蛋白質熱穩定性預測與分析

學生：梁涵塑　　指導教授：黃鎮剛

高明達

國立交通大學 生物科技系 博士班

中文摘要

蛋白質熱穩定性的研究不論是學術理論或是產業應用上都是一個極重要的科學議題。過去累積二十多年的相關研究後，已知可提高蛋白質熱穩定性的因素有數十種之多，多數是以結構觀點的同源比較所推論出來；因而在應用上，都需要有蛋白質結構資訊為前提，用以選擇增加或減少這類有效的因素的突變位置；一旦缺乏結構資訊將極難選擇適當突變位置。而且同源比較所得的結論往往只適用於特定結構，這些因素並不通用於各種結構狀況，因此造成實際應用上的困難。符合實際一般應用，需要發展只根據蛋白質序列資訊即可預測蛋白質熱穩定性的方法，進而做提高熱穩定性的突變位置選擇。在此論文中我們提出兩個以序列資訊為基礎的熱穩定性量化指標：二級結構亂度 (structural entropy)，胺基酸對偶特徵(amino acid-coupling patterns)。二級結構亂度是以SCOP-35中的二級結構資訊進行統計，而我們發現其亂度與蛋白質的穩定實驗值ΔT_m有高度負相關(r = 0.72)；胺基酸對偶特徵及胺基酸組成生長溫度回歸皆是以微生物的最適生長溫度統計不同溫度分群的特徵差異，以點突變的穩定改變方向預測驗證，*MCC* 最高可達到0.47。以上的方法皆可用來預測蛋白質序列中的穩定區域，並以真實的實驗驗證得到不錯的預測結果。我們將各式各樣的熱穩定性因素整合成單一量化指標，進而在點突變熱穩定性的選擇上可以有效加以應用。應用這些統計基礎開發的熱穩定預測網站工具TheCUP，可提供使用者一個可能的方向，以統計的經驗法則方式取得較為耐熱的酵素基因序列，提供任何蛋白質酵素改造耐熱的可能。這些資訊整合進序列趨勢profile，因此只需要蛋白質序列，即可快速得到以高溫生物基因組資訊為基礎的耐熱程度趨勢，如果有研究計畫需要以基因點突變的方式改造酵素，即可得到所需要的序列可能組合方式以提高耐熱效率，在有結構資訊狀況下可減少分析結構資訊時所需分析的複雜度，甚至沒有結構資訊的狀況下利用這些方法也依然有機會設計有效的突變方式。未來希望能真實應用過去研究的基礎，進行實際的實驗驗證我所提出的理論，並應用在生技產業酵素之上。

Protein thermostability prediction and analysis

Student : Han-Kuen Liang Advisor : Jenn-Kang Hwang

Ming-Tat Ko

Department of Biological Science & Technology

National Chiao Tung University

ABSTRACT

The study of protein thermostability is a very important scientific issue for both academic research and industrial applications. It is known throughout the last twenty years that most factors to thermostability are based on the comparisons of structural homology. However, it is difficult to select mutation sites and changing amino acid types to enhance thermostability when structural information is not available for a protein. Several statistical based methods including structural entropy, and amino acid coupling patterns, have been developed to evaluate protein thermostability. Structural entropy has a good linear relationship between the average structural entropy and the melting temperatures (r = 0.72), and the feature of amino acid coupling and amino acid composition regression had been validated by the single mutation stability change $\Delta\Delta G$ prediction (MCC = 0.47). The factors to thermostability are considered in these methods and applied in the selection of mutation sites. Thus, we can apply those methods to develop a web service tool TheCUP using the proposed method to suggest appropriate mutation site and changing amino acid types while only have the sequence information. TheCUP provide users with opportunities to enhance the thermostability of a query enzyme based on propensity profiles generated from these analyses. Given a query protein sequence, TheCUP can suggest candidate mutation sites to improve the enzyme thermostability based on the profiles of thermophilic organisms. This tool not only makes enhancement of enzyme thermostability possible without any structural information but also reduces the analytical complexity when structural information is available. In the future, I hope that those methods can be verified by real experiment.

ACKNOWLEDGEMENT

終於到了要畢業的時候，其中有很多的不捨與感動.....

首先要感謝的是中研院資訊所的高明達老師，能夠指點與收留我這實驗背景的學生；再來要感謝黃鎮剛老師的提拔，很多分析與觀念的建立都來自黃老師高明的指點與帶領，真希望以後做研究時能像老師一樣神奇。

再來最要感謝的是中研院高明達實驗室一起合作的嘉茂，有你的才華才能有今天的成果；雖然現在你離開了生物資訊領域，但還是希望以後能在其他領域上繼續合作。還有鎮熊學長，有你的參與才能讓我們這麼快完成第一篇文章。再來是交大實驗室的各個學長、同學與學弟妹，雖然我大多數時間都在中研院，在新竹相處的時間不多，不過有你們的幫忙或是生活上的各種交流，都是令人感動且難以忘懷的。

最後感謝我的家人，有你們的包容才能讓我無後顧之憂的從事研究工作。

希望不會辜負你們的期望，能夠用這論文所發展的一些成果，作為開創我人生道路的基石。

CONTENTS

中文摘要.....i

ABSTRACT ii

ACKNOWLEDGEMENT.....iii

CONTENTS.....iv

CHAPTER 1 Introduction.....1

1.1 Research Status of Thermostability.....2

1.2 Motivation.....4

CHAPTER 2 Major Features in Thermophilic Enzymes.....6

2.1 More Salt Bridges (Ion Pairs).....6

2.2 Side Chain- Side Chain Hydrogen Bonds.....10

2.3 Dipole-Dipole Interactions or Cation-π Interactions.....10

2.4 Higher Degree of Hydrophobic Packing in the Core Regions.....11

2.5 Increased Packing Density.....11

2.6 Disulfide Bridges 12

2.7 Shorter Loop Structures or Higher Conformational Rigidity.....12

2.8 More Secondary Structural Elements Such as α-Helix or β-Sheet....12

2.9 Bias in Amino Acid Composition.....13

CHAPTER 3 Methods…..……..…….……19

3.1Sequence Derived Structural Entropy (SDSE)……..……..19

3.2 Amino Acid Coupling Patterns..……..……..…....23

3.2.1 Coupling Patterns [*XdZ*]…….……....23

3.2.2 Thermophilic Dominant (Thermo) and Mesophilic Dominant (Meso) Patterns Score *T* and *M*……...…….……25

3.2.3 Conditional Probability Score *R*...……..……..……25

3.2.4 Coupling Strength Score *C*……...…..……26

3.2.5 Dominant Pattern Profile (*T*, *M*, *T-M* Score) …….....……27

3.3 Optimal Growth Temperature Dependent Amino Acid Composition (OGT_{Comp}) ……28

CHAPTER 4 Results and Discussion……....29

4.1 Sequence Derived Structural Entropy (SDSE) …….……29

4.1.1 Case 1: Methanococcal Adenylate Kinases……...…...…29

4.1.2 Case 2: Ribonuclease HI……..……32

4.1.3 Case 3: Holocytochrome c551…….……33

4.1.4 The Relationship Between Structural Entropy and Thermal Stability…….34

4.2Amino Acid Coupling Patterns in Thermophilic Proteins……...39

4.2.1 The ρ Profiles of Amino Acid Coupling Patterns……...……...39

4.2.2 The ρ Profiles of $[xdZ]$ ……...…41

4.2.3 $[xdE]$ and $[xdV]$ ……...…41

4.2.4 $[xdP]$ and $[xdC]$ ……...……42

4.2.5 $[xdQ]$, $[xdT]$ and $[xdH]$ ……..……42

4.2.6 Other Coupling Patterns ……...…..…43

4.2.7 The Significant Amino Acid Coupling Patterns……..…55

4.2.8 Identification of Thermophilic and Mesophilic Proteins Using the Amino Acid Coupling Patterns ……..……...55

4.2.9 GC Contents and the Amino Acid Coupling Patterns…….60

4.3 Conclusion……..61

CHAPTER 5 Protein Mutation Stability Change Prediction Based on Sequence Information……...62

5.1 Methods and Implementation……..…...……..…62

5.1.1 Data Sets……..…...……...…62

5.1.2 Support Vector Machines Predictor……..…62

5.1.3 Sequence Coding Schemes in SVMs……..……..……..…63

5.1.3.1 The Amino Acid Composition……...……..…63

5.1.3.2 The Dipeptide Composition……..…64

5.1.3.3 The Coupling Composition……..…..……..…64

5.1.3.4 Experimental Environment Features (Temperature, pH)…....…64

5.1.4 Scoring and Performance……..…65

5.2 Results and Discussion……...……....…65

5.3 Conclusion…….……69

CHAPTER 6 TheCUP: a Web Tool for Protein Thermostability..…..70

6.1 Methods and Implementation……...……..…71

6.1.1 Thermostability Profiles…….……..…71

6.1.2 Input and Output Format……..…..……..……..…..…71

6.2 Conclusion.....……...……...…76

CHAPTER 7 Conclusion and Future Works……..…….……77

8. REFERENCS...……....……....80