Evolution of My Scientific Research Interests (2020)

Period I. Structural Biology (Approximate period: 1962-1998, supported primarily by US National Institutes of Health, USA)

During most of my scientific career, I was interested in experimentally discovering the atomic details of three-dimensional (3D) structures of molecules in living cells by X-ray crystallography to understand how they may work in cells. Many of these molecules are key players in molecular communication networks for transmitting signals for normal or abnormal cell growth (such as in tumors). The structures determined during this period by my group covered wide ranges of known or unknown functions at the time, including those involved in translation of the genetic code (transfer RNA), molecular communication in normal cells (cellular Ras proteins, protein kinases, cyclins, chemotaxis receptors, and sweet tasting proteins) and in cancer cells (oncogenic Ras proteins, oncogenic protein kinases). What I learned during this period: (1). Almost all of these structures revealed wonderful surprises with unexpected features that helped understand some or many of their functions, have highly intricate architectures composed of, in general, many motifs, and are esthetically beautiful; (2). The studies also revealed that the total number of architectural motifs in these structures is very large and the combinatorial assembly of the motifs can generate a huge diversity of protein structures; (3). The degree of diversity of proteins and complexity of their three-dimensional structures may be beyond what our human brain can predict; (4). When there is need for one or more new functions for an organism to survive, "Nature" seems to select, in most cases, one or more proteins from a large pool of the existing diverse proteins or assemble one or more new proteins from existing protein motifs (through genes and gene motifs), then optimize the structures of the selected or newly assembled proteins for new functions for survival.

Period II. Structural Proteomics (Approximate period: 1998-2008, supported initially by US Department of Energy, then by US National Institutes of Health for most of the period)

Realization of the complexity and diversity of the 3-D structures of proteins and nucleic acids led my curiosity to ask this question: How big is the "Universe" of protein structures and that of nucleic acids, and how are they populated? During this period my major focus was in determining the 3-D structures of a large number of proteins especially those proteins of unknown structures and/or functions (i.e., the proteins with no sequence similarity to those with known structures or functions). Because the task was too large, we focused on one *organism with one of the smallest genomes, Mycoplasma genitalium*, which has only about 500 genes coding for mostly proteins. This project involved a large number of scientists and technicians to clone and express large number of genes of this organism or their homologs from other organisms, and determine their 3-D structures. The results of this project, combined with the 3-D structures of Period I and other known protein structures provided me with a "global" view of the "Universe" of protein 3-D structures. Similar studies on nucleic acids

(DNAs and RNAs) resulted in also a “global” view of the “Universe” of the conformational space (shape variations) of building blocks of nucleic acids, revealing possible pathways of conformational transitions found in DNA and RNA. During this period I learned: (1). About half of the new structures have their protein “folds” (the architectural motifs of the backbone structure) similar to those already observed among the proteins of known structures. This highlighted the fact that two proteins of different amino acid sequences can have the same or similar protein fold; (2). The proteins of similar folds often have related functions; (3). The “Universe” of protein architectural (fold) motifs is finite and sparsely populated, but their combinatorial assembly generates a huge diversity of protein structures; (4). Among the five protein fold classes (α, β, α+β, α/β and coil), the α/β fold class, proteins consisting of mostly alternating α-helix and β-strand, evolved most recently from α+βclass, the proteins consisting of random mixture of α-helices and β-strands, and became the most populated fold class.

Period III. Drug discovery (Approximate period: 2001-2011, supported by Plexxikon, Inc. Berkeley, CA, USA)

My previous experiences with 3-D structures of proteins suggested that, with the dramatic technical development in X-ray crystallography during the last two decades, it should be possible to develop small molecule drugs in a much shorter time and with reduced cost than prevailing (around year 2000) approaches by major pharmaceutical companies. With two colleagues, we founded a small company in Berkeley, CA, Plexxikon, and developed a process called “Scaffold-based drug design.” Using this approach, Plexxikon discovered (during 2001-2005) a drug candidate against metabolic syndrome, including type 2 diabetes, in a very short time and at low cost. The candidate was licensed out to a large pharmaceutical company, which was subsequently acquired by another large company, that terminated the project. The second drug developed (2006 –2011) was against malignant melanoma, one of the deadliest cancers, again in a very short time and at a much lower cost. This drug (Vemurafenib which has the commercial name of Zelboraf) was one of the first “personalized and targeted drugs” approved by US FDA through a “fast track” process. Success of this discovery highlighted the important role of individual genomic susceptibility to diseases and treatment responses. This realization directed my current interest in genomic variations in Humans (see below).

Period IV. Computational Genomics (Approximate period: 2008–present, supported by World Class University Program, Republic of Korea (“South Korea”) and a gift grant to University of California, Berkeley).

A “global (wide-angle)” view of the “Universe” of the architecture of protein structure (see above) raised my curiosity to the possibility of constructing an “even wider-angle” view of all living organisms, describable by the whole genome sequences that became available for many organisms covering the three domains of life (Archaea, Bacteria, and Eukarya). Since each organism can be represented by a “book” written with four-letter alphabets (A, T, C, G) with no spaces, we have developed a method, **“Feature Frequency Profile (FFP)”**

method, to quantitatively calculate the degree of difference between any two books. This method is an adaptation of "Word Frequency Profile" method, commonly used to compare two books with, for example, 26-letter alphabets using **Natural Language Analysis** algorithms based on Information Theory.

A. Whole-proteome "Tree of Life": a new phylogeny of all extant organisms.

An "Organism Tree of Life (ToL)" can be considered as a metaphorical and conceptual tree to capture a simplified narrative of the complex and unpredictable evolutionary courses of all living organisms. Currently, the most common approach has been to construct a "gene ToL", as a surrogate for the organism ToL, by selecting a group regions (that can be aligned with high reliability) of each of the select genes/proteins to represent each organism. Such selected regions, however, account for a small fraction of all genes/proteins and even smaller fraction of the whole genome of an organism. During the last decades, whole-genome sequences of many extant organisms became available, providing an opportunity to construct a "whole-genome or whole-proteome ToL" using our FFP method without sequence alignment. We have been able to construct a "Whole-proteome ToL" for over 4,000 extant organisms, for which whole genome sequences are available in the public genome database. The most surprising and unexpected feature of our ToL was that the founders of all 5 Kingdoms of all living organisms (Bacteria, Archaea, Fungi, Plants, and Animals) emerged in a **"deep Burst"** near the root of the ToL, a feature not observed in all earlier ToLs constructed based on a set of selected gene or protein regions that can be multiply aligned. Encouraged by this observation, we have started to construct whole genome/proteome ToLs for separate groups of organisms at the phylogenic levels of Phylum, Class, and Order.

B. Whole genome variation of Human species.

Most regions of genomes of normal human cells have been found to have the same sequences among individuals, but a small fraction, spread throughout the genome, have variations within a population group. Of these, the **single nucleotide variations** (SNVs) account for the largest number of variations and, have been identified in over 80 million genomic positions out of 3 billion positions (loci) in a whole haploid genome. It has been widely accepted that the analysis of SNVs may be able to allow one to predict the genomic component of the **disease susceptibility** of individuals to complex diseases such as cancers, neurological diseases, autoimmune diseases and other traits. So far, the results from the current analysis methods (e.g. Genome-wide Association Studies method) and interpretation of the results have yielded information of limited predictive value of practical utility for making health-related decisions at individual or population level without information of family histories.

Since prevention and early diagnosis of cancer are the most effective way of avoiding psychological, physical, and financial suffering from cancer, we developed a machine-learning method for statistically predicting individuals' *inherited susceptibility* (and environmental/lifestyle factors, by inference) for acquiring the most likely type among a panel of 20 major common cancer types plus one "healthy" type. The results show that, depending on the type, about *33 to 88%* of a cancer cohort has acquired its cancer type primarily due to inherited genomic susceptibility, and the rest primarily due to environment/lifestyle factors. These genomic susceptibilities with associated probabilities, at the cohort level, may provide practical information for health professionals and health policy makers working in the fields of prevention and/or early intervention of cancer. We are in the process of extending this approach to predict the susceptibility at the individual level.

C. Genomic studies of world's ethnic populations.

An ethnic population has different meanings to different people, but, generally, is a group of people who have a "perceived notion" that its members share a set of unique **inherited (genomic)** and **acquired (non-genomic)** traits, such as ancestry, social and cultural norms, religion/belief, language and life style. Thus, ethnic group identity has a strong emotional component that divides the people into opposing categories of "us" and "them", one of the primary causes for human conflict and suffering.

Recent availability (from Simons Genomic Diversity Project) of genomic sequences of a large number of ethnic populations throughout the world (over 160 ethnic groups) provides an opportunity to estimate quantitatively the fraction of whole genome that may account for the **inherited genomic component** of the ethnicity and to find a relationship, if any, between ethnic grouping and genomic grouping. We have applied our FFP method by representing each individual as a "book" written in alphabets of the genotypes of whole-genome variations (Single Nucleotide Variations (SNVs)). This new approach is starting to reveal a new landscape of the grouping pattern and the order of emergence the human ethnic populations in a genomic space.