Predicting, Designing and Altering Protein Function and Specificity Using Computational Approaches

Summary:

A fundamental and demanding biological challenge is to address the question of how the linear information encoded in the genome affects a biological process. Toward this broad goal, understanding how and why proteins perform function becomes vital. Although proteins with similar structures are likely to have similar functions, small differences between globally similar structures express variations in the specificity and regulation. In the case of enzymes, it is widely accepted that the flexible loop regions have a critical functional role. Lack of knowledge about the binding site flexibility has led to failures both in predicting protein function and in protein-ligand docking. This unsolved problem demands a quick solution as current structural genomics efforts are set to markedly increase the number of structures of proteins with unknown functions.

The research plan described here centers around the theme of predicting, designing and altering protein function and specificity. The plan would employ a range of computational techniques from staring at the protein structures to modeling structures to performing sophisticated analyses on genome wide sequences both at the molecular and biological systems levels. Further, collaboration would be established with experimental laboratories to effectively design and alter protein functions on specific model systems of mutual interest. This would enable an understanding of the origin of functional diversity as well as the mechanisms by which protein structure determines functional specificity.

Specifically, the problem would be tackled through, but not limited to, (i) bioinformatics analysis on protein sequences and structures at the molecular and genome wide levels to identify potential correlations between evolutionary conservation, binding site flexibility and function. This would include developing and analyzing databases of apo and holo structures and of proteins, which pose problems to function prediction and the understanding of disease mechanisms, such as ‘natively unstructured’ or ‘disordered’ proteins and proteins performing multiple functions (‘moonlighting proteins’); (ii) Carrying out parallel tempering molecular dynamics simulations to effectively sample the loop conformations with the aim to understand and predict the flexibility at the binding site; (iii) Elucidating the structure, function, and specificity of proteins of interest to collaborative experimental research to test hypotheses and generalize functional mechanisms in the context of cellular function and biological process; and (iv) A systems biology approach to understand the network of interactions, feedback mechanisms, and how and why proteins interact.

Specific Aim:

The research plan focuses on addressing the question of how function and specificity is encoded in the sequence and structure of the proteins. The goal is to infer general rules and develop algorithms that have predictive power and can be used for the rational design and alteration of protein function and specificity. The following questions would be tackled:

- How is functional specificity incorporated into protein sequences and structures?
- What amino acid sequence or structural modifications need to be done in order to increase or decrease the functional specificity?
- How can the binding site flexibility or conformations sampled by functional loops in the biological time scale be explored and predicted?
- What knowledge based rules can be learned from the available sequence and structural databases to effectively design functional motifs, such as the zinc-finger?
- How do some proteins perform multiple, seemingly unrelated, functions ('moonlighting proteins') and how do they evolve?
- How do 'disordered' proteins, which are structured only upon binding to their interacting partner molecule(s), perform their functions?
- How can we predict the interacting partner molecule(s) of a given 'disordered' protein from its sequence?

Significance:

Proteins act together to perform function. It is essential to have a coherent picture of molecular and organismal structure, function, evolution, and interaction in order to understand and develop effective and target based therapeutic drugs. Insights resulting from bioinformatics approaches, such as addressing the questions described above are essential for biologists to pose and answer precise scientific questions about systems and organismal biology. The development of tools will have practical utility for pharamcogenomics and genetic engineering.

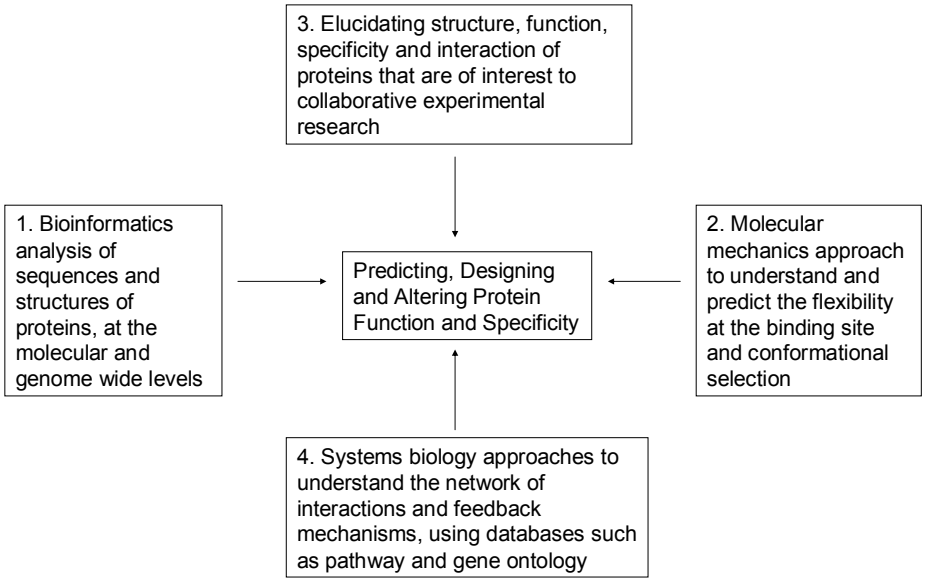
Designing and altering protein function strategy could be used for varied applications such as changing the specificity and efficiency of existing enzymes or regulating activity. Manipulation of protein function has many biotechnological applications, including the construction of enzymes, biosensors, genetic circuits, signal pathways, and chiral separations (Koh, 2002; Looger et al., 2003; DeGrado, 2003). The ability to design catalyst with enzyme-like properties could result in materials that could be used in innumerable applications (Tann et al., 2001). For example, chemical modifications of enzymes has been shown to be useful for preparing biosensors that couple the response to external stimuli, such as light and specific metal ions, to a perturbation in enzyme activity (Hamachi et al., 2001). Design of biocatalysts demands ability to engineer specific interactions, such as C-H---O hydrogen bonds, between the protein and its cognate substrate.

Protein-ligand engineering is a tremendously important issue with regard to the creation of new tools for the manipulation and study of biological systems. The technique may ultimately be applied to the design of custom pharmaceuticals for certain genetic disease due to the fact that many human genetic diseases are associated with mutations to receptors that impair ligand binding. For instance, Ye et al. (2001) demonstrated that a thyroid hormone mimic could be designed to complement a mutationally impaired receptor associated with a human genetic disease. For nearly two decades, chemical and genetic methods for manipulating molecular function have been used to alter enzyme substrate specificity or to generate new ligand-receptor pairs that can selectively regulate transcription, apoptosis, genetic recombination, signal transduction, or motor protein function (Koh, 2002).

Recent studies on small molecules influencing protein-protein interactions are even more encouraging. These studies suggest that the scope of ligand-receptor engineering may extend to the control of the assembly of protein-protein interfaces, leading to the potential treatment of certain human genetic disease (Tian et al., 1998).

Understanding and predicting binding site flexibility has direct implications in predicting protein function and in developing scoring functions for protein-ligand docking algorithms. Currently, docking and binding site prediction algorithms have met with failures mainly due to the difficulty in handling binding site flexibility (Carlson, 2002). In cases where even moderate conformational changes take place upon ligand binding, it has proven to be a difficult task for docking algorithms to work successfully.

The following multiple and redundant approaches would be adopted to address the questions



Approach 1: Bioinformatics analysis of sequences and structures at the molecular and genome wide levels to probe conformational change upon ligand binding and protein function

Bioinformatics addresses biological questions through the integration of computer software tools, databases and analyses. A comprehensive structural database consisting of ligand bound and unbound protein structures determined through X-ray crystallography and Nuclear Magnetic Resonance would be developed. The database would be designed to provide the details of the conformational changes taking place at the side chain and backbone levels upon ligand binding. The type of ligand considered would be from small molecules to domains to proteins. This will be done by going through the published structures and various structural databases such as Protein Data Bank. The aim here is to have a well curated database that relates to details of the experiments and the functional network of biological process.

Additionally, for all the members present in the above structural database, multiple sequence alignment would be carried out using the sequences obtained from a BLAST search and from the PFAM database at the family, super-family and fold levels. The sequence database thus obtained would be analyzed for evolutionary pressure, such as amino acid conservation, using traditional sequence conservation analysis methods, evolutionary trace algorithms and a phylogenetic tree scheme. The results would be combined with analysis of the network of residue-residue interactions in the structure to understand and predict how short and long range communication of interactions modulate binding site flexibility or conformational changes.

Further, proteins that have unique functional features and that provide a challenge in the understanding of disease mechanisms, such as proteins that perform multiple functions and disordered proteins, would be cataloged and analyzed. This would broaden the understanding of mechanisms by which proteins act. The increasing number of single protein that can 'moonlight' or perform multiple, apparently unrelated, functions challenges the interpretation of the human genome sequence and the annotation of protein sequence databases (Jeffery, 2003a,b). Natively unstructured or disordered proteins pose a problem to structural genomics initiatives as they do not crystallize without their partner molecule(s) (Dunker et al., 2001). The natively unstructured proteins are structured only upon binding to their partner molecule. The lessons and rules learned would be used in various predictions, such as predicting the partner molecule of a given disordered protein.

Approach 2: Molecular mechanics approach to sample loop conformations at or near biological time scale to foresee the conformational flexibility at the binding site

Loop flexibility in enzymes plays a vital role in correctly positioning catalytically important residues. This strong relationship between enzyme flexibility and function provides an opportunity to engineer new substrates and inhibitors. It further allows the design of site-directed mutagenesis experiments to explore enzymatic activity through the control of the flexibility of functional loops (encompasses the binding site and undergoes conformational change upon ligand binding). Molecular dynamics (MD) simulations would be carried out to understand the relationship between the evolutionary pressure or sequence conservation and the mechanism of conformational change upon ligand binding. The recently discovered parallel tempering or replica exchange method allows one to carry out long time MD simulations to effectively sample all minima in the energy landscape (Pitera and Swope, 2003). The parallel tempering method would be used to explore the flexibility of the functional loop at or near biological time scale. Correlations would be established between the residues involved in the dynamics and those that are evolutionarily conserved. The significance of the backbone conformational freedom of the residues involved in correctly positioning the catalytically important residues at the binding site would be assessed through this approach.

Approach 3: Collaborate effectively with experimental laboratories to test hypothesis and to design and alter protein function on specific model systems of mutual interest

A combination of theoretical and experimental approaches is the most rewarding way to predict, design, and alter protein function and specificity. A relentless effort will be made to establish collaboration with experimental labs, where the aim of this research plan ties together the collaborators' interests. Therefore, the choice of the system in this approach will depend on the collaboration. The aim here is to put the results from a specific case into the context of cellular function and biological process. Collaborations will be established within as well as outside the hiring institution. Once the system is identified, a battery of approaches, as described here, would be used towards understanding the functional mechanism and specificity. In particular, bioinformatics analysis, and detailed parallel tempering MD simulations would be carried out to understand how the flexibility is incorporated into the polypeptide chain. The roles played by residues to help position the amino acids important for activity and binding would be examined thoroughly.

A further step would be taken towards putting the choice case in the context of cellular environment and identifying potential applications in disease, genetic disorders, etc. This will be done through bioinformatics analysis, pathway examination, and protein-protein interactions map studies. Essentially, the successful collaboration will lead to the development of a coherent picture of a biological process in which the specific case would be involved.

Approach 4: Systems biology approach to understand the network of interactions and how and why proteins interact

In contrast to the conventional biological approach of looking at a particular protein and its interaction and pathway, systems biology approach looks at the whole organism and strives to understand how different parts work together. In addition to examining the currently available databases such as KEGG pathway, gene ontology, and many others, a customized protein interactions network database would be developed. The database would be built to suite the systems under investigation such as on those where collaboration with experimentalists has been established. The details of the interactions would be extracted from the literature using an automated algorithm and manually curated. The database would contain additional information like, pathways, diseases, molecular summary and much more. The key features include the validation tool, relational database architecture and Java based interactive visualization tool. This database would be used to examine how altering the function of a protein affects a biological process in the network of interactions. The information obtained would be used to learn about the feedback mechanisms in the systems, and to build models and test hypotheses.

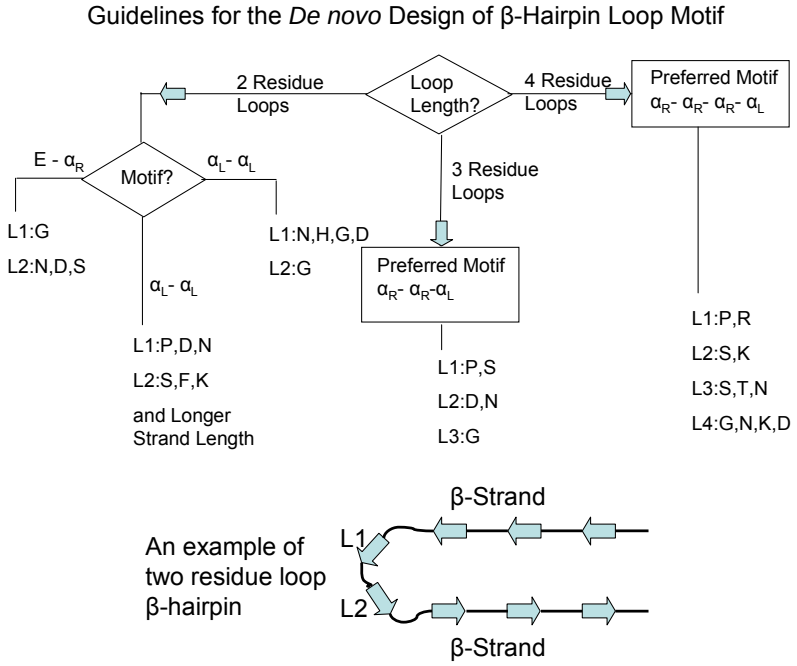
Further, the database of interacting proteins (DIP) together with the gene ontology (GO) database provides enormous opportunity to examine highly connected partner proteins *versus* those that connect to only a few. In other words, a systematic comparison could be made between the peripheral and centrally located proteins in the network of protein-protein interactions. This would lead to understanding of the mechanism by which some proteins are able to interact with many proteins in contrast to those that interact with only a few.

Appendix

Past Experience and Preliminary Work Supporting the Approaches

Approach 1: Bioinformatics and database analyses to derive rules for designing mini proteins and functional motifs

De novo protein design is an attractive approach for studying the structure and function of proteins. *De novo* design confronts the issue of how to specify a protein's fold and function. *De novo* design of several functional motifs, such as the EF hand and zinc-finger, has proven to be a valuable tool for biological research and gene therapy. The *de novo* designed motifs and proteins have potential applications in receptors, enzymes, and ion channels and biomimetic polymers with properties unprecedented in nature. The design also provides a stringent test of our understanding of enzyme function. The success of *de novo* design relies heavily on the ability to design relatively short stretches of polypeptides that can adopt stable secondary structures. Analysis carried out on 250 non-homologous high-resolution protein crystal structures, with a view towards expanding the scope and nature of the connecting loops, suggested that the rational design of loops > 2 residues may indeed be possible (shown in the diagram; Gunasekaran et al., 1997; 1998). The rules learned have been successfully applied in the experimental design of turns in β-hairpin motifs (for example see Das et al., 2001).



Legend: Analysis of high-resolution structures reveals preferred motifs, with frequently occurring amino acids, connecting anti-parallel β-strands. Three conformational motif types are preferred for two residue loops, while only one motif type is frequently found for three or four residue loops. E – Extended conformation, α_R - right-handed helical conformation, α_L - left-handed helical conformation; L1, L2, L3, L4 refers to position in the loops; an example of two residue loop is shown at the bottom.

Further analysis revealed that the conformational variations in β-turns are indeed observed in protein crystal structures and such changes may be an important dynamic feature in solution, such as in the case of HIV-1 protease (Gunasekaran et al., 1998).

In yet another example of the analysis of structural database leading to implications for *de novo* design, analysis of ordered and disordered protein complexes led to identification of striking structural features discriminating stable and unstable monomers (Gunasekaran et al., 2004b). The study made use of the fact that natively unstructured proteins, which undergo a disorder-to-order transition upon binding their partner, and stable monomeric proteins that exist as dimers only in their crystal form, provide examples of two vastly different scenarios. The analysis showed that ordered monomers can be distinguished from disordered monomers based on the per-residue surface and interface areas, which are significantly smaller (< 80 $Å^2$) for ordered proteins. Significantly, the chemical characteristics (polar and non-polar solvent accessible area composition) at the surface showed a large difference between the ordered and disordered proteins. In contrast to the disordered proteins, the ordered proteins clustered narrowly around 0.5 in the fraction of polar or non-polar exposed surface area composition. This observation suggests that stable proteins are designed to balance the polar and non-polar composition at the exposed surface equally. This finding is useful in *de novo* protein design where an amino acid sequence intended to fit a structural model could be examined for the exposed surface area composition through model building exercises.

Approach 2 & 3: Conservation analysis and MD simulations on a model system to explore functional loop flexibility and mechanism – an experimental and computational study

Loop flexibility in enzymes plays a vital role in correctly positioning catalytically important amino acids for enzymatic reaction. β1,4-galactosyltransferase (β4Gal-T1) has been shown to undergo a large conformational change to create binding sites for oligosaccharides and α-lactalbumin. Comparison of the substrate bound and unbound crystal structures reveals a large conformational change (displacement of up to 20Å) in a long loop comprising residues Ile345 to His365 (Ramakrishnan et al., 2004). This model system was identified through collaboration with an experimental group headed by Dr. Pradman Qasba at the National Cancer Institute, Frederick, MD.

Molecular dynamics (MD) simulations of the wild-type β4Gal-T1, carried out with an implicit solvent model and with explicit water, identified a high degree of flexibility in the long loop region, His347 to Glu368, and in a smaller loop, residues Tyr311 to Gly316, which contains a conserved tryptophan residue, Trp314 (Gunasekaran et al., 2003a). Further analysis of covariance of spatial displacement of the residues revealed that coupled motions occur between the residues in these two loops (Gunasekaran et al., 2004a). Thus the computational work pointed to correlation between the residues involved in the dynamics and in the conservation. This led us to hypothesis that the Trp314 residue was influencing the flexibility of the long loop. Consistently, an experimentally designed Trp314Ala mutant, which folds properly *in vitro*, lost nearly 99% of the wild-type catalytic activity (Ramasamy et al., 2003; Ramakrishnan et al., 2004). Importantly, the long loop conformation was significantly affected. Further, when Trp314 is mutated to Ala314, the long flexible loop of the Trp314Ala

mutant remains in a nearly open conformation and is more accessible to limited proteolysis by Glu-C or Lys-C proteases, indicating that Trp314 motion plays a vital role in the overall motion of the long flexible loop.

The observation made in the case of β4Gal-T1 was later extended to two other proteins, enolase and lipase (Gunasekaran et al., 2003b). The correlation between the positions involved in the dynamics and conserved residues was seen in all three cases: β4Gal-T1, enolase and lipase. Hence, evolution appears to select residues that drive the functional loop to a large conformational change. These observations suggest that altering the selected loop-loop interactions might modulate the movements of the functional loop.

Approach 4: Systems biology: Analysis of central *versus* peripheral proteins in the network of protein-protein interactions

In order to understand the function of multi-protein complexes, or whole proteomes, it is vital to have a model as proteins in a cell do not work in isolation. A model will have to be developed to enable us to observe how and why different proteins interact with each other and other molecules to carry out their respective functions in the cell. In order to understand the feedback mechanisms, such a model must be dynamic and not provide us with a static picture. In this regard, it is important to understand the difference between proteins that interact with many partners and those that interact with only a few. Fortunately, the DIP and GO databases are rich in information about connectivity and function. A preliminary analysis of central (high connectivity; many partners) and peripheral (low connectivity; few partners) proteins from *yeast* do not reveal any sequence and structural features discriminating them. There is no correlation between the connectivity *versus* sequence length, number of domains, secondary structure content, conserved surface patches, etc. In other words, there are no clear structural markers for high or low inter-molecular connectivity of the proteins. However, the analysis shows that the binding sites of highly connected proteins can be of very different sizes and similar folds can accommodate high or low connectivity. This work is currently under progress to further quest for the answer.

References:

Carlson HA (2002) Protein flexibility is an important component of structure-based drug discovery. *Curr Pharm Des* **8**, 1571-8.

Das C, Naganagowda GA, Karle IL, and Balaram P (2001) Designed β-hairpin peptides with defined tight turn stereochemistry. *Biopolymers* **58**, 335-46.

DeGrado WF (2003) Computational biology: Biosensor design. *Nature* **423**, 132-3.

Dunker AK, Lawson JD, Brown CJ, et al. (2001) Intrinsically disordered protein. *J Mol Graph Model.* **19**, 26-59.

Jeffery CJ (2003) (a) Multifunctional proteins: examples of gene sharing. *Ann Med.* **35**, 28-35. & (b) Moonlighting proteins: old proteins learning new tricks. *Trends Genet.* **19**, 415-7.

Gunasekaran K, Ramakrishnan C and Balaram P (1997) β-hairpins in proteins revisited: Lessons for *De novo* Design. *Protein Engg* **10**, 1131-1141.

Gunasekaran K, Gomathi L, Ramakrishnan C, Chandrasekhar J and Balaram P (1998) Conformational interconversions in peptide β-turns: Analysis of turns in proteins and computational estimates of barriers. *J Mol Biol* **284**, 1505-1516.

Gunasekaran K, Ma B, Ramakrishnan B, Qasba PK, and Nussinov R (2003a) The interdependence of backbone flexibility, residue conservation and Enzyme function: A case study on β1,4galactosyltransferase. *Biochemistry* **42**, 3674-4687.

Gunasekaran K, Ma B, and Nussinov R (2003b) Triggering loops and enzyme function: Identification of loops that trigger and modulate movements. *J. Mol. Biol.* **332**:143-159.

Gunasekaran K and Nussinov R (2004a) Modulating functional loop movements: The role of highly conserved residues in the correlated loop motions. *ChemBioChem* **5**:224-230.

Gunasekaran K, Tsai C-J, and Nussinov R (2004b) Analysis of ordered and disordered protein complexes reveals structural features discriminating stable and unstable monomers. *J. Mol. Biol.* **341**:1327-1341

Hamachi I, Watanabe JI, Eboshi R, Hiraoka T, Shinkai S (2000) Incorporation of artificial receptors into a protein/peptide surface: A strategy for on/off type of switching of semisynthetic enzymes. *Biopolymers* **55**, 459-68.

Koh JT (2002) Engineering selectivity and discrimination into ligand-receptor interfaces. *Chem Biol.* **9**, 17-23.

Looger LL, Dwyer MA, Smith JJ, Hellinga HW (2003) Computational design of receptor and sensor proteins with novel functions. *Nature.* **423**, 185-90.

Pitera JW, Swope W (2003) Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proc Natl Acad Sci USA* **100**, 7587-92.

Ramakrishnan B, Boeggeman E, Ramasamy V, Qasba PK (2004) Structure and catalytic cycle of β1,4-galactosyltransferase. *Curr Opin Struct Biol* **14**, 593-600

Ramasamy V, Ramakrishnan B, Boeggeman E, Qasba PK (2003) The role of Tryptophan 314 in the conformational changes of β1,4-galactosyltransferase-I. *J Mol Biol* **331**, 1065-76.

Tann CM, Qi D, Distefano MD (2001) Enzyme design by chemical modification of protein scaffolds. *Curr Opin Chem Biol* **5**, 696-704

Tian SS, Lamb P, King AG, et al. (1998) A small, nonpeptidyl mimic of granulocyte-colony-stimulating factor. *Science* **281**, 257-9.

Teaching Statement

My general teaching interests are in the area of computational biology, bioinformatics and biophysics. My primary goal in teaching is to foster critical thinking so as to prepare the students for careers in these rapidly developing fields.

As a teaching assistant for a graduate level course on conformational analysis of peptides and proteins, I filled the gap between the students and faculty. I was then a graduate student at the India's prominent institute of Indian Institute of Science, Bangalore. I enjoyed interacting with students and directing them to reading materials from where they can find the answers for their assignments. Since the class consisted of students from vastly different disciplinary, I provided concise and clear hints for those who were finding the problems to be difficult to solve. I purposefully avoided in helping the students in all aspects of solving their assignment problems.

When teaching a course, I would explore several possible ways to organize the materials. The materials must be well thought-out and adaptive. Upon finding a way that fosters the learning process and critical thinking, I would read through the materials several times trying to think of the central ideas. I would then explore how the ideas can be conveyed to the students in a way that is interdisciplinary. One possible method of achieving this is through discussion of practical real world examples. I consider the discussion of examples to be vital component of the learning process.

Before each class I would make available the lecture outlines, which would then be presented in a PowerPoint format during the class. This would help the students to visualize the structure of the class ahead of the time. Importantly, the outlines would provide the central ideas and critical examples. The examples would be discussed in the class in detail. Further, once the fundamental concept is conveyed, students would also be engaged in the discussion of the examples in the class. At the end of the class, students would be left with further examples to explore on their own. They would also be directed to materials that would help those curious students who seek further understanding and wanting to develop their skills.

I would assess and evaluate my teaching effectiveness in achieving the objectives through constant evaluation in the form of class and web based online tests and quizzes. The tests will also be designed to engage the students to develop their skills. I would commit myself to make sure that no student is left behind in the class. Both the effectiveness of the teaching and learning skills of the students could be improved through this iterative process.

I would be excited by the opportunity to work with the students as I consider teaching to be a necessary component of an academic career. It fosters critical thinking not only for the students but also for the teacher.