國立交通大學

生物資訊系

碩 士 論 文

綜合分析蛋白質的打結

Comprehensive analysis of knots in proteins

研 究 生：賴彥龍

指導教授：黃鎮剛　教授

中 華 民 國 九 十 六 年 六 月

綜合分析蛋白質的打結
Comprehensive analysis of knots in proteins

研 究 生：賴彥龍　　Student：Yan-Long Lai

指導教授：黃鎮剛　　Advisor：Jenn-Kang Huang

國 立 交 通 大 學
生 物 資 訊 系
碩 士 論 文

A Thesis
Submitted to Department of Bioinformatics
College of Biologic Science and Technology
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Bioinformatics

June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

國　立　交　通　大　學

博碩士論文全文電子檔著作權授權書

本授權書所授權之學位論文，為本人於國立交通大學 生物資訊 系所 ＿＿＿＿＿＿ 組， 95 學年度第 2 學期取得碩士學位之論文。

論文題目：

指導教授：

■ 同意　□不同意

本人茲將本著作，以非專屬、無償授權國立交通大學與台灣聯合大學系統圖書館：基於推動讀者間「資源共享、互惠合作」之理念，與回饋社會與學術研究之目的，國立交通大學及台灣聯合大學系統圖書館得不限地域、時間與次數，以紙本、光碟或數位化等各種方法收錄、重製與利用；於著作權法合理使用範圍內，讀者得進行線上檢索、閱覽、下載或列印。

論文全文上載網路公開之範圍及時間：	
本校及台灣聯合大學系統區域網路	■ 中華民國 96 年 6 月　日公開
校外網際網路	■ 中華民國 96 年 6 月　日公開

授 權 人：賴彥龍

親筆簽名：賴彥龍

中華民國　96　年　6　月　　日

國　立　交　通　大　學

博碩士紙本論文著作權授權書

本授權書所授權之學位論文，為本人於國立交通大學 生物資訊 系所 _____ 組， 95 學年度第 2 學期取得碩士學位之論文。

論文題目： Comprehensive analysis of knots in proteins

指導教授：黃鎮剛

■ 同意

本人茲將本著作，以非專屬、無償授權國立交通大學，基於推動讀者間「資源共享、互惠合作」之理念，與回饋社會與學術研究之目的，國立交通大學圖書館得以紙本收錄、重製與利用；於著作權法合理使用範圍內，讀者得進行閱覽或列印。

本論文為本人向經濟部智慧局申請專利(未申請者本條款請不予理會)的附件之一，申請文號為：_____，請將論文延至____年____月____日再公開。

授 權 人：賴彥龍

親筆簽名：賴彥龍

中華民國　96 年　　月　　日

國家圖書館博碩士論文電子檔案上網授權書

ID:GT009451509

本授權書所授權之論文為授權人在國立交通大學 生命科學 學院 生物資訊 系所 _____ 組 95 學年度第 2 學期取得碩士學位之論文。

論文題目： Comprehensive analysis of knots in proteins

指導教授：黃鎮剛

茲同意將授權人擁有著作權之上列論文全文（含摘要），非專屬、無償授權國家圖書館，不限地域、時間與次數，以微縮、光碟或其他各種數位化方式將上列論文重製，並得將數位化之上列論文及論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

※ 讀者基於非營利性質之線上檢索、閱覽、下載或列印上列論文，應依著作權法相關規定辦理。

授權人：賴彥龍

親筆簽名：賴彥龍

民國 96 年　　月　　日

1. 本授權書請以黑筆撰寫，並列印二份，其中一份影印裝訂於附錄三之二(博碩士紙本論文著作權授權書)之次頁；另一份於辦理離校時繳交給系所助理，由圖書館彙總寄交國家圖書館。

國立交通大學

論文口試委員會審定書

本校 生物資訊 碩士班 賴彥龍 君

所提論文:

Comprehensive analysis of knots in proteins

綜合分析蛋白質的打結

合於碩士資格水準、業經本委員會評審認可。

口試委員：黃鎮剛

何信瑩

盧錦隆

指導教授：

黃鎮剛

所長：何信瑩

中 華 民 國 96 年 月 日

綜合分析蛋白質的打結

學生：賴彥龍　　指導教授：黃鎮剛

國立交通大學生物資訊學系（研究所）碩士班

摘　　要

近年來，隨著在 Protein Data Bank (PDB)已知的結構越來越多，蛋白質打結的結構也被不斷的發現出來。研究指出打結的區域有助於配體的結合以及酵素的活性，蛋白質中植物光敏色素的色基結合的範圍以及酮醇酸還原異位酶或 SPOU 甲基轉移酶。此外，在文獻中有著很多被定義錯誤的打結，因此，我們做出了 protein Knot 的 database 以及可提供使用者更容易去觀察 protein 是否打結，在 pKNOT 上面使用者可以輸入 pdb id 或者上傳 pdb structure 的座標。pKNOT 將會使用 Taylor's smoothing algorithm 去判斷是否有打結。所有偵測 protein 的 knot 都會以 Java 的圖形介面方式呈現，讓使用者更容易去得知其位置跟結構。我們深信 pKNOT 對於生物結構以及生物領域，會有著很大的幫助。

Comprehensive analysis of knots in proteins

student：Yan-Long Lai　　Advisors：Dr.Jenn-Kang Hwang

Department（Institute）of Bioinformatics
National Chiao Tung University

ABSTRACT

Knotted proteins are more commonly observed in recent years due to the enormously growing number of structures in the Protein Data Bank(PDB). Studies show that the knot regions contribute to both ligand binding and enzyme activity in proteins such as the chromophore-binding domain of phytochrome, ketol–acid reductoisomerase or SpoU methyltransferase. However, there are still many misidentified knots published in the literature due to the absence of a convenient web tool available to the general biologists. Here, we present the first web server to detect the knots in proteins as well as provide information on knotted proteins in PDB—the protein KNOT (pKNOT) web server. In pKNOT, users can either input PDB ID or upload protein coordinates in the PDB format. The pKNOT web server will detect the knots in the protein using the Taylor's smoothing algorithm. All the detected knots can be visually inspected using a Java-based 3D graphics viewer. We believe that the pKNOT web server will be useful to both biologists in general and structural biologists in particular.

誌　　謝

首先誠摯的感謝指導教授黃鎮剛博士，老師悉心的教導使我得以一窺生物資訊領域的深奧，不時的討論並指點我正確的方向，使我在這些年中獲益匪淺。老師對學問的嚴謹更是我輩學習的典範。

本論文的完成另外亦得感謝多媒體系所的志偉先生大力協助，及醫學資訊所佳昇先生的支持。因為有你們的體諒及幫忙，使得本論文能夠更完整而嚴謹。

兩年裡的日子，實驗室裡共同的生活點滴，學術上的討論、言不及義的閒扯、讓人又愛又怕的宵夜、趕作業的革命情感、因為睡太晚而遮遮掩掩閃進實驗室.....，感謝眾位學長姐、同學、學弟妹的共同砥礪(墮落?)，你/妳們的陪伴讓兩年的研究生活變得絢麗多彩。

感謝建華、草爸、水蛭、書瑋學長、蔚倫學姐們不厭其煩的指出我研究中的缺失，且總能在我迷惘時為我解惑，也感謝鍾錡同學的幫忙，恭喜我們順利走過這兩年。實驗室的松桓學弟當然也不能忘記，他的幫忙及搞笑我銘感在心。

最後，謹以此文獻給我摯愛的雙親。

Content

中文提要	……	i
英文提要	……	ii
誌謝	……	iii
目錄	……	iv
表目錄	……	v
圖目錄	……	vi
1	Introduction……	1
2	Dataset……	3
3	Method……	4
3.1	Data set and pre-computed knots……	4
3.2	Input format……	5
3.3	Output format and visualization of chains and knots	6
4	Result……	7
4.1	The trefoil knot……	7
4.1.1	Methyltransferase……	7
4.1.2	Transcarbamylase……	9
4.1.3	Sunperantigen YPMa - the dynamic knot……	9
4.1.4	Methionine Adenosyltransferas……	10
4.1.5	Carbonic anhydrase……	10
4.2	The figure-eight knot……	11
4.2.1	Bacterial phytochrome……	11
4.2.2	The inner layer core protein of bluetongue virus	12
4.2.3	The LIM-1db1-LID chimeric protein	12
4.2.4	Ketol-acid reductoisomerase……	13
4.3	The 5_2 knot……	13
4.3.1	UbiquitinC-terminal hydrolase……	14
5	Discussion……	15
	Acknowledgement……	17
參考文獻	……	18

表目錄

Table 1 The proteins with a knot in the PDB…… 21
Table 2 Mtase or MTase-like proteins with a βαβ knot motif… 24

圖目錄

Figure 1 …… 21
Figure 2 …… 22
Figure 3 …… 23
Figure 4 …… 24
Figure 5 …… 25
Figure 6 …… 26
Figure 7 …… 27
Figure 8 …… 28
Figure 9 …… 29
Figure 10 …… 30
Figure 11 …… 31
Figure 12 …… 32
Figure 13 …… 33
Figure 14 …… 34
Figure 15 …… 35
Figure 16 …… 36

1. Introduction

The proteins with knotted conformations become more common in the Protein Data Bank (PDB). We carried out comprehensive search of all protein structures from PDB as of 2006/5/3 and identified all knotted proteins. We found that the formation of a knot can be roughly classified into two types in terms of its origin: one type of knot can be clearly identified to arise from a local knotting structural motif, while the other type arises from the global dispositions of the structural elements. In the X-ray structures, the proteins with a knot of type I are (1) methyltransferase and (2) transcarbamylase; the proteins with a knot of type II are (1) carbonic anhydrases, (2) ketol-acid reductosiomerase, (3) ubiquitin hydrolase (4) methionine adenosyl transferase, (5) the chromophore bining domain of bacterial phytochrome and (6) the inner core shell component protein of bluetongue virus. We found a number of knotted hypothetical proteins, all of which have a knot of the first type, i.e., a knot arising from a local knotting structural motif. Interestingly, we found the presence of both knotted and unknotted states in the solution structures of two proteins - one is a chimeric protein comprising LIM domains of LMO2 and the LID domain of ldb1, and the other is a microbial protein belonging to the family of superantigens. Together with another recent study that knotted protein can fold efficiently and behaves remarkably similarly to other unknotted proteins under equilibrium conditions, our results suggest that the knotted and unknotted conformation may dynamically transform from one state to the other in some proteins.

Though, mathematically, the identification of a general knot is a difficult problem, it is relatively easy to detect knots in proteins, since only 3 types of knots are currently being identified in the PDB: the trefoil knot, the figure-eight knot and a knot type with 5 crossings (denoted by 5_2) [1,2]. Still, the knots in proteins were sometimes erroneously identified in the case of, for example, the

SET domain2,3. The reasons may be due to the presence of mobile loops or hydrogen-bonded loops, missing residues, or just visual error in tracing the entangled protein chains.

2. Datasets

We used the PDB dataset dated 03 May 2006. This dataset consists of 36646 proteins including 4350 NMR protein structures. We scan every chain of X-ray structures and every model of NMR structures for the knots. The chains with breaks of discontinuities are visually checked for their relevance in knot formation. If the proteins have a missing gap so large that it is improper to simply connect the two ends of the missing fragment to complete the chain, these proteins will not be taken into consideration.

3. Method

The pKNOT web server detects the knot in a protein by smoothing the protein chain using the Taylor's algorithm . The algorithm first fixes both N and C termini in space, then repeatedly smoothes and straightens the protein chain. The chain is reduced in such a way that, with details of the chains eliminated, the knot can be easily detected. If the protein does not contain a knot, the chain will simply shrink into a straight line. The Taylor's algorithm formally goes as follows: Let the protein chain of length N be described by $(r_1,r_2,\ldots,r_N)$, where r_i is the coordinate of the i-th C_α atom. A new coordinate r'_i is taken to be $r'_i = (r_{i-1}+ r_i+ r_{i+1})/3$, where $2\le i\le N-1$. The termini remain fixed, i.e. $r_i=r_1$ and $r'_n= r_n$. The iterative procedure will continue to progressively smooth the chain. The main idea is to prevent the chains from passing through each other. This is done by checking that the triangles defined by $\{r_{i-1},r_i, r'_i\}$ and $\{ r_i, r'_i ,r_{i-1}\}$ do not intersect any line segments defined by $\{r_{i-j},r_j\}$ for $j<i$ and $\{r_j,r_{j+1}\}$ for $j>i$. In practice, most protein chains reduce to a straight line defined either by two termini or to an obvious knot in less than 50 iterations. However, there are cases that will take 500 or more iterations to converge. Figure 1~3 shows a typical example of a chain-smoothing procedure from the original structure of the chromophore-binding domain of bacterial phytochrome (1ZTU) to the final smoothed chain that can be easily identified to contain a figure-eight knot.

3.1 Data set and pre-computed knots

To speed up the web server, we pre-computed all proteins in the PDB as of January 12, 2007, which consists of 41013 proteins comprising 34971 X-ray structures and 6042 NMR protein structures. The crystal structures of homologous protein chains (even those with identical sequences) as well as the solution structures of the same protein were checked for the presence of knots. The chains with breaks or discontinuities are visually checked for their relevance in knot formation. If the proteins have a missing gap so large that it is

improper to simply connect the two ends of the missing fragment to complete the chain, the identified knots will be disregarded. All final smoothed chains that appear to form a knot, i.e. not a simple straight line, were visually examined to decide whether these knots are authentic knots, slipknots or artificial knots caused by large breaks in the chains. The knots in proteins are quite simple in that they can be visually identified, and no sophisticated analysis [such as the Jones polynomials or others] is required. In summary, pKNOT provides information about all knotted proteins, such as their protein classes, their knotted types and the cores and depths of the knotted regions. The core is the smallest region that will remain knotted when the residues are successively deleted from both ends, and the depth is the product of the number of residues that must be deleted from both ends in order to free the knot. Users can also upload the protein structure coordinates in the PDB format and the pKNOT server will progres-sively smooth the chains on the fly and then present the final smoothed chain as well as the original chain in a JAVA-based 3D graphics viewer AstexViewer for users to inspect.

3.2 Input format

The web page of the pKNOT web server is shown in Figure 4. The users can either type in the PDB ID or upload a structural file in the PDB format. In the latter case, the default iteration number is set to 500 and the collision threshold, to 0.5A ˚ . The user can either ignore or preserve the breaks in the chain when smoothing the chain. The former option will close the breaks by using the shortest line segment connecting the breaks, while the latter option preserves the breaks in the chain and smoothes each individual segment, keeping the endpoints of each segment fixed. The default is set to ignore the breaks in the chain. The users can also choose from the pull-down menu the number of iterations to smooth the chain. The collision threshold is the distance threshold to determine whether a line segment will intersect the triangle during the

smoothing procedures.

3.3 Output format and visualization of chains and knots

Upon query, pKNOT will return a table of the CHAIN, LENGTH, KNOT TYPE and DISPLAY STRUCTURE Figure 5~7. When clicking on the column of KNOT TYPE, the server will return a list of all the proteins of the given knot type. PKNOT also provides the molecular viewer AstexViewer so that the users can visualize and manipulate in real time the protein structure and the knot in the protein. Both the original structure and the knot are shown in the same graphics window and the user can toggle on and off one of them for easy inspection.

4. Results

We found that the formation of a knot can be roughly classified into two types in terms of its strorigin: one type of knot can be clearly identified to arise from a local knotting structural motif, while the other type arises from the global dispositions of the structural elements. In the X-ray structures, the proteins with a knot of type I are (1) methyltransferase and (2) transcarbamylase; the proteins with a knot of type II are (1) carbonic anhydrases, (2) ketol-acid reductosiomerase, (3) ubiquitin hydrolase (4) methionine adenosyl transferase, (5) the chromophore bining domain of bacterial phytochrome and (6) the inner core shell component protein of bluetongue virus.

4.1 The trefoil knot

The trefoil knot (or the threefoil or overhand knot) is the simplest knot of all, which is characterized by 3 crossings. It is mathematically denoted as a 3_1 knot, where the subscript 1 indicates that there is only one prime knot with 3 crossings. The proteins with a trefoil knot are (1) methyltransferase, (2) transcarbamylase, (3) methionine adenosyltransferase, (4) carbonic anhydrase and (5) YMPa supantigen.

4.1.1 Methyltransferase

The S-adenosyl-L-methionine (AdoMet)-dependent methyltransferase (MTase) can be grouped into the knotted type and the unknotted. The majority of MTases, like the classical MTase (1vid and 1mwi), the SET domain[1] (1ml9) or others (1nth and 1msk), do not have a knot[3]. But other MTases like SpoU[4,3] and TrmD[5,6] (or collectedly called as SPOUT) are known to have a 3_1 knot at the AdoMet-binding site[7,3]. We have identified 23 knotted MTases or MTase-like hypothetical proteins in the PDB (Table 1). Interestingly, 43% of them (i.e., 10

out of 23) are annotated as 'hypothetical protein' in PDB. We clustered them into 8 cluster groups using BLASTClust[8] with 30% sequence identity as the similarity threshold. The knot of MTase arises from a local structural motif that is of a βαβ topology. The knot motif is made up of 3 parallel β-strands arranged in the order of β_2, β_1 and β_3. A typical example is shown in Figure 8~11. Because of the right-handedness of the βαβ motif, the helix $\alpha_{1,2}$and $\alpha_{2,3}$ occur on the opposite sides of the plane spanned by β_1, β_2 and β_3 strands. The unique feature of this motif is that the peptide segment comprises β_3 and its entailing loop threads through the peptide hoop comprising β_1, $\alpha_{1,2}$ and the loop between them and this results in a 3_1 knot. We refer to this motif as the βαβ-knot motif. This knot motif is important in contributing most of the AdoMet-binding and catalytic activity[6].

CATH[9] and SCOP[10] define an alpha/beta knot superfamily for MTase. Both data sets currently contain identical sets of proteins (19) in alpha/beta knot superfamily. Interestingly, there is one protein, 1oy5[11], which does not have a knot. Though 1oy5 belongs to the TrmD MTase family known to have a knot, its structure does not have a knot. 1oy5 has a βαβknot-like as that of a typical knotted MTase 1mxi:A[4] Figure 12, but its chain segment composed of β_3 and its entailing loop fails to thread through the hoop formed by β_1, $\alpha_{1,2}$ and the loop between them. However, there is concern[3] that the absence of a knot in 1oy5 may be due to its poor structural quality or its misfolded inactive structure[10]. At present, there are 5 knotted MTases (1x7o, 1x7p, 1zjr, 1v6z and 2cx8; see Table 1) that have not yet been included in SCOP or CATH data sets.

4.1.2 Transcarbamylase

There are only 3 transcarbamylases (1yh0, 1yh1 and 1js1) have a 3_1 knot. The knot of transcarbamylase arises from a local structural motif similar to that of MTase. The knot motif is also of the βαβ topology, as in the case of methyltransferase. An example (1yh1) is shown in Figure 13. In this particular example, the knot motif comprises 4 parallel β strands arranged in the order of $\beta_3\beta_2\beta_1\beta_4$. The segment composed of β_4 strand and its entailing loop goes through the hoop made up of β_1, α_{12} and the loop between them. Another unqiue feature is that the loop between β_1 and α_{12} contains lots of prolines. It is instructive to compare the knotted region of the knotted transcarbamylase with that of the unknotted one. One of the conspicuous differences between these two structures is that 1yh1 has a much longer, proline-rich loop (TYHPKPLN)[12] while 1als has a much shorter loop (GDGN). The other knotted transcarbamylase (1yh0 or 1yh1), which has 31% sequence identity with 1js1, also have a long proline-rich loop (TWAPHPRPKPQ) [13]. The proline-rich loop is rather rigid and, due to its longer length, forms a much larger hoop for a peptide segment to thread through it to form a knot. Additionally, the proline-rich loop also functionally important since it forms part of the active site[13,12].

4.1.3 Sunperantigen YPMa – the dynamic knot

YPMa[14] is a microbial protein belonging to the family of superantigens, which are able to excessively activate T cells by binding to the T cell antigen receptor (TCR). Both the solution structure (1poq) and the crystal structure (1pm4) of YPMa have been solved[14]. Thought the crystal structure does not contain a knot, we found that one of its solution structures contains a 3_1 knot. It will be interesting to see how the knot comes about in the solution structures.

We compare the different NMR structures of YPMa. In the knotted conformation, the N ad C termini thread through the loop (in cyan), creating 3 crossover points characterics of a 3_1 knot. If any one of them is missing, the knot will not form. This type of knot is different from what we have considered so far and may be termed as the transient knot, since its formation depends on the conformational fluctuation. The knots in MTase, transcarbamylase or proteins in the later section may be termed as the intrinsic knots, since they do not depend on the dynamics of the protein conformation. Since X-ray structure of YPMa does not contain a knot, one may tend to rule out the possibility of a knot in the solution structure. However, since YPMa packs as a trimer in the crystal structure, but behaves as monomer in solution[14], the crystal structure may not reflect some distinguished features inherent in the solution structure. On the other hand, though, topologically speaking, this knot is a real one, whether a knot physically exists in YPMa depends on the quality of its solution structures. Obviously, further experimental work is required to resolve this issue.

4.1.4 Methionine Adenosyltransferase

Methionine Adenosyltransferase (MAT) is the solely enzyme that is responsible for the synthesis of AdoMet, the most important methyl donor present in living organisms. MAT contains a 3_1 knot[15,16]. The knot is due to the global fold of MAT, instead of a local knot motif like MTase or transcarbamylase. There are 12 MAT proteins (or 26 chains) in the PDB (see supplementary material).

4.1.5 Carbonic anhydrase

Carbonic anhydrase has long being recognized to contain a 3_1 knot[17] Figure 14. We have identified 230 chains of carbonic anhydrase in PDB. Like MAT, carbonic anhydrase has a knot due to its global fold instead of a isolated knot

motif.

4.2 The figure-eight knot

The figure-eight knot is characterized by 4 crossover points, alternately under and over. There is only one prime knot with 4 crossings and is denoted as the 4_1 knot. The proteins with a 4_1 knot are (1) bacterial pyhtochrome, (2) the core protein of bluetongue virus, (3) ketol-acid reductoisomerase and (4) a LIM-ldbl-LID chimeric protein.

4.2.1 Bacterial phytochrome

Recently, Wagner *et al.*[18] have identified a light-sensing trefoil knot in the chromophore-binding domain (CBD) of phytochrome. This knot originates from the N-terminal peptide fragment passing through a lasso-like loop. The knot region was implicated in chromophore binding and in protein folding. The unique topology of the knot region is shown to be common to all red/far-red photochromic phytochromes and plays a key role in light sensing. However, we found the knot in CBD is in fact a figure-eight knot instead of a trefoil knot. The N-terminal segment (in blue) is seen to pass through the lasso-like loop (in orange). But it is a bit harder to discern other crossover points in the protein. There are disordered missing residues in the model, but these residues play little role in the cause of the knot. It is interesting to investigate why the figure-eight knot in BCD is mistaken for a trefoil knot. The figure-eight knot is characterized by 4 crossover points, alternately under and over. If we trace the protein chain from the N terminus, the chain first goes under at the 1st crossing, goes over at 2nd, then again goes under at the 3rd and finally goes over at the 4th. It will be relatively easy to trace the chain at the 1st and 2nd crossover points, but, since the other part of the protein chain is rather entangled, the chain may be incorrectly

traced at the 3rd and 4th crossover points. Consequently, if the chain is incorrectly traced to go over at the 3rd crossing and under at the 4th, the total crossing number will reduce by 1. The knot in BCD will then be misidentified to be a trefoil knot, which is characterized by 3 crossover points.

4.2.2 The inner layer core protein of bluetongue virus

The inner core proteins of bluetongue virus presents an interesting example that, of the two chains A and B with identical sequences in the protein complex, only chain A has a knot while chain B does not. The inner core, VP3(T2), of bluetongue virus is an icosahedral shell[19], which consists of 60 copies of two sets, 2btv_A and 2btv_B. Both X-ray structures have been solved (2btv_A: residues 57-901 and 2btv_B: 1-803, 814-901). Give that these two chains have identical sequences, their structures are significantly different and the root-mean-square deviation for Cαs is around 3.0 Å. Most interestingly, 2btv_A does not have a knot, while 2btv_B has a figure-eight knot or a 4_1 knot[17]. In 2btv_B, the last 7 residues (895-901) of the C terminus threads through the hoop consisting of residues 228-267, but in 2btv_A, the C terminal chain only lie on the same side of the corresponding hoop region. It has been pointed out that the structural differences between 2btv_A and 2btv_B are due to their completely different structural environments[19]. Our results have an intriguing implication: the same protein may adopt both knotted and unknotted structures.

4.2.3 The LIM-1db1-LID chimeric protein

FLIN2 (PDB ID: 1j2o) is a chimeric protein[20] comprising the 80-residue LIM domains of LMO2, followed by an 11-residue flexible linker (GGSGGHMGSGG), and then the 40-residue ldb1-LID. Its structure has been solved by the NMR method[20]. We find that there is a 4_1 knot contained in one of

the solution structures. The knot originates from both ldb-LID domain and the linker threading through the loop comprising the residues 40-47. This loop is highly dynamic and undergoes certain chemical exchange process on the microsecond-millisecond time scale[20]. We observe that, 6 out of 20 solution structures, the ldb1-LID domain threads through the loop. At present, FLN2 as well as 1YPMa are only two examples of the so-called 'transient knot' in the PDB. The figure-eight knot of FLN2 is mathematically authentic, though whether a knot physically exists in FLN2 depends on the quality of its solution structures – this is similar to the case of 1YPMa.

4.2.4 Ketol-acid reductoisomerase

Ketol-acid reductoisomerase (KARI) is a key enzyme involved in the biosynthesis pathway of the branched chain amino acid such as valine, leucine and isoleucine[21]. KARI exists in 2 forms: the short form (class I)[22] and the long form (class II)[23], which is 150-160 residues longer. Both forms have a NADPH-binding N-terminal domain and an all-helical C-terminal domain. The C-terminal domain is important in contributing most of the ketol-acid substrate-binding site. The long form has a duplication of the C-terminal domain found in the short form[22]. The dimeric C-terminal domain of the long form gives rise to a 4_1 knot Figure 15 in the long form[17]. The short form does not have a knot but may form a knot if its C-terminal domain is fused with that of another short form[22]. It has been suggested[17,22,23] that the C-terminal domain of the long form is derived *via* domain duplication. There currently are 3 knotted KARIs currently in the PDB: 1qmg, 1yve and 1yrl.

4.3 The 5_2 knot

There are 2 types of knot with 5 crossings: the 5_1 and 5_2 knots Figure 16. Only the 5_2 knot has been identified in the protein structure and, as of writing, no proteins with 6 or more crossings have been identified in the PDB.

4.3.1 UbiquitinC-terminal hydrolase

Ubiqutine C-terminal hydrolase (UCH) is the only protein that has a knot with 5 crossings, which is a 5_2 knot (Fig. 10)(18). UCH possess the sequence signature of cysteine protease: a conserved catalytic triad of cysteine, histidine and aspartic acid[24] . The knot region is important in bringing the catalytic residues into popper relative positions in space: the fragment (in orange) where both catalytic histidine and aspartic acid are located threads the hoop (color ramped from cyan to green) where the catalytic cysteine is located, thus bring these three catalytic residues close together into a triadic formation.We have identified 4 knotted UCH (1cmx, 1uch, 1xd3 and 2etl) in the PDB.

5. Discussion

It will be interesting to compare our results with those of the recent work by Lua and Grosberg[25]. This work identified 19 knot proteins using the RANDOM method from the PDB-REPRDB[26]. This data set contains 4716 representative protein chains whose pair sequence identity < 30% and 3D structural similarity index $D_{\max} \geq 30\text{Å}$. The structural similarity index $D_{\max}$ is defined as the maximum Cα distance between the corresponding residue pairs in the compared structures. However, we believe that 5 of the identified 19 knotted proteins (i.e., 1t0h:B, 1gku:B, 1u2z:C, 1m72:B and 1xi4:C) are questionable due to very large structural breaks in these proteins (for example, in some case, the structural gap is around 50 Å wide). The knots in fact arises from arbitrarily connecting the structural discontinuities with virtual bonds: (1) The X-ray structure of 1t0h:B has missing residues 414-424. A knot will be created if we simply connect a straight virtual bond between 413 and 425 with a distance of around 32 Å between them. (2) The X-ray structure of 1gku:B has many missing residues near the N terminus (7, 13, 22-32), and, if these ends are simply connected to each other, a knot will be created. (3) Though 1u2z catalyze AdoMet-dependent methylation of histone H3 Lys79[27], it does not have a typical βαβ knot motif. The X-ray structure of 1u2z:C has missing residues 570-573 and 575. If these structural gap (around 52 Å in distance) are simply connected by virtual bonds, a 4_1 knot arises. Furthermore, the 1u2z:A, which has identical sequence and similar missing residues in the X-ray structure with 1u2z:C, does not have a knot even with the structural gaps connected by virtual bonds. (4) The structure of 1m72:B misses residues 192-199, creating a gap of around 58 Å in the structure. The knot arises from connecting the Arg191 and Thr200. (5) The PDB file of 1xi4T contains unusual format which will give a distance of around 150 Å between Leu838 and Gln839 and around 150 Å between His1279 and Ala2780 for the chain C. If these unrealistic connections

are removed, the knot disappears.

In general, we have identified two types of knots existing in the structures in PDB. The first type is referred to as the intrinsic knots. The second type is referred to as the dynamic knot.

Acknowledgement

This work is supported by National Science Council, National Research Program of Genomic Medicine, and the University System at Taiwan-Veteran General Hospital Grant. We are grateful to both the hardware and software supports of the Structural Bioinformatics Core Facility at National Chiao Tung University.

References

1. Jacobs SA, Harp JM, Devarakonda S, Kim Y, Rastinejad F, Khorasanizadeh S. The active site of the SET domain is constructed on a knot. Nat Struct Biol 2002;9:833-8.

2. Taylor WR, Xiao B, Gamblin SJ, Lin K. A knot or not a knot? SETting the record 'straight' on proteins. Comput Biol Chem 2003;27:11-5.

3. Mosbacher TG, Bechthold A, Schulz GE. Structure and function of the antibiotic resistance-mediating methyltransferase AviRb from Streptomyces viridochromogenes. J Mol Biol 2005;345:535-45.

4. Lim K, Zhang H, Tempczyk A, Krajewski W, Bonander N, Toedt J, Howard A, Eisenstein E, Herzberg O. Structure of the YibK methyltransferase from Haemophilus influenzae (HI0766): a cofactor bound at a site formed by a knot. Proteins 2003;51:56-67.

5. Ahn HJ, Kim HW, Yoon HJ, Lee BI, Suh SW, Yang JK. Crystal structure of tRNA(m1G37)methyltransferase: insights into tRNA recognition. Embo J 2003;22:2593-603.

6. Elkins PA, Watts JM, Zalacain M, van Thiel A, Vitazka PR, Redlak M, Andraos-Selim C, Rastinejad F, Holmes WM. Insights into catalysis by a knotted TrmD tRNA methyltransferase. J Mol Biol 2003;333:931-49.

7. Zhang X, Tamaru H, Khan SI, Horton JR, Keefe LJ, Selker EU, Cheng X. Structure of the Neurospora SET domain protein DIM-5, a histone H3 lysine methyltransferase. Cell 2002;111:117-27.

8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-402.

9. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH- A Hierarchic Classification of Protein Domain Structures. Structures 1997;5:1093-1108.

10. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226-9.

11. Liu J, Wang W, Shin DH, Yokota H, Kim R, Kim SH. Crystal structure of tRNA (m1G37) methyltransferase from Aquifex aeolicus at 2.6 A resolution: a novel

methyltransferase fold. Proteins 2003;53:326-8.

12. Shi D, Morizono H, Yu X, Roth L, Caldovic L, Allewell NM, Malamy MH, Tuchman M. Crystal structure of N-acetylornithine transcarbamylase from Xanthomonas campestris: a novel enzyme in a new arginine biosynthetic pathway found in several eubacteria. J Biol Chem 2005;280:14366-9.

13. Shi D, Gallegos R, DePonte J, 3rd, Morizono H, Yu X, Allewell NM, Malamy M, Tuchman M. Crystal structure of a transcarbamylase-like protein from the anaerobic bacterium Bacteroides fragilis at 2.0 A resolution. J Mol Biol 2002;320:899-908.

14. Donadini R, Liew CW, Kwan AH, Mackay JP, Fields BA. Crystal and solution structures of a superantigen from Yersinia pseudotuberculosis reveal a jelly-roll fold. Structure 2004;12:145-56.

15. Gonzalez B, Pajares MA, Hermoso JA, Guillerm D, Guillerm G, Sanz-Aparicio J. Crystal structures of methionine adenosyltransferase complexed with substrates and products reveal the methionine-ATP recognition and give insights into the catalytic mechanism. J Mol Biol 2003;331:407-16.

16. Komoto J, Yamada T, Takata Y, Konishi K, Ogawa H, Gomi T, Fujioka M, Takusagawa F. Catalytic mechanism of guanidinoacetate methyltransferase: crystal structures of guanidinoacetate methyltransferase ternary complexes. Biochemistry 2004;43:14385-94.

17. Taylor WR. A deeply knotted protein structure and how it might fold. Nature 2000;406:916-9.

18. Wagner JR, Brunzelle JS, Forest KT, Vierstra RD. A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome. Nature 2005;438:325-31.

19. Grimes JM, Burroughs JN, Gouet P, Diprose JM, Malby R, Zientara S, Mertens PP, Stuart DI. The atomic structure of the bluetongue virus core. Nature 1998;395:470-8.

20. Deane JE, Mackay JP, Kwan AH, Sum EY, Visvader JE, Matthews JM. Structural basis for the recognition of ldb1 by the N-terminal LIM domains of LMO2 and LMO4. Embo J 2003;22:2224-33.

21. Biou V, Dumas R, Cohen-Addad C, Douce R, Job D, Pebay-Peyroula E. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65 A resolution. Embo J 1997;16:3405-15.

22. Ahn HJ, Eom SJ, Yoon HJ, Lee BI, Cho H, Suh SW. Crystal structure of class I acetohydroxy acid isomeroreductase from Pseudomonas aeruginosa. J Mol Biol 2003;328:505-15.

23. Tyagi R, Duquerroy S, Navaza J, Guddat LW, Duggleby RG. The crystal structure of a bacterial class II ketol-acid reductoisomerase: domain conservation and evolution. Protein Sci 2005;14:3089-100.

24. Misaghi S, Galardy PJ, Meester WJ, Ovaa H, Ploegh HL, Gaudet R. Structure of the ubiquitin hydrolase UCH-L3 complexed with a suicide substrate. J Biol Chem 2005;280:1512-20.

25. Lua RC, Grosberg AY. Statistics of knots, geometry of conformations, and evolution of proteins. PLoS Comput Biol 2006;2:e45.

26. Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res 2003;31:492-3.

27. Sawada K, Yang Z, Horton JR, Collins RE, Zhang X, Cheng X. Structure of the conserved core of the yeast Dot1p, a nucleosomal histone H3 lysine 79 methyltransferase. J Biol Chem 2004;279:43296-306.

Table 1. The proteins with a knot in the PDB

Protein	PDB ID
The 3_1 knot	
Methyltransferase or methyltransferase-like	1to0_A 1to0_B 1to0_C 1to0_D 1to0_E 1to0_F 1to0_G 1to0_H 1o6d_A 1vh0_A 1vh0_B 1vh0_C 1vh0_D 1vh0_E 1vh0_F 1ns5_A 1ns5_B 1x7o_A 1x7o_B 1x7p_A 1x7p_B 1ipa_A 1gz0_A 1gz0_B 1gz0_C 1gz0_D 1gz0_E 1gz0_F 1gz0_G 1gz0_H 1uaj_A 1uak_A 1ual_A 1uam_A 1p9p_A 1zjr_A 1v2x_A 1j85_A 1mxi_A 1v6z_A 1v6z_B 2cx8_A 2cx8_B 1k3r_A 1k3r_B 1vhk_A 1vhk_B 1vhk_C 1vhk_D 1vhy_A 1vhy_B 1nxz_A 1nxz_B 1z85_B
Transcarbamylas	1yh0_A 1yh1_A 1js1_Y 1js1_Z 1js1_X
YPMa superantigen	1poq_A
Methionine adenosyltransferase	1o90_A 1o90_B 1o92_A 1o92_B 1o93_A 1o93_B 1o9t_A 1o9t_B 1qm4_A 1qm4_B 1fug_A 1fug_B 1mxa_ 1mxb_ 1mxc_ 1p7l_A 1p7l_B 1p7l_C 1p7l_D 1rg9_A 1rg9_B 1rg9_C 1rg9_D 1xra_ 1xrb_ 1xrc_

Carbonic anhydrase	1y7w_A 1y7w_B 1z93_A 1z97_A 1znc_A 1znc_B 1jcz_A 1jcz_B 1jd0_A 1jd0_B 1rj5_A 1rj5_B 1rj6_A 1rj6_B 1v9i_C 2cab_ 12ca_ 1avn_ 1azm_ 1bcd_ 1bic_ 1bv3_A 1bzm_ 1ca3_ 1cah_ 1cai_ 1caj_ 1cak_ 1cal_ 1cam_ 1can_ 1cao_ 1cay_ 1caz_ 1cnw_ 1cnx_ 1cny_ 1cra_ 1crm_ 1czm_ 1dca_ 1dcb_ 1eou_A 1flj_A 1fql_A 1fqm_A 1fqn_A 1fqr_A 1fr4_A 1fr7_A 1fr7_B 1fsn_A 1fsn_B 1fsq_A 1fsq_B 1fsr_A 1fsr_B 1g0e_A 1g0f_A 1g6v_A 1hca_ 1hcb_ 1hea_ 1heb_ 1hec_ 1hed_ 1hug_ 1huh_ 1hva_ 1j9w_A 1j9w_B 1jv0_A 1jv0_B 1kwq_A 1kwr_A 1lg5_A 1lg6_A 1lgd_A 1lzv_A 1moo_A 1ray_ 1raz_ 1rza_ 1rzb_ 1rzc_ 1rzd_ 1rze_ 1t9n_A 1tb0_X 1tbt_X 1te3_X 1teq_X 1teu_X 1tg3_A 1tg9_A 1th9_A 1thk_A 1xeg_A 1xev_A 1xev_B 1xev_C 1xev_D 1yo0_A 1yo1_A 1yo2_A 2ax2_A 2cba_ 2cbb_ 2cbc_ 2cbd_ 2cbe_ 2foq_A 2fos_A 2fou_A 2fov_A 2foy_A 2foy_B 4ca2_ 4cac_ 5ca2_ 5cac_ 6ca2_ 7ca2_ 8ca2_ 9ca2_ 1a42_ 1am6_ 1bn1_ 1bn3_ 1bn4_ 1bnm_ 1bnn_ 1bnq_ 1bnt_ 1bnu_ 1bnv_ 1bnw_ 1ca2_ 1ccs_ 1cct_ 1ccu_ 1cil_ 1cim_ 1cin_ 1cnb_ 1cnc_ 1cng_ 1cnh_ 1cni_ 1cnj_ 1cnk_ 1cva_ 1cvb_ 1cvc_ 1cve_ 1cvf_ 1f2w_A 1g1d_A 1g3z_A 1g45_A 1g46_A 1g48_A 1g4j_A 1g4o_A 1g52_A 1g53_A 1g54_A 1h4n_ 1h9n_ 1h9q_ 1i8z_A 1i90_A 1i91_A 1i9l_A 1i9m_A 1i9n_A 1i9o_A 1i9p_A 1i9q_A 1if4_A 1if5_A 1if6_A 1if7_A 1if8_A 1if9_A 1lug_A 1okl_ 1okm_ 1okn_ 1oq5_A 1ttm_A 1v9e_A 1v9e_B 1xq0_A 1yda_ 1ydb_ 1ydc_ 1ydd_ 1ze8_A 1zsa_ 1zsb_ 1zsc_ 2abe_A 2ca2_ 2h4n_ 3ca2_ 1uga_ 1ugb_ 1ugc_ 1ugd_ 1uge_ 1ugf_ 1ugg_ 1xpz_A 2znc_ 3znc_ 1mua_ 1cvd_ 1cvh_ 1dmx_A 1dmx_B 1dmy_A 1dmy_B 1keq_A 1keq_B 1urt_ 1kop_A 1kop_B 1koq_A 1koq_B
The 4_1 knot	
The core proteins of bluetongue virus	2btv_B
Phytochrome	1ztu_A
FLIN2	1j2o_A
ketol-acid reductoisomerase	1qmg_A 1qmg_B 1qmg_C 1qmg_D 1yve_I 1yve_J

	1yve_K 1yve_L 1yrl_A 1yrl_B 1yrl_C 1yrl_D
The 5_1 knot	
Ubiquitin Hydrolase	1cmx_A 1cmx_C 1uch_ 1xd3_A 1xd3_C 2etl_A 2etl_B

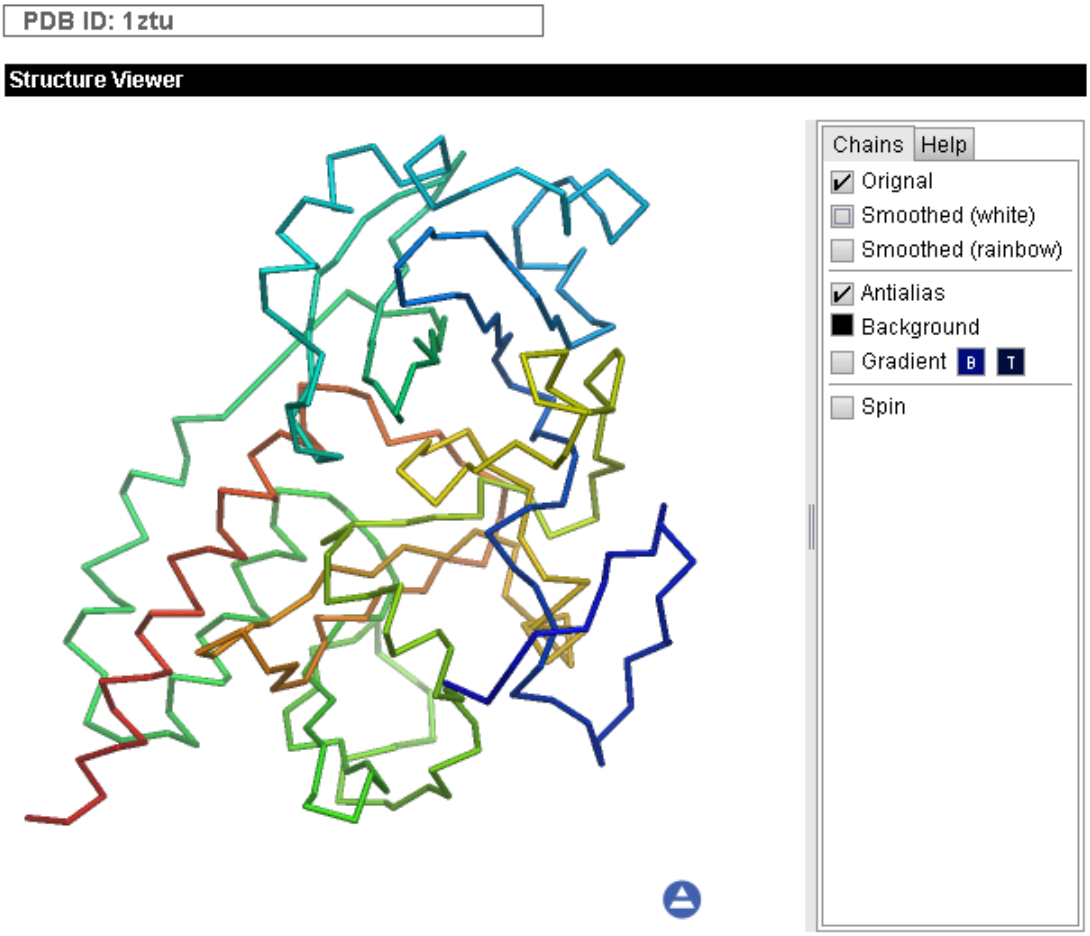


Figure 1
A typical example of a chain-smoothing procedure(orignal).

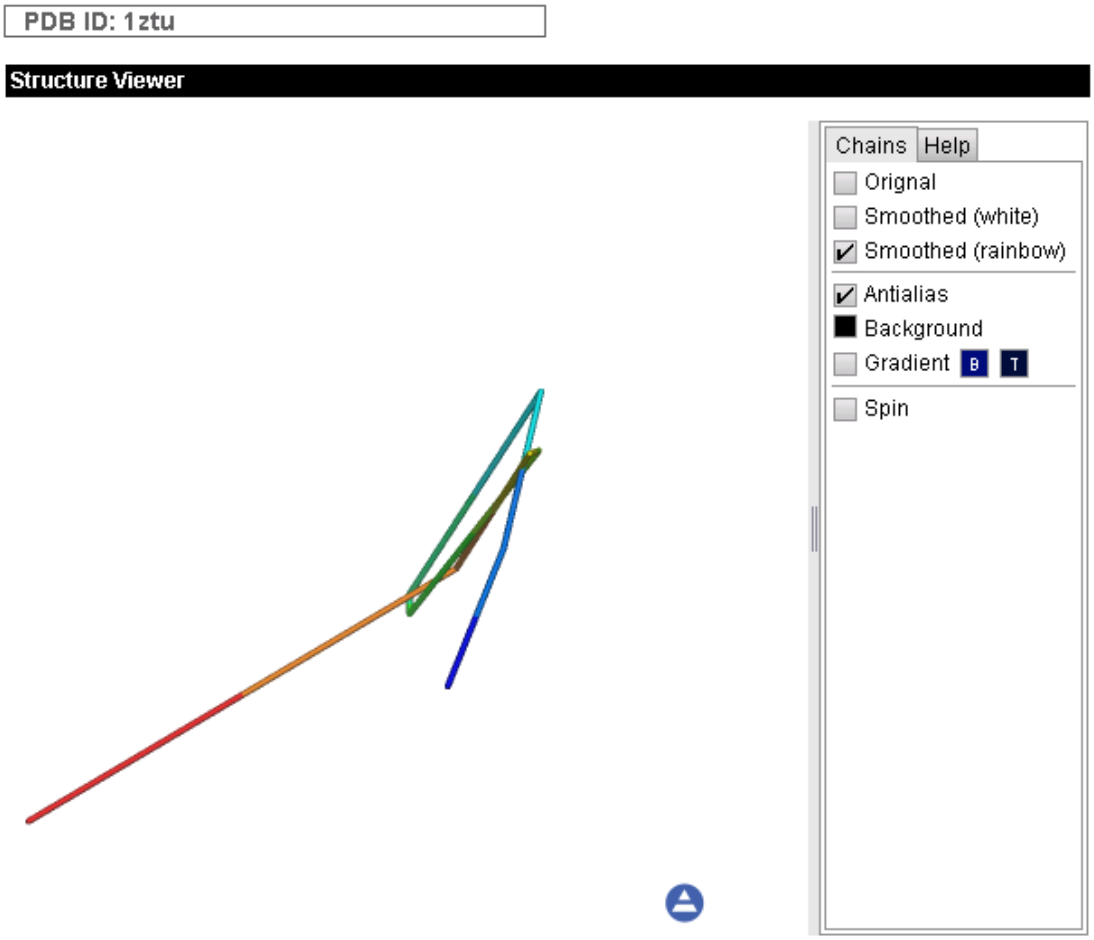


Figure 2
A typical example of chain-smoothing procedure(smoothed).

Figure 3
A typical example of chain-smoothing procedure(original & smoothed).

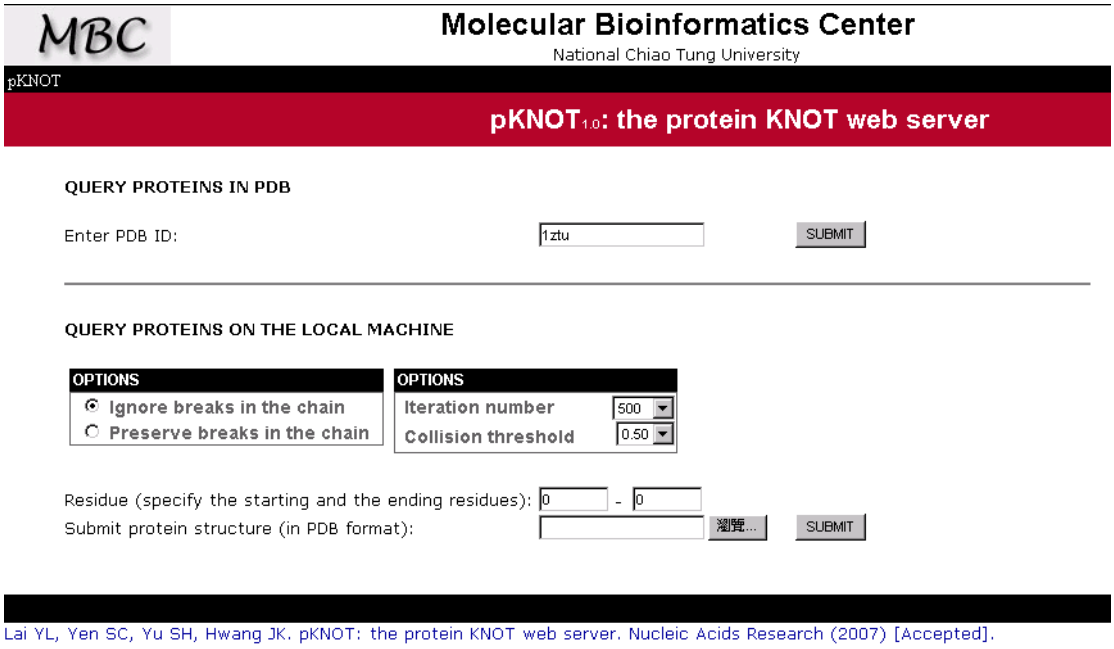


Figure 4
The web page of the pKNOT web server.

3-1 knot

PDB ID	CHAIN	METHOD	LENGTH	CORE	DEPTH	PROTEIN CLASS
1to0	A	X-Ray	157	73-125	2442	Methyltransferase or methyltransferase-like
1to0	B	X-Ray	157	73-125	2442	Methyltransferase or methyltransferase-like
1to0	C	X-Ray	155	74-125	2325	Methyltransferase or methyltransferase-like
1to0	D	X-Ray	157	73-125	2442	Methyltransferase or methyltransferase-like
1to0	E	X-Ray	153	73-125	2146	Methyltransferase or methyltransferase-like
1to0	F	X-Ray	154	74-125	2250	Methyltransferase or methyltransferase-like
1to0	G	X-Ray	157	73-125	2442	Methyltransferase or methyltransferase-like
1to0	H	X-Ray	157	73-125	2442	Methyltransferase or methyltransferase-like
1o6d	A	X-Ray	147	66-117	2077	Methyltransferase or methyltransferase-like
1vh0	A	X-Ray	157	74-125	2475	Methyltransferase or methyltransferase-like
1vh0	B	X-Ray	157	73-125	2442	Methyltransferase or methyltransferase-like
1vh0	C	X-Ray	157	73-125	2442	Methyltransferase or methyltransferase-like

Figure 5
pKNOT will return a table of the CHAIN, LENGTH, PROTEIN CLASS and DISPLAY STRUCTURE by clicking PDB(3-1 knot).

4-1 knot

PDB ID	CHAIN	METHOD	LENGTH	CORE	DEPTH	PROTEIN CLASS
2btv	B	X-Ray	901	209-895	1470	The core proteins of bluetongue virus
1ztu	A	X-Ray	325	34-303	805	Phytochrome
1j2o	A	NMR(14)	114	43-95	880	FLIN2
1qmg	A	X-Ray	595	327-526	22960	ketol-acid reductoisomerase
1qmg	B	X-Ray	595	327-526	22960	ketol-acid reductoisomerase
1qmg	C	X-Ray	595	327-526	22960	ketol-acid reductoisomerase
1qmg	D	X-Ray	595	327-526	22960	ketol-acid reductoisomerase
1yve	I	X-Ray	595	327-526	22960	ketol-acid reductoisomerase
1yve	J	X-Ray	595	327-526	22960	ketol-acid reductoisomerase
1yve	K	X-Ray	595	327-526	22960	ketol-acid reductoisomerase
1yve	L	X-Ray	595	327-526	22960	ketol-acid reductoisomerase
1yrl	A	X-Ray	489	225-422	15368	ketol-acid reductoisomerase
1yrl	B	X-Ray	488	228-432	13053	ketol-acid reductoisomerase
1yrl	C	X-Ray	489	196-416	14578	ketol-acid reductoisomerase
1yrl	D	X-Ray	489	228-366	28396	ketol-acid reductoisomerase

Figure 6

pKNOT will return a table of the CHAIN, LENGTH, PROTEIN CLASS and DISPLAY STRUCTURE by clicking PDB(4-1 knot).

5-2 Knot

PDB ID	CHAIN	METHOD	LENGTH	CORE	DEPTH	PROTEIN CLASS
1cmx	A	X-Ray	234	10-231	44	Ubiquitin Hydrolase
1cmx	C	X-Ray	236	10-230	77	Ubiquitin Hydrolase
1uch	_	X-Ray	230	6-226	35	Ubiquitin Hydrolase
1xd3	A	X-Ray	230	8-226	45	Ubiquitin Hydrolase
1xd3	C	X-Ray	230	8-226	45	Ubiquitin Hydrolase
2etl	A	X-Ray	223	6-217	49	Ubiquitin Hydrolase
2etl	B	X-Ray	223	6-217	49	Ubiquitin Hydrolase

Figure 7

pKNOT will return a table of the CHAIN, LENGTH, PROTEIN CLASS and DISPLAY STRUCTURE by clicking PDB(5-2 knot).

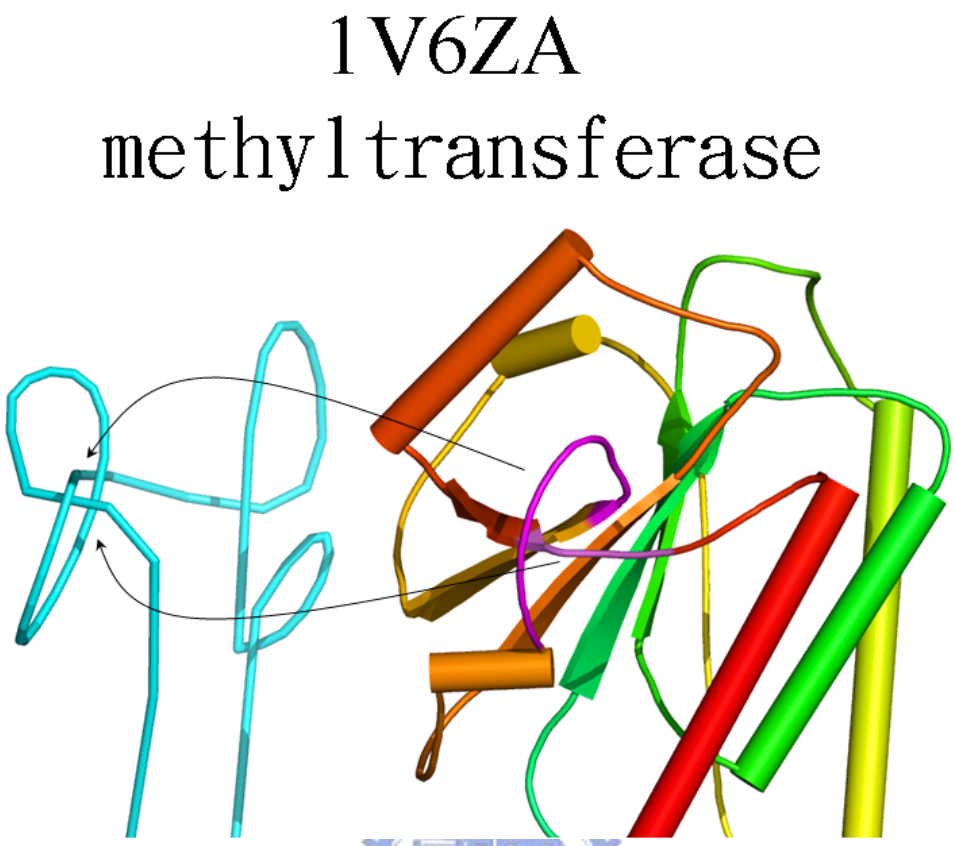


Figure 8
The knot of MTase arise from a local structural motif that is of a topology(1V6ZA).

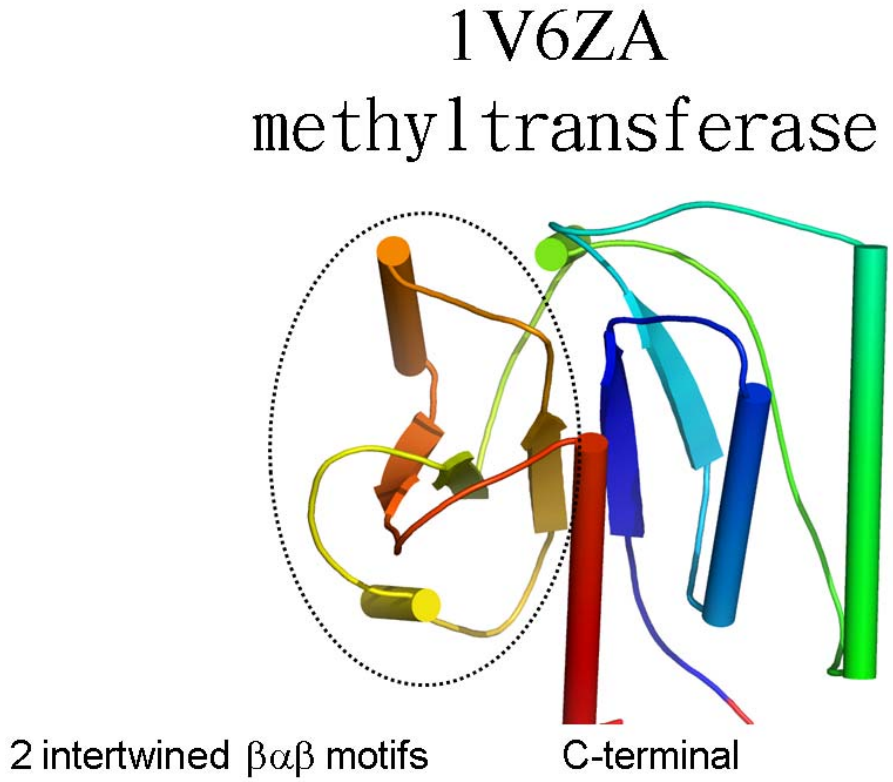


Figure 9
There are 2 intertwined βαβ motifs in the knot of Mtase.

1V6ZA

Two right-handed βαβ motifs:
β1–α1–β2
β2–α2–β3

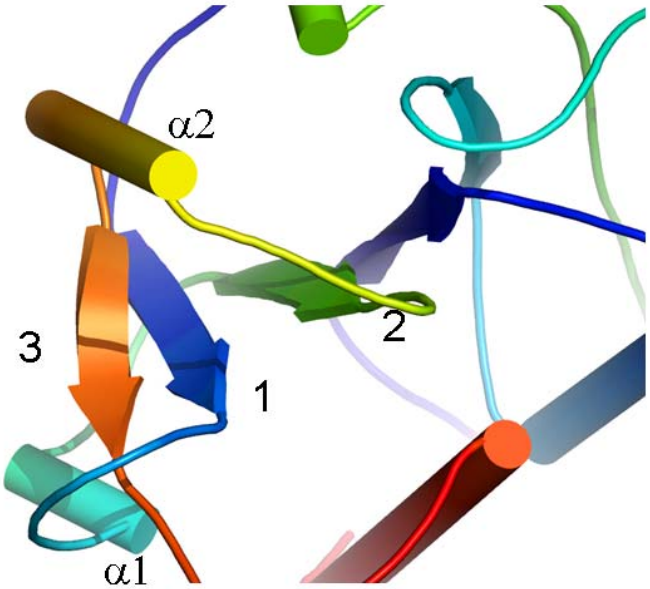


Figure 10
The knot motif is made up of 3 parallel β-strands arranged in the order of β_2, $\beta_1$and β_3.

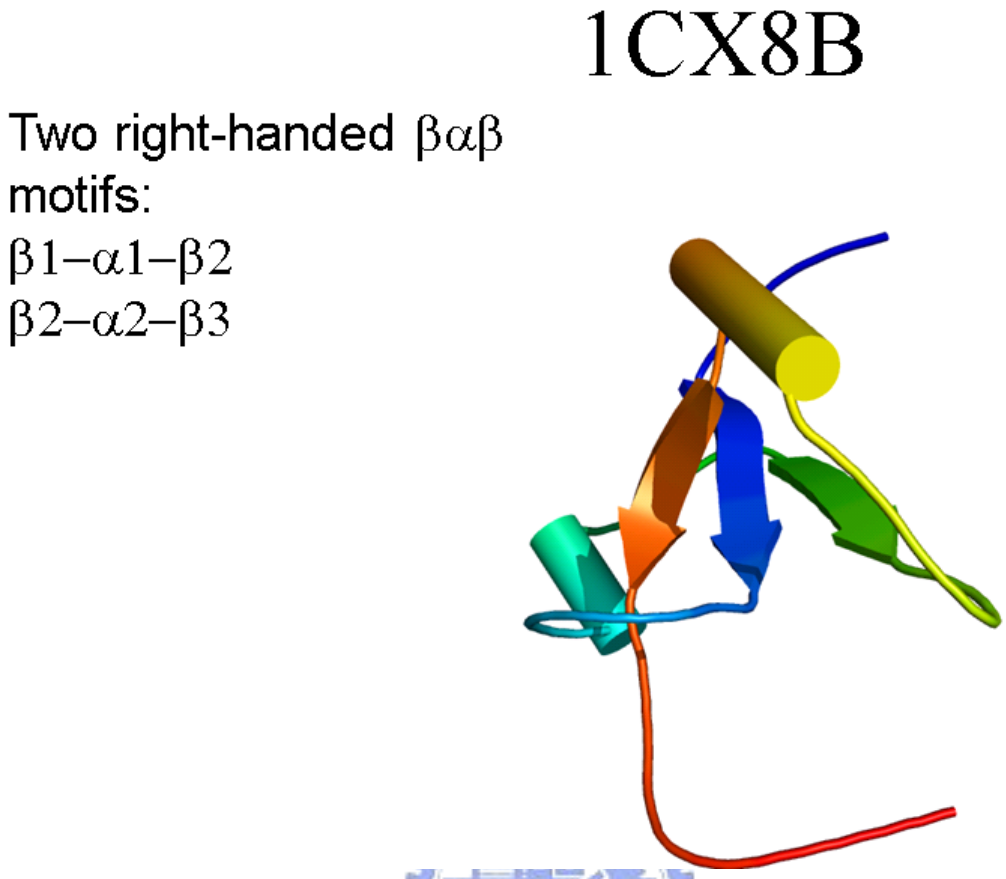


Figure 11

Because of the right-handedness of the βαβ motif, the helix $\alpha_{1,2}$and $\alpha_{2,3}$ occur on the opposite sides of the plane spanned by β_1, β_2 and β_3 strands. The unique feature of this motif is that the peptide segment comprises β_3 and its entailing loop threads through the peptide hoop comprising β_1, $\alpha_{1,2}$ and the loop between them and this results in a 3_1 knot.

1MXI:A
Methyltransferase

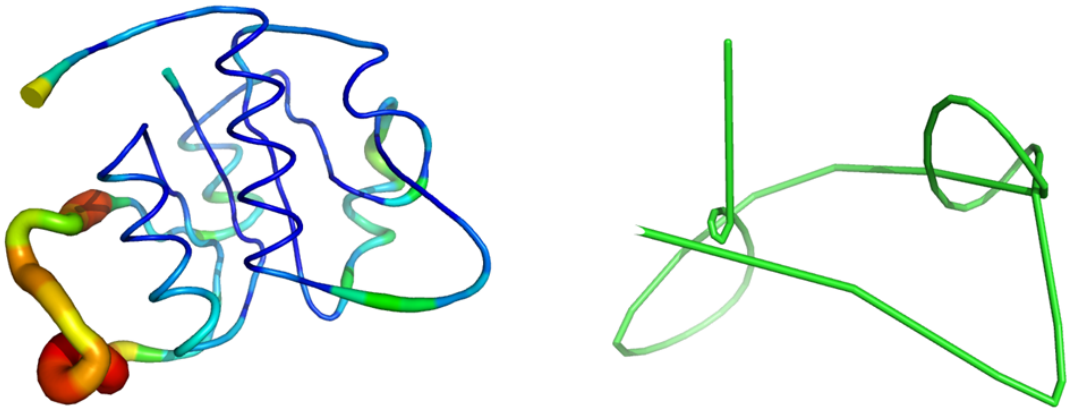


Figure 12
Methytransferase(1MXI:A)

Acetylornithine Transcarbamylas
1yh1_A

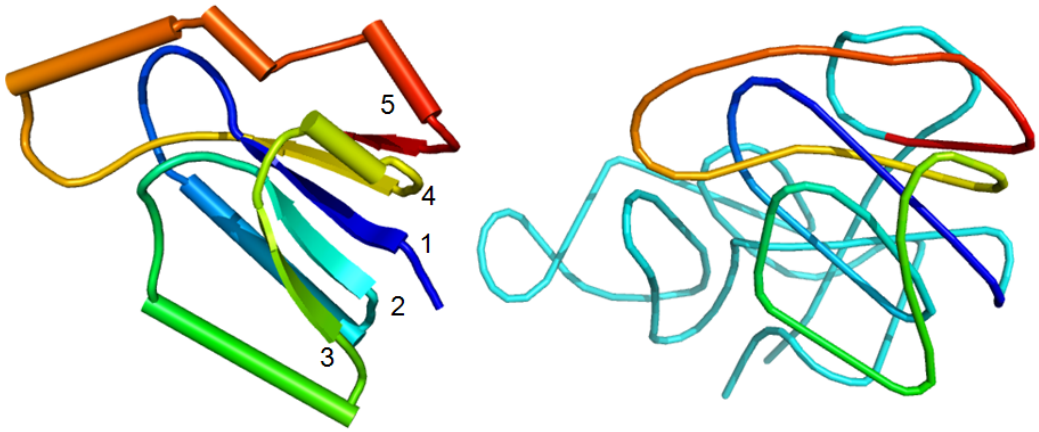


Figure 13
There are only 3 transcarbamylases (1yh0, 1yh1 and 1js1) have a 3_1 knot. The knot of transcarbamylase arises from a local structural motif similar to that of MTase. The knot motif is also of the βαβ topology, as in the case of methyltransferase.

Compound

– Carbonic anhydrase

Conforrmation

Knots

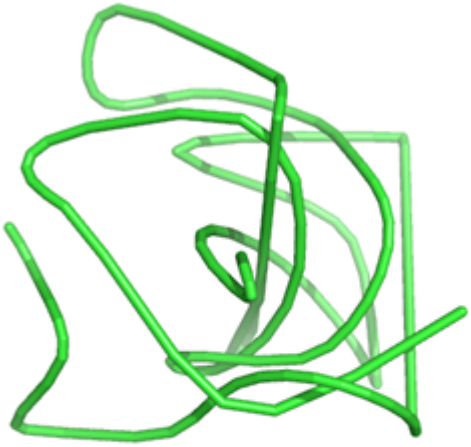
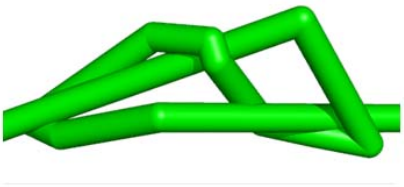


Figure 14
Carbonic anhydrase has long being recognized to contain a 3_1 knot

- **Compound**
  - ketol-acid reductoisomerase
  - Release 2005
- **Conformation**
  - C-domain & N-domain
    - C-domain comprising mainly **helices**
- **Knots**
  - C-domain
  - The figure-eight knot
  - 210-381;279-350+413-486



- **Homologues**
  - **1yrl_A** 1yrl_B 1yrl_C 1yrl_D

Figure 15
The dimeric C-terminal domain of the long form gives rise to a 4_1 knot.

- **Compound**
  - Ubiquitin Hydrolase
- **Conformation**
- **Knots**
  - 5_2

- **Homologues**
  - 1cmx_A 1cmx_C 1uch_ **1xd3_A** 1xd3_C 2etl_A 2etl_B

Figure 16
There are 2 types of knot with 5 crossings: the 5_1 and 5_2 knots