Discovery of Novel Gene Functions by Chemistry-Guided Targeted Sequence Analysis

Mina Rho*†‡ and Woon Ju Song*§

†Department of Computer Science and Engineering, Hanyang University, Seoul 04763, Republic of Korea
‡Department of Biomedical Informatics, Hanyang University, Seoul 04763, Republic of Korea
§Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea

The advent of next-generation sequencing technology, along with recent advances in bioinformatics algorithms, have accelerated the acquisition and analysis of massive genome and metagenome data. Numerous genes have been identified in the biosynthesis of natural products, metabolic pathways, and chemical signaling pathways between inter- and intraspecies, and therefore, identification and elucidation of gene functions are essential.

A conventional and traditional biochemical approach to explore discovered genes would be to start from a series of gene clusters that exhibit intriguing functions (Figure 1a). Determination of protein structures, reaction mechanisms, and intermolecular interactions with cognate substrates allow us to pinpoint the functions of the genes and the roles of amino acid residues at the molecular level. Multidimensional information such as kinetic and spatial data can be obtained, although the number of genes that can be explored is relatively low.

A bioinformatic approach, in contrast, is a top-down analysis of large-scale sequence data (Figure 1b). A homology-based search, multiple sequence alignment, entropy analysis, sequence similarity networks, and phylogenetic analysis provide us with a list of sequences with the highest likelihood. While a bioinformatic approach is applicable to an essentially unlimited number of data sets, it is less useful for the genes if no or only a little number of the analogous genes have been reported and collected for analysis.

Therefore, iterative integration of both biochemical and bioinformatic data, named as “chemistry-guided targeted sequence analysis” herein, would be valuable to understand the massive data in chemical contexts and to decipher a high number of novel gene functions with greater accuracy (Figure 1c). We can reflect biochemical information on the massive sequence analysis data to factor in the relative importance of conserved or seemingly less strictly conserved residues, and the presence of adjacently located genes of known functions. For example, depending on the degrees of amino acid conservation and relative spatial position of the sequences, residues can be grouped and examined. The residues that are highly conserved and located at the active sites of the previously characterized proteins are likely to be important in the identification of protein superfamily or its functions. Residues that are conserved but located distant from the active sites are likely to be related to protein folding, structure, and solubility. In contrast, novel gene functions can be discovered by the close inspection of the genes having the sequences that are less conserved from the previously characterized genes but located at the vicinity of the active sites in the analogous protein structures. In the massive sequence analysis, the genes with such seemingly not-conserved residues might be clustered within a superfamily, implying that the sequence variations may produce discrete chemical reactivities as an indicator for novel functions. In this regard, the Enzyme Function Initiative constructs an iterative pipeline in a multidisciplinary sequence and structure-based strategy. Such efforts, in conjunction with recent advances in the studies of functional annotation, have assisted identifying genes with diverse functions.

Balskus et al. established a chemically guided functional profiling strategy to integrate biochemical understandings with
metagenomic results.\(^2\) They discovered ubiquitous and abundant glycy1 radical enzymes from the gut microbiome by prioritizing and clustering genes. In particular, they identified uncharacterized gene clusters by inspecting sequences that exhibit no conservation to the active sites that were previously characterized for substrate-binding or catalysis. Then, uncharacterized genes were investigated by various bioinformatic and biochemical methods, demonstrating that they are \(\text{trans}-4\)-hydroxyl-L-proline dehydratases.

Alternatively, novel gene functions can be identified by searching for the genes that are in close proximity to those of known functions when their chemical functions are potentially related to each other. Seyedsayamdost et al. explored the vast biosynthetic landscape in \(\text{Streptococci}\) by devising a genetic context-dependent bioinformatic search.\(^7\) They examined the genes that are neighboring with those related to quorum sensing to discover the genes involved in the post-translational modifications of peptide secondary metabolites. One such example is a radical \(S\)-adenosylmethionine enzyme, which has unprecedented double-cross-linking activities in a regio- and stereospecific manner.

We also carried out a chemistry-guided analysis of 45 rifamycin ADP-ribosyltransferase genes that are associated with resistance to rifamycin antibiotics.\(^7\) Sequence analysis indicates that most of the residues constituting the catalytic domains are predominantly conserved, whereas some of the residues at the active site pocket are not strictly conserved. Although low sequence conservation often indicates little importance in function, they were in close contact with the substructures of rifamycin analogs, indicating that the variations in the sequence might be responsive to the structural changes of the antibiotics, concurrently forming altered intermolecular interactions. Indeed, we demonstrated that the residues are related to the distinct reactivities with rifamycin analogs. Our results imply that adaptation to changing chemical environments such as the introduction of novel antibiotics into biological systems could trigger substitution of amino acids, resulting in substantial variations in the chemical and biological functions.

The last example may suggest that chemistry-guided targeted sequence analysis is necessary to elucidate the native function of the genes. Recently, genes in the biosynthesis pathway of lantibiotic NA1-107 have been identified, including a halogenase (MibH).\(^8\) Although MibH exhibits sequence similarity with \(\text{FADH}_2\)-dependent tryptophan halogenases, it exhibits no reactivity with a free tryptophan. Instead, it reacts with an indole ring within a polypeptide. The unique activity may be related to the fact that the aromatic substrate-binding region is composed of less strictly conserved amino acids, and the sequences are less explicitly explored. Therefore, further chemistry-guided sequence analysis on the substrate-binding sites may identify the residues that determine the chemical and biological functions of potentially more diverse and abundant halogenase genes.

The examples listed above suggest that the integration of bioinformatics and biochemical analysis is a powerful and indispensable approach to understand diverse gene functions. By filtering the massive sequence data with biochemical viewpoints, novel genes and essential residues can be discovered. Iterative and integrated targeted sequence analysis in chemical contexts, therefore, will enrich the scope and improve our understandings of protein sequence—structure—function relationships to expedite discovery and applications of novel genes.

**REFERENCES**


**AUTHOR INFORMATION**

**Corresponding Authors**

*E-mail: minarho@hanyang.ac.kr.*

*E-mail: woonjusong@snu.ac.kr.*

**ORCID**

Woon Ju Song: 0000-0003-0434-2684

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

The authors are supported by the Collaborative Genome Program of the Korea Institute of Marine Science and Technology Promotion (KIMST) funded by the Ministry of Oceans and Fisheries (MOF) (20180430) and the National Research Foundation (NRF) of Korea (NRF-2019R1C1C1003863).