely. All beta sheets have their strand twisted once. The twist always has the same handedness as shown in Figure 3, which is defined as right-handed twist.

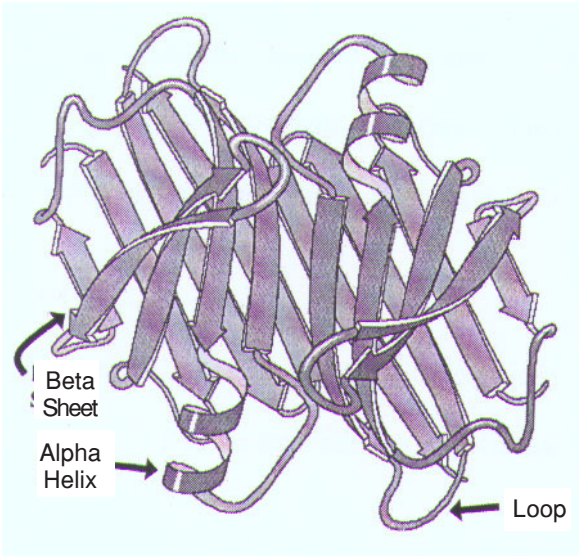

Figure 3 The protein secondary structure, which consists of alpha helices, beta sheets and loops [7]

2.4 Characteristics of Loops

Loop regions occur at the surface of the protein molecule. The main chain CO and NH groups of the loop regions, which generally do not form hydrogen bonds with each other, are exposed to the solvent and can form hydrogen bonds with water molecules. Loop regions exposed to solvent have large quantities of charged and polar hydrophilic residues. It is possible to predict loop regions with higher accuracy than alpha helices or beta sheets. In homologous amino acid sequences, it is found that insertions or deletions of a few residues occur almost only in the loop regions. This is because during evolution protein cores are much more stable than loops (which are at the surface).

3.0 SECONDARY STRUCTURE PREDICTION RESEARCH

The most widely used accuracy index for secondary structure prediction is the three-state perresidue accuracy (Q_3) which gives the percentage of correctly predicted residues in either of three states (classes), alpha helix, beta strand or loop region [8]

$$Q_3 = [P_\alpha + P_\beta + P_{\text{loop}})/ T] \times 100$$

where P_α, P_β and P_{loop} are number of residues predicted correctly in state alpha helix, beta strand and loop respectively while T is the number of residues.

There are three simple measures for assessing the quality of predicted secondary structure segment (or states): the number of segments in the protein, the average segment length and the distribution of the number of segments with length. Prediction methods need to meet four requirements. Firstly, no significant pair wise sequence identity between proteins used for training and test set (<25%). Secondly, all available unique proteins should be used for testing (since proteins vary in structural complexity, certain features are easier to predict than others). Regardless of which data sets are used for s particular evaluation, a standard set should be used for which results are also reported. Finally, test set should never be used before the method is set up [8].

The Protein Structure Prediction Center at Lawrence Livermore National Laboratory, California, USA occasionally organizes experiments on the Critical Assessment of Techniques for Protein Structure Prediction (CASP). It has held three such experiments since 1994. The goal is to obtain indepth and objective assessment of current abilities and inabilities in the area of protein structure predictions.

4.0 METHODS OF SECONDARY STRUCTURE PREDICTION

Basically there are three generations of secondary structrure prediction methods. Each new generations has overall accuracy of about ten percent higher than methods from previous generations [8].

4.1 First Generation

Most methods of the first generations were based on single residue statistics. From the limited proteins database (in 1960-1970s), evidence was obtained for the preference of particular residues for particular secondary structure states. The performance accuracy of these methods had been overstated, examples include Chou-Fasman algorithm [9] and GOR algorithm [10].

In Chou-Fasman algorithm, form the 15 amino acid sequence and the corresponding conformations known at the time (1974), Chou and Fasman computed frequencies with which each amino acid appears in alpha helices, in beta sheet and beta turns. For an alpha helix for example, each amino acid was classified as helix-

former, helix-neutral or helix-breaker based on the computed frequencies. The same was done for beta sheets and turns. Chou and Fasman used this information to predict statistically the secondary structures in others protein given their primary sequences. Their prediction was claimed to be 50–60 percent correct.

In GOR algorithm, the prediction of secondary structures is a way of assigning each residue in the primary sequence one of four states–alpha helix, beta sheet, beta turn or loop–and it is completely determined statistically by the residues within the same primary sequence. This algorithm gave a better prediction than Chou-Fasman method. The claimed Q_3 is 57.0% while CAP2 Q_3 is 54.4%.

4.2 Second Generation

The main improvement of the second generation of prediction techniques was a combination of a larger database of protein structures and the use of statistics based on segments. The accuracy levels were slightly higher than 60%. Mainly used algorithms were based on statistical informations, physico-chemical properties, sequence patterns, multi-layered neural networks, graph-theory, multivariate statistics, expert rules and nearest-neighbour algorithms.

In GOR3 algorithm [11], the framework of information theory provides a mean to formulate the influence of local sequence upon the conformation of a given residue. The first-level approximation drawn from the theory, involving single-residue parameters, marginally improved (compared to GOR algorithm) by an increase on the database. The second-level approximation, involving pairs of residues, provides a better model. This new version of the GOR method claimed Q_3 is 63.0%.

Qian and Sejnowski used neural network based algorithm to predict the secondary structure in 1988 [12]. They used the back propagation algorithm to predict the alpha helix and beta sheets of 15 test protein. The neural they used had three layers and with an optimal number of 40 hidden units and 13 input residues. They added a second network whose inputs were sequences of output from the first network. The claimed Q_3 is 64.3%.

The first and second generation methods shared at least two of the following problems (mostly all three) i.e. three-state per-residue accuracy was below 70%, beta strands were predicted at levels of 28 – 48% and predicted helices and strands were too short [8]. The first problem (<100% accuracy) may have arisen from two sources, i.e. secondary structure differ even between different crystals of the same protein, secondary structure formation is partially determined by long range interactions, i.e. contacts between residues that are not visible by any method based on segments of 11 – 21 adjacent residues. The second problem (beta strands <50% accuracy) is caued by the fact that beta sheet formation is determined by more non-local contacts than is alpha helix formation.

4.3 Third Generation

Method in the third generation are superior (in terms of accuracy) to their predecessors becouse information from homologous sequences is used. In addition, problems of first and second generation methods were addressed.

In Zvelebil algorithm, information available from a family of homologous sequences is used [13]. The approach is based on averaging GOR secondary structure preferences for aligned residues and on the observation that insertions and high sequence variability tend to occur in loop regions between secondary structures [10]. As a result, a statistical algorithm first aligns a family of sequences and a value for the extent of sequence conservation at each position is obtained. This value modifies GOR prediction on the averaged sequence to yield the improved prediction. This algorithm claimed Q_3 to be at 66.1%.

DSC algorithm ued linear statistics in its implementation. It identifies residue conformational propensities, sequence edge effects, moments of hydrophobicity, position of insertions and deletions in aligned homologous sequence, moments of conservation, auto-correlation, residue ratios, secondary structure feedback effects, and filtering [14]. It use simple nd explicit structure of the prediction, which allows the method to be reimplemented easily. This algorithm claimed Q_3 is 70.1% while CASP2 Q_3 is 69.5%.

In PREDATOR algorithm, the secondary structure prediction is based on local pairwise alignment of the sequence to be prediction with each related sequence rather than utilization of a multiple alignment [15]. Secondary structure propensities are based on both local and long-range effects, utilization of similar sequence information in the form of carefully selected pairwise alignment fragments, and reliance on a large collection of known protein primary structures. Its claimed Q_3 is 75%.

NNSSP is program tht predicts secondary structure based on neural networks and nearest-neighbour techniques [16]. The main idea of the nearest-neighbour approach is the prediction of the secondary structure state of the central residue of a test segment, based on the secondary structure of similar segments from proteins with known three-dimensional structure. The information coming from the different templates is scored according to their similarity (according to the sequence or other properties) with the test segment. NNSSP is an enhancement of the algorithm designed by Yi and Lander, which selects the neighbours by the mean environmental score and combines by the mean of neural network predictions made with different parameters (environmental scores, length of nearest-neighbours) [17}. In addition, it incorporates information coming from multi ple aligned sequences (by averaging their score for the weighting of each nearest-neighbour). It claimed Q_3 is 72.2% while it CASP Q_3 is 67.7%.

PHD is a program that is composed of everal cascading neural networks (previously trained on proteins of known structures) [18]. A first network takes as input a set of vectors representing the sequences present in a window sliding along the

multiple alignment. Its output is composed of a vector representing the probabilities for each of the three states of the residue central to the window. Since the secondary structure of a residue is not independent to that of its neighbours, second step takes into account these local interactions. A neural network takes a input the vectors present in a window sliding along the previous output. Its output is a refined three-states probabilities vector. Another step consist of averaging (for each state) the outputs from independently trained networks. Finally a "winner take all" decision assigns the secondary structure state. No explicit rules are included in the algorithm. PHD my generate its own alignment with the submitted sequence. PHD's claimed Q_3 is at 72.2% while its CASP2 Q_3 is 71.6%.

5.0 CONCLUSIONS

Looking at the recent and more accurate algorithms, it is evident that they are based on neural networks. Hybrid techniques incorporating neural networks and other such as expert systems or genetic algorithms would be a good ground for further research, as there has not been much work done on this area.

After more than three decades of research, theoretical biology can still not predict protein structure from DNA sequence with 100% accuracy. Nevertheless, most breakthrough in protein structure prediction were achieved over the last seven years. Hence, although general prediction problem cannot be solved, significant progress has been made. Only continued perseverance in structure prediction research can contribute to a better accuracy in the prediction.

6.0 ACKNOWLEDGMENTS

The author would like to thank Dr Zaharah Ibrahim and Dr Zainoha Zakaria for reviewing this paper.

REFERENCES

[1] Zhang X. 1994. A Hybrid Algorithm for Determining Protein Structure. *IEEE Expert*. Vol. 94, p.: 66 – 74.

[2] Metfesel B. A., and Saurugger P.N. 1993. Pattern Recognition in the Prediction of Protein Structural Class. *Proceedings of the Twenty-Sixth Hawaii International Conference on System Sciences*. Vol. 1, p.: 679 – 688.

[3] Fleischmann R. D., Adams M. D., White O., Clayton R. A., Kirkness E. F., Kerlavage A. R., Bult C. J., Tomb J. F., Dougherty B. A., Merrick J. M., McKenney K., Sutton G., Fitzhugh W., Fields C. A., Gocayne J. D., Scott J. D., Shirley R., Liu L. I., Glodek A., Kelly J. M., Weidman J. F., Phillips C. A., Spriggs T., Hedblom E., Cotton M. D., Utterback T. R., Hanna M. C., Nguyen D. T., Saudek D. M., Brandon R. C., Fine L. D., Fritchman J. L., Fuhrmann J. L., Geoghagen N. S. M., Gnehm C. L., McDonald L. A., Small K. V., Fraser C. M., Smith H. O., Venter J. C. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, Vol. 269, p.: 496 – 512.

[4] Bairoch A., and Apweiler R. 1996. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucl Acids Res*. Vol. 24, p.: 21 – 25.

[5] Bernstein F. C., Koetzle T. F., Williams G. J. B., Meyer E. F., Brice M. D., Rodgers J. R., Kennard O.,

Shimanauchi T., and Tasumi M. 1977. The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol*. Vol. 112, p.: 535 – 542.

[6] Branden C., and Tooze J. 1991. *Introduction to Protein structure. Garland Publishing*. Inc. p.: 3 – 29.

[7] Protein Structure Image Gallery at Biochemistry Department Website of Duke University. USA. (http:// kinemage. biochem.duke.edu/).

[8] Rost B., and Sander C. 1999. Third generation prediction of secondary structure. *Protein Structure Prediction: Methods and Protocols*. Humana Press. p.: 71 – 95.

[9] Chou P. Y., and Fasman G. 1974. Conformational parameter for mino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*. Vol. 13, p.: 211 – 222.

[10] Garnier J., Osguthorpe D. J., Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol*. Vol. 120, p.: 97 – 120.

[11] Gibrat J. F., Robson B., and Garnier J. 1987. Further development of protein secondry structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol*. Vol. 198, p.: 425 – 443.

[12] Qian N., and Sejnowski T. J. 1988. Predicting the Secondary Structure of Globular Proteins using Neural Network Models. *J. Mol. Biol*. Vol. 202. p.: 865 – 884.

[13] Zvelebil M. J., Barton G. J., Taylor W. R., Sternberg M. J. E. 1987. Prediction of protein secondary structure and active site using the alignment of homologous sequences. *J. Mol. Biol*. Vol. 195, p.: 957 – 967.

[14] King R. D., Sternberg M. J. E. 1996. Indentification and application of the concept important for accurate and reliable protein secondary structure prediction. *Prot Sc*. Vol. 5, p.: 2298 – 2310.

[15] Frishman D. and Argos P. 1997. Seventy-five percent accuracy in protein secondary structure prediction. *Protein*. Vol. 27, p.: 329 – 335.

[16] Salamov A. A., and Solovyev V. V. 1995. Prediction of protein secondary structure by combining nearest-neighbour algorithms and multiple sequence alignment. *J. Mol. Biol*. Vol, 247, p.: 11 – 15.

[17] Yi T. M., and Lander E. S. 1993. Protein Secondary Structure Prediction Using Nearest-Neighbour Methods. *J. Mol. Biol*. Vol. 232, p.: 1117 – 1129.

[18] Rost B. and Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol*. Vol. 232, p.: 584 – 599.

Satya Nanda Vel is a Master student at the Faculty of Computer Science and Informtion Systems, Universiti Teknologi Malaysia. He received a B.Eng. In Electronics Engineering from the same university in 2000. His research interests include bioinformatics, genomics, intelligent agents and neural networks.

Safaai Deris is an Associete Professor at the Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia. He is a member of the IEEE, AAAI, Management Science and Operations Research Society, Malaysia (MSORSM), and Artificial Intelligence Society, Malaysia. Prior to joining the univesity he was a system analyst at Ministry of Agriculture, Malaysia for 10 years. He received a B.Sc. in 1978 from the Agriculture University of Malaysia, an M.Eng. in Industrial Engineering in 1989 from the University of Osaka Prefecture, Japan and a D.Eng. in Computer and System Sciences from the same university in 1997. His research interests include intelligent agents and applications of constraint-based reasoning, genetic algorithms, and fuzzy systems in planning and scheduling.

Rosli Md Illis is a Bioprocess Engineering lecturer at Faculty of Chemical Engineering and Natural Resources Engineering, University Teknologi Malaysia. He received a B.Sc. in Microbiology in 1992 from Nationl University of Malaysia and Ph.D. in Genetic Engineering from University of Edinburgh, Scotland in 1997. His research interests include biotechnology, molecular biology including gene manipulation, protein structure and biochemistry in industrial enzyme production.