Chapter 7
Predicting Protein Folding Pathways

**Mohammed J. Zaki, Vinay Nadimpally,
Deb Bardhan, and Chris Bystroff**

Summary

A structured folding pathway, which is a time-ordered sequence of folding events, plays an important role in the protein folding process and hence in the conformational search. Pathway prediction thus gives more insight into the folding process and is a valuable guiding tool for searching the conformation space. In this chapter, we propose a novel "unfolding" approach for predicting the folding pathway. We apply graph-based methods on a weighted secondary structure graph of a protein to predict the sequence of unfolding events. When viewed in reverse, this process yields the folding pathway. We demonstrate the success of our approach on several proteins whose pathway is partially known.

7.1 Introduction

Proteins fold spontaneously and reproducibly (on a time scale of milliseconds) into complex three-dimensional (3D) globules when placed in an aqueous solution, and the sequence of amino acids making up a protein appears to completely determine its three-dimensional structure [16, 249]. At least two distinct though interrelated tasks can be stated.

1. *Structure Prediction Problem:* Given a protein amino acid sequence (i.e., linear structure), determine its three-dimensional folded shape (i.e., tertiary structure).
2. *Pathway Prediction Problem:* Given a protein amino acid sequence and its three-dimensional structure, determine the time-ordered sequence of folding events, called the folding pathway, that leads from the linear structure to the tertiary structure.

The structure prediction problem is widely acknowledged as an open problem, and a lot of research in the past has focused on it. The pathway prediction problem, on the other hand, has received almost no attention. It is clear that the ability to predict folding pathways can greatly enhance structure prediction methods. Folding pathway prediction is also interesting in itself since protein misfolding has been identified as the cause of several diseases, such as Creutzfeldt-Jacob disease, cystic fibrosis, hereditary emphysema, and some cancers. In this chapter we focus on the pathway prediction problem. Note that while there have been considerable attempts to understand folding intermediates via molecular dynamics and experimental techniques, to the best of our knowledge ours is one of the first works to *predict* folding pathways.

Traditional approaches to protein structure prediction have focused on detection of evolutionary homology [13], fold recognition [56, 370], and where those fail, ab initio simulations [372] that generally perform a conformational search for the lowest energy state [369]. However, the conformational search space is huge, and, if nature approached the problem using a complete search, a protein would take millions of years to fold, whereas proteins are observed to fold in milliseconds. Thus, a structured folding pathway, i.e., a time-ordered sequence of folding events, must play an important role in this conformational search [16]. The nature of these events, whether they are restricted to "native contacts," i.e., contacts that are retained in the final structure, or whether they might include nonspecific interactions, such as a general collapse in size at the very beginning, were left unanswered. Over time, the two main theories for how proteins fold became known as the "molten globule/hydrophobic collapse" (invoking nonspecific interactions) and the "framework/nucleation-condensation" model (restricting pathways to native contacts only).

Strong experimental evidence for pathway-based models of protein folding has emerged over the years, for example, experiments revealing the structure of the "unfolded" state in water [276], burst-phase folding intermediates [82], and the kinetic effects of point mutations ("phi values" [300]). These pathway models indicate that certain events always occur early in the folding process and certain others always occur later (Figure 7.1).

Currently, there is no strong evidence that specific nonnative contacts are required for the folding of any protein [75]. Many simplified models for folding, such as lattice simulations, tacitly assume that nonnative contacts are "off pathway" and are not essential to the folding process [227]. Therefore, we choose to encode the assumption of a "native pathway" into our algorithmic approaches. This simplifying assumption allows us to define potential folding pathways based on a known three-dimensional structure. We may further assume that native contacts are formed only once in any given pathway.

Knowledge of pathways for proteins can give important insight into the structure of proteins. To make pathway-based approaches to structure prediction a reality, plausible protein folding pathways need to be predicted.