Understanding the energy landscape that guides protein folding using machine learning and molecular simulations

Wenfa Ng

Citizen scientist, Singapore, Email: ngwenfa771@hotmail.com

Abstract

Natural proteins fold with a smooth energy landscape into their lowest energy state within a short timeframe. On the other hand, *de novo* designed proteins, especially ones with unnatural amino acids may encounter a disjointed energy landscape during folding; hence, resulting in unfolded proteins or proteins stuck in a higher energy and unstable state. Hence, what defines a smooth energy landscape for protein folding? Presence of specific sequence motifs seems to be an enabling factor. Hence, one possible research direction is to elucidate the defining amino acid sequence motifs that help chart a smooth energy landscape for protein folding. To this end, the amino acid sequence and structural fold of natural proteins in a bacterial proteome such as that of *Escherichia coli* would serve as training data for machine learning algorithms designed to identify patterns in amino acid sequence that correlate with specific structural motifs. While many proteins require chaperones to aid folding, the correlations between amino acid sequence and structural fold obtained would still serve as a good starting point for understanding the influence of amino acid sequence on protein folding. In essence, the correlations obtained help provide an understanding of the energy landscape of a protein. Subsequently, the work could move on to delineate design principles and rules to help guide *de novo* design of proteins or proteins incorporating unnatural amino acids. Using multiple sequence alignment combined with energy calculations, the work could attempt to identify sequence motifs in proteins that account for a smooth energy landscape. Specifically, it will answer the question of what amino acid sequence will smooth the kinks in an energy landscape that guides the folding of proteins. Finally, design rules elucidated could be tested experimentally by the design of genes encoding the particular amino acid sequence of a designed protein. Given that poorly folded proteins tend to form aggregates such as inclusion bodies, the state of folding of designed proteins could be assessed based on their solubility in the cytoplasm of the expression host through a SDS-PAGE analysis of the fraction of the protein in the pellet and cell lysate. Designed proteins incorporating unnatural amino acids could be tested similarly. Overall, results and knowledge emanating from the research will expand our understanding of protein folding processes and the role of amino acid sequence in determining the speed at which proteins fold and the energy state at which they will rest. Furthermore, knowledge gained would also help guide the design of *de novo* proteins with unusual amino acid sequence or unnatural amino acid incorporated, with the goal of helping *de novo* proteins fold in an expeditious manner to a low energy state.

Keywords: amino acid sequence, energy landscape, structural motifs, design rules, protein folding, folding speed, unstable state, unnatural amino acids, solubility, chaperones,

Subject areas: biochemistry, computational biology, machine learning, biotechnology, structural biology,

Conflicts of interest

The author declares no conflicts of interest.

Funding

No funding was used in this work.