INVESTIGATION OF GÖDEL-TYPE CODING OF PROTEIN PRIMARY STRUCTURES

By
Ashvi Patel and Daniel J. Graham
Department of Chemistry and Biochemistry
Loyola University Chicago
ABSTRACT

The folded structures of proteins, along with their biological functions, are controlled by one-dimensional sequences of amino acids, aka protein primary structures. At present, there are over 130 million protein primary structures on record in international databases. Such information motivates much of genome research for organisms and viruses in modern day. Proteins are the molecules that make life possible. They are the principal targets in drug therapy. The list of important protein properties is a long one.

Protein primary structures are encoded traditionally as letter strings, for example, the following ubiquitin ligase (PDB: 4ZFI):

MOIPASEQETLVRPKPLLKLKSVGAQKDTYTMKEVLFYLGQYIMTKRLYDEKQQHIVYCSNDLGLDLFGVPSFSVKEHRKIYTMIRNLVVNQQ

As is well-familiar to biochemists and molecular biologists, each symbol refers to one of the twenty standard amino acids encoded by a genome: A⟷alanine, V⟷valine, etc. The folded structure and biological functions hinge on the nuances of string length, component diversity and order.
INTRODUCTION

Gödel’s coding quantifies a protein sequence with a numerical value. It starts with giving a number to the first symbol of a sequence based on a pre-determined second sequence due to the position of the symbol. It links each symbol together with each of the symbol’s number and then forms a larger number that quantifies the whole sequence.

In this research, a symbol can be seen as an amino acid within the sequence which is known as the protein sequence. Another second sequence is referred to as the amino acid alphabet in which allows to quantify the amino acid within the protein sequence based on its position in the alphabet.

Gödel’s coding is efficient since it allows for comparison of protein sequences, that can be many amino acids long, into a single number that can be assessed via graphs to compare proteins on a mathematical scale.

Our research explores alternative coding methods for protein primary structures. The motivation is to illuminate structure/function properties that are otherwise obscure using the traditional coding methods. The methodology makes appeal to Gödel-type coding. Critically, it offers up to 20! (≈ 2.418) expressions for a single primary structure. Such numbers vastly exceed those of database archives.
METHODOLOGY

The standard twenty amino acids offer 20! lexicographic orders. Four from the ocean of possibilities express as:

WCHANMFRDPETGIOSYKVL
LVKYSOIGTEPDRFMNAHCW
AMPSWNHCGDTRFIEVOKLY
LVYQKERFSNGATIPMHWDCC

The ocean depth is determined by combinatoric relationships among factorials. There are 20 choices for the first letter of a lexicographic order, 19 for the second, 18 for the third, and so forth. The possibilities reach $2.4^{10}$, and indeed grow exponentially larger if post-genome processing protein sites are included. Gödel-type coding combines a given lexicographic order with the prime numbers: 2, 3, 5, 7, etc. taken in order. The coding requires a prime number for every amino acid unit in a primary structure. The coding application re-packages information in a protein letter string as a single positive integer. The method is perfectly reversible for all proteins, real and hypothetical, and for lexicographic orders. This means that information about a primary structure is never lost during the coding process. It is always possible to retrace steps to recover the original sequence.
A demonstration summarizes the methodology. Revisit the ubiquitin ligase and the first of the above lexicographic orders:

\[
\text{MQIPASEQETLVRPKPLLKLKKSVGAQKDYTMTKEVLFLGQYIMTKR}L\text{DEKQHQIVYCSNDL}G\text{DLFGVPSFSVKEHRKIYTM}Y\text{RNLLV}V\text{V}\text{NQQ}
\]

\[
\text{WCHANMF}RD\text{PETGI}Q\text{SYKVL}
\]

The integer function \( G(k) \) is constructed by forming a product of the prime numbers \( p \), each raised to a power determined by the lexicographic order. The index \( k \) numbers the amino acid units in a protein: 1, 2, 3, ..., viz.

\[
\text{MQIPASE...} \quad G(k) = 2^6 \times 3^{15} \times 5^{14} \times 7^{10} \times 11^{14} \times 13^{16} \times 17^{11} \times ...
\]

The method is reversible because of the fundamental theorem of arithmetic: every integer can be expressed as a unique product of prime numbers. Thus, integer \( G \) has a given prime factor representation determined by the protein sequence and the rank of each amino acid in the lexicographic order. The 20! allowed orders equate with the number of possible \( G \)-integers for a given protein. Due to time and energy limits, it is possible to examine only a small fraction of the allowed \( G \).
METHODOLOGY CONT.

To diversify our results, we randomized the amino acid alphabets by placing the overall principle of Gödel's coding into a program. Using C++, we used a random generator provided by a library, and from there were able to randomly generate $2^{19}$ amino acid alphabets. Out of these alphabets we used the one that produces the highest Gödel's value and one that produces the lowest Gödel's value for a given protein.

We focused on the E3 ubiquitin ligase known as MDM2. We specifically used this protein to assess one that is considered a druggable protein.

For the MDM2, we assessed through multiple graphs the highest producing and lowest producing amino acid alphabets.

MDM2
MAX ALPHABET: WCHANFRDGMPESIYQKVL
MIN ALPHABET: LVKQYISTEPMGDRFNAHCW
RESULTS

The graph is for $\log_e G$ based on random alphabets. It has a Gaussian curve and shows the maximum and minimum values using alphabets based on the amino acid distributions in the MDM2 protein. The alphabets derived are from the previous page for the MDM2 protein based on the highest and lowest Gödel’s values from the graph.
The maximizing and minimizing amino acid alphabets were further assessed to confirm their relationship to the MDM2 protein.

There was a further extension that is presented in the graph below. The graph shows the relationship of $\Delta \log G_{\max} / \Delta \log G_{\min}$ behavior and weighed it against the secondary structure of the MDM2 protein.
DISCUSSION

The binding region of MDM2 has mainly nonpolar amino acids (LFYLGQYIMTKR). MDM2 must have binding interactions with other proteins via hydrophobic interactions due to the nonpolar attachment site. It can also be a phosphorylation site as it has many amino acids that can be phosphorylated such as tyrosine and threonine.

The graph on the left strengthens the relationship of the maximizing and minimizing amino acid alphabets. And due to being selected from a group of randomized alphabets, our experiment has randomized results.

The amino acid alphabets show how reversing the amino acids alphabets change the Gödel's value which in turn relates to the overall protein. If there's a low/high Gödel's value, one can predict which amino acids are more prevalent and less prevalent, and how these amino acids play a role in the overall protein function.
DISCUSSION CONT.

MDM2 is an important protein to focus on as it is considered a druggable protein. MDM2 plays an important role in the degradation of p53 in normal cells, thus if mutated in cancer cells, it can allow for cell to inhibit p53 degradation which can cause a cascade of reactions that ultimately allow for cancer cell progression. MDM2 is a current topic of study as by having a drug target the protein, it can possibly allow to decrease the activity of p53 and in turn rapid cell progression. The C-terminus of MDM2 is the area of many phosphorylation sites, and once MDM2 is phosphorylated it can allow for the protein to enter the nucleus and inhibit p53 degradation. However, the interaction of p53 and MDM2 occurs mainly near the N-terminus of the protein, which even differs from the graph on the right on Slide 8. MDM2 can be suggested to have multiple binding sites, but further research would be needed. Understanding the complexity of the primarily structure of MDM2 is important to assess in order to understand its functioning in diseases and cancer.