TM-1252

挑戦！　境界領域創造
めざせ！　タンパク質の
立体構造予測の新方法開発

鬼塚　健太郎

March, 1993

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191～5

Institute for New Generation Computer Technology

挑戦！境界領域創造

・・めざせ！ タンパク質の立体構造予測の新方法開発・・

(財) 新世代コンピュータ技術開発機構 鬼塚健太郎 (おにづかけんたろう)

1993 年 3 月 17 日

1 はじめに

筆者の現在の研究分野は、分子生物学とコンピュータ科学、或いは情報科学の接点とも言うべき分野である。今回は、二つの分野の境界領域の研究とはいかにあるべきかについて、筆者自身の研究方針を例として書いてみたい。

筆者の大学時代の専攻は物理学であるから、コンピュータに関しても、生物学に関してももともと専門ではなかった。だからここでは中立の立場で書いてみたいと思う。

2 コンピュータ科学が科学を変える

「生物学者は烏合の衆だ」とは、かの遺伝子 DNA の姿を解明した一人 Watson が、当時のイギリスの生物学者たちを指して、名著「二重らせん」([Watson 68] あるいは [Watson S43]) の中で語った言葉だ。おそらく、子供の頃から野の草花に生命の神秘を感じて生物学者となった人々の当時の研究の在り方は、生命現象すら単なる物質のザワメキとして捕らえられる冷酷無慈悲かつ合理的理性的解析能力をもつ物理学者たちの間で研究していた Watson の目からすれば、さぞかしナイーブなものと写ったことだろう。

前世紀から今世紀にかけての物理学の革命的進歩は、ほとんど全ての学問領域に多大な影響を及ぼした。現代物理学が確立した電磁気学、量子力学などの理論体系、そして数理解析的方法論があらゆる学問領域で研究の在り方に本質的な変革をもたらしたのだ。量子化学の一つの応用分野として始まった分子生物学は、その後、生物学のあらゆる分野に進出し、いたるところで成功してきた。かつては山師のように化石探しをしていた人々によって担われてきた古生物学の分野でも、最近では化石から採取されたコラーゲンの成分分析による生物種の系統の決定などといういかにも分子生物学的な解析的分析方法が横行し始めている。既に化石から遺伝子を抽出し恐竜を復活させるなどという SF 小説のようなことが半ば真面目に議論され始めてもいる。

とすれば、Watson の過激な指摘は結論としてはやはり正しかった。この大きな波に取り残された古典的生物学者は、生物学の本道では生き残れなかった。

現在、或いは近い将来においてかつての物理学のように他の学問領域への大きな影響力をもつ学問領域があるとしたら、それはコンピュータ科学、あるいはより広く情報科学にほかならない。実際、あらゆる学問領域においてコンピュータの利用はごく自然なこととなりつつあり、また情報科学で理論的に確立した方法論も次第にその応用分野を増やしつつある。

Watson・Crick らは古典的生物学に現代物理学的方法論をもって殴り込みをかけた人々だった。そしてナイーブな生物学者らの伝統的研究方法を蹴散らして分子生物学を確立した先駆者だ。境界領域の研究者は、激しい他流試合を真剣勝負で行う訳であり、こうした中で我々は、

将来、例えば「生物情報科学(Bio-informatics)」などと呼ばれることになろう新しい研究分野を開拓しているのだ。そして、それはすでに分子生物学に根本的な変革をもたらし始めている。

3　境界領域研究は根っこから考える

トツゼン話は変わるけれど、最近「ネオテニー（幼体成熟）」という言葉が気に入っている。本来両生類は肺呼吸を行う成体に変態するが、ウーパールーパー（学名はアホロートル）は変態せず鰓呼吸のコドモのまま生殖能力を持ち、繁殖してしまう。このようにコドモの状態で繁殖能力を持つことネオテニーという。

これが面白いのは、ネオテニーが生物の劇的進化の発端となっていることを示唆しているからだ。例えば、非常に繁栄している昆虫類は、一般的な多足類の節足動物の幼体、つまり三対しか足をもたない状態からのネオテニーらしい。また脊椎動物は、成体になると固着生活になるホヤような生物のオタマジャクシのように水中を泳ぎ回る幼体からのネオテニーかもしれない。我々人間もまた、様々な点で類人猿の乳幼児に似ているので、猿のネオテニーだという学者もいる。

ネオテニー的進化の意味するところは興味深い。発達し、適応において特化したオトナの状態より、未熟なコドモの状態がむしろ新しい方向への進化の発端となっていることを示唆しているからだ。

生物の進化では、ネオテニーに限らず一般により下等かつ未発達なもののほうが、未知の方向への爆発的進化を起こしやすいことあるように思われてくる。つまり、適応において特化したものではなく、未適応な進化系統的に幹に近いものほど新しい方向への進化の始点となっているのである。現在哺乳動物以上に繁栄している鳥類も、中生代三畳期以来のもっとも形態的には下等な恐竜である獣脚類の一種が、突如として羽毛の生えた翼を発達させたものだという説が出されている(興味のある方は、始祖鳥の骨格と、白亜期初期の最近大人気の獣脚類デイノニクスの骨格を比べてみると面白い。あれは大きさ以外はそっくりおんなじものだ。きっとデイノニクスも羽毛に覆われていたんだろう)。また、知能がもっとも発達している（と少なくとも自ら思っている）人間も、手足にそれぞれ五本の指趾をもつなど、その身体的特徴は、陸生の脊椎動物としては最も下等な形態を留めており(尾が無いのはカエルも同じ)、非常に下等な食虫哺乳動物(ネズミのようなもの)から突如進化してきたらしいことが推測できる。(以上の部分は[毛利　92]の要約と一部加筆。)

このようなことを考えているうちに、生物の進化に限らず、ものごとの進歩発展には、いたるところにネオテニー的進歩があることに気付く。近年のコンピュータ技術の進歩に関するもっとも典型的なネオテニー的発展進歩の例は、RISC技術とそこから発展したスーパースカラーなどの技術であろう。これは、様々な状況に対応するための複雑な命令セットを扱えるように発達して来たCISCにおけるマイクロプログラムなどの技術を一切取り外して、CPU本来の純粋な姿を追及した結果得られた新技術である(異論もあるだろうが)。

これから言えることは、技術や方法論の研究開発においても、革新的技術はむしろ非常に未発達状態のものから生まれてくるということを示している。大きな進歩はそれまでの適応特化の延長上にはないのである。だから研究開発にたずさわる者は、ときには自分が特殊な分野の特殊な問題解決のために特殊な法論を探求していないだろうかと振り返り、見直すことが大切だということを教えてくれる。多くの人は、先端研究と、末端研究を取り違えている人が多い。「温故而知新」など色々なコトワザにもあるけれど、本当の先端は、根っこにあるものなのだ。

とすれば、境界領域研究のように、異なる研究分野の研究者が格闘して一つの学問領域を確立し、相互理解していく過程では、両者それぞれが自分達の研究の在り方を根っこから見直す必要がある。だからこそ、境界領域は爆発的に進歩発展する可能性をもっているとも言えよう

(ちょっとコジツケかも知れないけれど)。

4 タンパク質の二次構造予測問題

さて、本題に入ろう。筆者はいま、タンパク質の立体構造予測問題に取り組んでいる。で、ここでは筆者の最も生々しい進行中の研究を紹介したいと思う。

タンパク質の立体構造を知ることはタンパク質の機能がどのように実現されているかを考える上で極めて重要だ。しかし、実際のタンパク質の立体構造を実験的に直接決定するのは難しい(連載3月号を参照)。現状で実質300種類程度のタンパク質の立体構造が決定されているが、これは大変少ない。というのも、タンパク質のアミノ酸配列決定は自動化されているため、既に5万本以上の配列がデータベースに登録されている訳だから。

タンパク質はそのアミノ酸配列によって決定されているから、配列が同じなら立体構造も同じ筈(Anfinsenが実験的に証明らしいのだが) — ならば、立体構造は原理的にアミノ酸配列からほぼ一意に決定でき、配列から立体構造の予測は原理的に可能な筈だ。それができれば、アミノ酸配列が決定されている数万のタンパク質の立体構造を一挙に知ることができる。1種類のタンパク質の立体構造決定が1研究室1年として人件費だけでも5千万円くらい費用がかかると思うし、タンパク質は人間に関わるものだけでも10万種類と言われているから、単純に計算して5兆円くらいの一次的な経済効果があると考えられる(これは凄い)。新薬の開発やタンパク質の機能解明などの二次的な効果を経済効果として考えるとそらく莫大なものになるだう。

というわけで、最近では、いろいろな人々がこの研究を始めた。ではどうするか。

タンパク質の立体構造決定の技術が確立する前から、タンパク質にX線を当てると奇妙な回折パターンが二種類現れるので、それぞれにαパターン、βパターンと名付けてきた。後にそれはL. Paulingらの研究で規則的ならせん状のhelixと、規則的な直線状のstrandであることがわかった。その他に特徴的なturnと呼ばれる折れ曲がり構造を含めて、これらをタンパク質の二次構造と呼ぶ(詳しくは本連載3,5月号)。

こんなキレイな規則的な構造が発見されたため、だれもが立体構造の予測には、まずどの部分がhelixで、どの部分がstrandかということをその部分のアミノ酸配列から予測するという、いわゆる二次構造予測が大切だと思うようになった。で、本来の立体構造予測は二次構造予測技術が確立したらやろうではないかという暗黙の了解ができてしまった(たとえば、[Cohen等 82]も含めて[Fasman 89]に詳しい)。

それではという訳で、みんながこの二次構造予測問題にハエのように群がり、確かに、だんだんと予測精度は上がってきた。かつては50%と言われたが、最近では、80%近い精度があると豪語する方法もある(厳密なテストをすると、60%をちょっと越える程度らしい)。いろいろな人がいろいろなフクザツな工夫をするから少しづつ精度が上がっていくのは当り前。でも、誰もが認める決め手となる方法は未だ見つかっていない。

情報科学の分野の研究者にとっても、定式化された二次構造予測問題は非常に扱い易いので、分子生物学入門のための材料としてもてはやされ、学習理論とかニューラルネットとか、音声認識とか、あらゆる分野の研究者が分子生物学のイロハも勉強せずによってたかって研究を始めた。かく言う筆者も群がっている一人ではあるが、やはりこの状態はちょっとおかしい。単なる二次構造予測が本当に意味があるのか、もう一度始めから考えてみる必要がある。

二次構造予測をベースとした立体構造予測は、「局所構造(二次構造)はその部分の局所配列によってのみ決定される」という仮説を大前提としている。だから、現状で精度が上がらないのは、まだ学習に使えるサンプルが少な過ぎるからだという言い訳ができる。ほんとうか。

そもそも大前提となっているこの仮説は物理化学的には決して証明などできるものではなく、むしろ、否定的なものだ。賢明な生物学者はこのことにとっくに気付いている(例えば、

[Branden と Tooze 91] にも書いてある)。とは言え、この仮説が崩れたときにどのような立体構造予測の方法論が考えられるかということについては、生物学者にとって難しい問題だ。だからここで本来情報科学の研究者が方法論を提供することができる筈 — かつての Watson のように、情報科学で知られている理論をひっさげて、生物学に他流試合を挑むよいチャンスなのだ。

しかし、筆者の知る限り、これまで情報科学者の側から根本的に新しい方法論を提案するということががほとんどなされてこなかった。情報科学の分野の研究者がこれほどこの問題に群がっているのに、基本方針としては生物学者が 20 年前に考えた本質的に否定的な仮説を大前提としいて、ひたすら予測精度の向上のために、とてつも無く高級でゼイタクな情報科学的理論方法論を投入し続けることに終始してきたのだ。いかにゼイタクな方法論を用いても、この前提となる仮説を用いている限り、生物学者が考えた単純な統計に基づく方法 (例えば原点とも言える [Chou 74] など) に比べて格別精度の良い方法を開発できる筈が無い。事実そうだった。

生物学の分野に情報科学の分野から進出するならば、それは、生物学者が古来から考えている手法の未発達な部分、ナイーブな部分を手直しして、ゴテゴテと化粧するためではあるまい。根本的に新しい方法論を情報科学的知見を拠り所にして提示することこそ、大きなな貢献となるはずだ。

勿論、コンピュータ科学での理論的研究成果がそのまま他の学術分野にすぐに応用できると考えるのは余りにもナイーブだ。綺麗な理論、方法論であればあるほど、実際の応用を考えるときには、応用にまつわるぐちゃぐちゃした膨大な工夫を加えないと、現実的な問題を解決できないのが普通だ。で、現実的問題と綺麗な理論との間を汚い工夫のかたまりで繋ぐと、結局は全体としては非常にキタナらしい応用になってしまう。ということは大変だけれども、対象となる問題に合わせて根っこから新しい方法論を作らなければならない。勿論、ここで既存の綺麗な理論体系は方法論の探求の上では極めて有益だヒントになることは間違い無い。でも、そのまま使ってもエレガントなものは出てこない。

5 いきなり立体構造を予測する

というわけで、筆者は二次構造予測を行わずにいきなり立体構造を予測する新しい方法を研究している最中だ。結局、これまでの二次構造予測では考慮されていなかったタンパク質の立体構造形成のファクターを考慮した構造予測方法を考えれば良い筈だ。

まず、局所配列とその部分の立体構造との制約関係は二次構造予測で用いられてきたもっとも重要なファクターであり、構造予測の原点でもある。しかし同時に、これまで考慮されてこなかった中規模大規模な部分構造と、その部分の配列との制約関係は大局的相互作用として考慮されなければならない。こういう配列と構造との様々な規模での制約関係を一次制約と呼ぶことにする。さらに、局所構造同士が 3 次元的にどう配置され得るかを空間的制約として考慮する必要がある。これを幾何学的制約と呼ぶことにする。これだけでも、今までの二次構造予測よりずっと色々なファクターを取り込んだことになる。だいぶ、根っこから研究していることになる。

さて、果たしてこれらのファクターを取り入れた立体構造予測の方法論が得られるのか。少なくともこれまで生物学者は諦めてきた。扱いが難しいからだ。

しかし、情報科学、コンピュータ科学は、このようなブヨブヨした問題でも解く方法論を与えてくれる。統計、多変量解析、形式言語理論、学習理論などの方法論、理論をヒントにすれば、解決不可能な問題ではない筈 — こういうときにこそ情報科学の様々なゼイタクな方法論を試すべきなのだ。

6　対象の記述形式が全てを決める

いちばん最初にやらなければならないのは、立体構造をどううまく記述するかだ。おそらく、これが一番重要な研究となる。これができれば、あとは、一直線に構造予測まで話を持っていける筈だ。勿論、与えられる元の座標表現ではどうにもならない。筆者は構造記述方法の研究に2年を費やした。

おそらく、情報処理のみならず、科学一般で言えることだと思うが、対象の記述形式というものが研究において一番重要なものだ。記述形式が優れていれば、それで記述された対象を扱う方法は自明になることが多い。結局、ものごとを認識するということはものごとの特徴を捕らえて何らかの記述形式に変換することであり、一度変換されればそれは如何様にでも処理できる。

膨大な試行錯誤があった。タンパク質立体構造の記述形式は、次の2つの条件を満たしている必要がある。1つめは、局所構造から中規模構造、大規模構造などの構造の階層的性質を旨く表現出来なければならない。2つめは、記述された構造が元の座標表現にある程度の誤差で復元できることである。階層的性質を考えることで、局所的な相互作用から大局的な相互作用までを取り入れた構造予測方法を考えることができる。また、復元性がある記述形式は、立体構造に関する情報をほぼ完全に保持しているから、これを扱うことはもとの立体構造を扱っているのと等価である。したがって、復元性のある記述形式から得られる様々な情報、知識、規則は実際の立体構造のそれらであることになる。この意味では、これまで使われてきた二次構造を用いた構造記述は全く適当でない — 復元性もなければ階層性もない。

階層的に記述するには、色々な大きさの局所構造をそれぞれ規模に応じて別々に分類することから始める必要がある。小さな構造、たとえば、アミノ酸残基が5個からなる構造、9個からなる構造, 以下、17個、33個、65個、129個、をそれぞれ別箇に分類してみる。こうすると、それぞれの規模のレベルで、構造と配列との制約関係を統計的に取り出すことができ、これがモデル化された一次制約になる。次に、それぞれ分類された構造が、空間的にどう重なり合うかを統計的に調べることで、幾何学的制約も取り出すことができる。もうこれでタンパク質の立体構造形成のファクターのモデリングが出来てしまう。

こうなると、構造を予測したいタンパク質の配列が与えられたら、その各規模のレベルで局所配列とその部分の構造との一次制約と、近傍の色々な規模の局所構造の間の幾何学的制約をもっとも良く満たすような構造の組み合わせを探索することで、立体構造予測を行うことができる。それも、直ぐに3次元座標表現が可能な形式で記述されたものを予測することが出来る訳だ。これは統計処理の結果を確率モデルに入れ込むことで、確率的な最適化問題として定式化できる筈だ。

これで研究のシナリオもできた。

7　で、根っこから研究を始めた..

..訳であるが、シナリオができても新しい方法論を確立するのは結構大変だ。なにせ誰もやっていないことだから、既存の研究をベースにして行くことはできない。できるだけ簡単な方法を考えて行く。

最初にやることは、いろいろな大きさの局所構造の分類だ [Onizuka 等 93]。ちょっと細かい話にはなるが、配列上連続するアミノ酸残基 N 個からなる局所構造は $3N-6$ 個の数値パラメータを用いて表現できる。(話を簡単にするために、1つのアミノ酸残基の位置はその C^{α} 原子の位置で代表させている)。3次元空間だからそれぞれのアミノ酸残基の位置の表現に3個のパラメータが必要で、N 個のアミノ酸残基の位置の表現には $3N$ 個のパラメータが必要だ。これから回転並進の自由度6を除くと $3N-6$ 個になる。でも、筆者たちはいろいろな大きさ

の局所構造を同じような基準で分類したいから、局所構造のアミノ酸残基の数が変わってもパラメータの数が変わらないように工夫しないと厄介なことになる。そうしないと、大きな構造の分類に何百ものパラメータを扱わなければならなくなるからだ。

それでどうするかというと、局所構造の大きさに応じて局所構造を表現するときの解像度を変化させることを考える訳だ。大きな構造は細かく見てもしょうがないから適当に表現し、小さな構造はきっちり表現することにすれば、どの規模の構造も同じ数のパラメータで表現できることになる。こういうことをしないから、今までは、2 残基とか 3 残基といった小さな部分構造の分類しか行われてこなかった ([Miller 等 93] などのように Ramachandran Plot を利用するのが一般的)。ここでは線形代数学のもっとも基本である、線形展開 (あるいは有限次元の

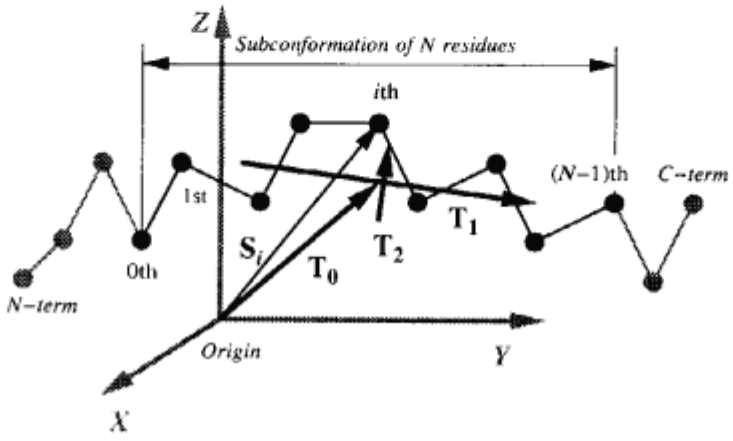


図 1: 局所構造と、線形特徴量としてのベクトル $\mathbf{T}_k$

線形基底変換) が使える。N 個のアミノ酸残基の座標データの配列 (部分構造の中の i 番目の残基の位置を位置ベクトル $\mathbf{S}_i$ で表そう) を線形展開して、部分構造の線形的な特徴量 (これの k 次のものをベクトル $\mathbf{T}_k$ で表そう) を取り出すということをやる。線形展開の基底には、規格

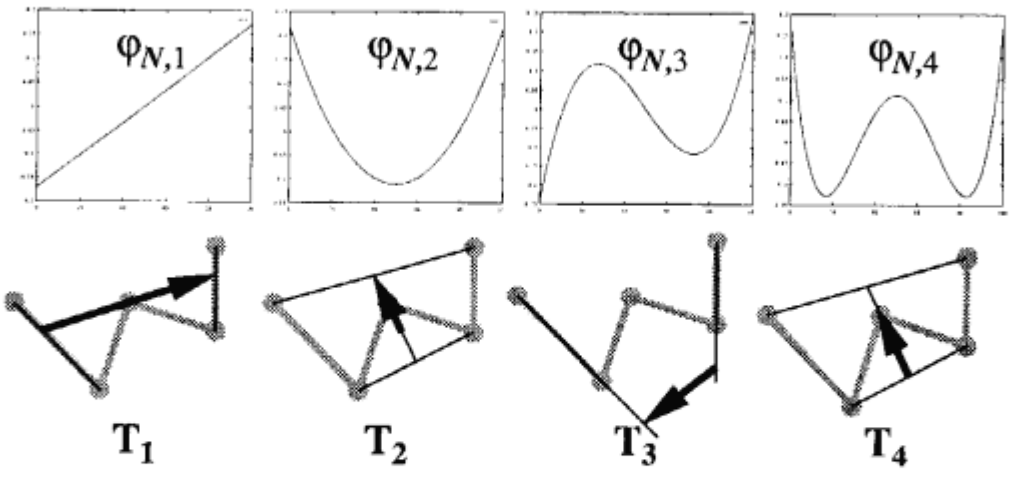


図 2: 規格直交基底 $\varphi_{N,ki}$ と対応する線形特徴量 $\mathbf{T}_k$

直交系 (成分の数 N の直交基底の k 次のものの i 番目の成分を $\varphi_{N,ki}$ と表そう) をつかうと、

線形展開は、下のような数式で表せる。

$$\mathbf{T}_k = \sum_{i=0}^{N-1} \varphi_{N,ki} \mathbf{S}_i. \tag{1}$$

で、逆展開は、下のようになる。

$$\mathbf{S}_i = \sum_{k=0}^{N-1} \varphi_{N,ki} \mathbf{T}_k. \tag{2}$$

逆展開が可能だから、復元性があるのは自明で、本質的に可逆な展開を考えることができる。ここで、$\mathbf{T}_0$ というのは、局所構造の平均位置を表しているので、これを無視すると、並進自由度を無視することができるし、$\mathbf{T}_1$、$\mathbf{T}_2$ をつかって、構造の向いている方向を規格化すると、回転自由度も無視することができる。一般に、線形展開では次数の小さい方から順番に、重要

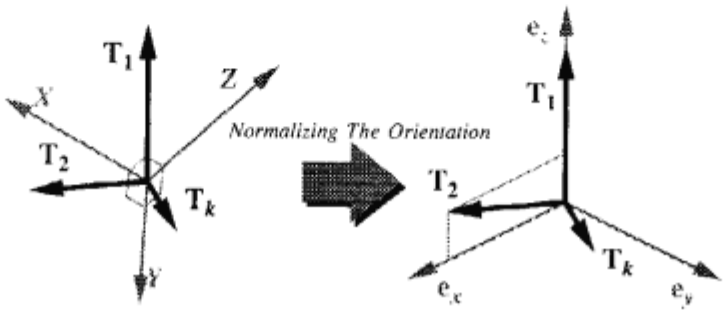


図 3: $\mathbf{T}_k$ の方向を規格化する

な展開係数が取り出せるから、適当な次数で展開を打ち切ると、必要な数だけのパラメータが得られる。勿論、こうすると完全な復元のための高次の展開係数の情報が失われるから復元は正確には行われない。しかし、我々の場合、階層的記述を行うので、いろいろな規模で展開しておけば、小規模な構造からの積み上げである程度の精度での復元は可能だ。というわけで、$\mathbf{T}_1$ から $\mathbf{T}_4$ までのベクトル線形特徴量をもらいて、X, Y, Z で 12 個のパラメータが得らる ($\mathbf{T}_1$ の長さは部分構造の大体の長さ、$\mathbf{T}_2$ は、部分構造の曲がり方、$\mathbf{T}_3$ はねじれ方、$\mathbf{T}_4$ はくねり方を表すと考えられる)。方向の規格化をすることで 3 つのパラメータが常に 0 に成るようにできるから、全部で 9 個のパラメータが得られることになる。どのような規模の局所構造もその大体の形状をこの 9 個のパラメータで上手に表現できることが分かった。これは、最近よく信号解析などで使われているトレンディな方法 "Wavelet 変換" ([Combes 等 89] に詳しい) をヒントに考えたものだが、最終的には全く異なるものになった。

こうしておいてから、それぞれの大きさの局所構造をその 9 個のパラメータを用いてクラスター分析して分類し、それぞれのクラスターに名前を付ければ、局所構造の分類はできたことになる。とは簡単に書いたが、これだけれも大変な作業だった。最初は、物理化学的に意味のあるクラスター分析 (つまり、化学結合の状況を考慮したような) をしないといけないと思っていたのが敗因で、どうにもうまくいかない。結局、形状の量子化なんだからと割り切ったところで、普通のベクトル量子化法を使ってクラスター分類ができた。ここで得た教訓は、分からないことはそのままにして、先に進めるということだ。先に進むと、先では何が必要で、何が必要でないかが分かってくるから、やらなければならないことと、そうでないことが切り分けられる。口絵 1 で、実際の 33 残基からなる部分構造が、分類されて復元されると、のっぺりとした構造になってしまうのが分かると思う。これは、階層的な復元はしていないから、こうなるのである。

次にやることは、実際のタンパク質の立体構造を分類された局所構造の組み合わせとして、分類記号を用いて記号記述することだ。およそ 400 個の厳選された良質 (?) のタンパク質の立体構造データ (データの抜け落ちが無いとか、精度が高いとか) を記号記述する。そうして、ここから前述の幾何学的制約の基となる事例をとりだすことになる。カラーの口絵 2 では、その階層的な記述が実際どのようになっているかを示している。

綺麗な形で有名な TIM(Triosephosphate Isomerase) というタンパク質の立体構造 (口絵 3 参照) の記述例を下に示す。各規模のレベルで、16 種類のクラスに分けたから、アルファベットの A から P を使って表現している。見て分かるように、5 個のアミノ酸残基からなる構造のレベルでは、二次構造が helix になっているところ、つまり、二次構造が h となっているところには、A という記号が来ていることがわかり、良く対応している。それから、s で示される strand になっているところには、F とか P という記号が来ていることが多い。口絵 4 で、5 個のアミノ酸残基からなる分類された局所構造の A と P,F を見れば、それが、helix、strand であることは直ぐに分かると思う。

0 から 49	0	1	2	3	4
二次構造		sssss	hhhhhhhhhhh		sssss hh
5 残基レベル	HLFPLPPPPMGCCHBAAAAAAAAAAAAGJFPDIECFPPPFCHBAGMAA				
9 残基レベル	DDDDDDIPPPNKEBBAAAAAAAAFFFLPPNNLHDDDGTKFABAAA				
17 残基レベル	KKGGMMMMOOOOOAAAAAADBBPPPLDDDHIIGGJMNNCOO				
33 残基レベル	FFFOOBNNDDDDDDFFFOONNNNNDDDDHEHHHK				
65 残基レベル	EEEEEJJJJJJJJJHH				
129 残基レベル					

50 から 99	5	6	7	8	9
二次構造	hhhhh	sssss		hhhhhh	ssss hhh
5 残基レベル	AAAGKFIBGJOFPODKOHDJEIGJDGJHOBAAAAANEPNKCLDKCBAAAA				
9 種基レベル	AAFFLNKMGLHDDGOPPPPNNNNMMNLLKEBBACCJGLPHOOIIEEBBAC				
17 残基レベル	OBBBPPLLEFFHHIKKGGGJJJPPLLEOOOABBPPLEFFDHIIGGGBBPP				
33 残基レベル	KKKKLLLLLLBBBBDDDIIIIIHHHNNLLMMGGGOOOONNNNDDDDDHHK				
65 残基レベル	HFFFFFFFFFFFFFFFJJJLLLLLLLLAAAAFFFFFFFJJJGGGLLLJJ				
129 残基レベル			PPJAAAAAAAAAAAAAAAAAEEEEEEEEPPPPPPPP		

100 から 140	0	1	2	3	4
二次構造	hh	hhhhhhhhhhhhh	ssssss	hhhhh	hhhhhhhhhhh
5 残基レベル	ANEPCBAAAAAAAAAAAAANECLPLPFFCHOBAAAANEMMAAAAAAAAAAA				
9 種基レベル	JGGBKEBBAAAAAAAACCJGLPHDDDDDIKEBBCCJGBEFAAAAAAAAAA				
17 残基レベル	PCCCCDCBAAAAABBBPPLEEFFHHKKKGGGJMNNLCCOOAAAAAAAAAA				
33 残基レベル	HFFOOBBBDDDDDMMMOOLLLLBJJJPPPIIIIIHFFOOBBBJJJPPPCC				
65 残基レベル	JJJJJJLLLHHHFEEEEEEFJGGGGGGGGKKKKKKLLLLLLLLLKKKKEE				
129 残基レベル	PAAAAAAAIJJJJFFFFFFFFJAAAAAAAAAALLLLPPPPPPPPPPPPPP				

150 から 199	5	6	7	8	9
二次構造	hh	hhhhhhssss		hhhhhhhhhhhhhhhhhh	
5 残基レベル	AAGJOBAAAAGJOPPFCOBAGMGIEJCLHBAAAAAAAAAAAAAAAAAAANHB				
9 残基レベル	ACFBNEBBAFFHHDDDIKEBBCCJGLHIKEBBAAAAAAAAAAAAAACJGGE				
17 残基レベル	BBBCCPPLDDDHHHKKGGGGMNLLLOOOOMMAAAAAAAAAABBBBPPPCC				
33 残基レベル	CCHFFOOLLLBJJJJPPPPIIIHHFFOOBBBBJJJEPPPPCCCCHFFOOL				
65 残基レベル	EOOOKKDDDDDDDKKKKKKKKKKNNKDDDLLLLKKKKKKKKKKKNNNE				
129 残基レベル	FFFFAAAAAAAAAAAAAAFFFFFFFFFFFFFPPPAAA				

200 から 240	0	1	2	3	4
二次構造	hhhhhh	sss	hhhhhh	ssss	hhhh hhhhhh
5 残基レベル	AAAAGIFPPFFIKPCBAGAAAAAGIBGJOJPLPDAAAGNBAAAAAAAG				
9 残基レベル	BBAFFLHDDIIPPNKEBBAAAFACCJGLHOOIIEBBCCEBBAAAAA				
17 残基レベル	CODDHHHHIIGGMMNNCCOOOBPPPLEEDFDIGGMMNNPCOO				
33 残基レベル	LLBBJJJDPPIIIIIIHHKKKKKLLLLNNNNDDD				
65 残基レベル	EEEEAAAJJJJJJJJAFF				
129 残基レベル					

こうして、タンパク質の構造をもとの立体構造に復元可能な階層的に記号記述する方法が完成した。これはある程度の誤差はあるが、元に復元できるわけで、3 次元構造に関する情報をちゃんと保持した記述形式になっている。

8　統計に基づく確率的制約解消問題

タンパク質の立体構造が、簡単な階層的な記号記述で表現できれば、あとは、ここから局所的な構造同士のいろいろな制約関係を統計情報として取り出せば良い [Onizuka, Asai と Ishikawa 93]。例えば、5 残基からなる構造で B と分類されているものが、9 残基からなる A と分類されている構造と先頭の残基が 2 ずれて重なっていたら、「B5 は A9 と 2 だけずれて重なる」という表現ができる。で、それぞれの重なり具合がどれくらいの頻度で現れるかを実際のタンパク質で調べて統計処理する訳だ。これが幾何学的制約のモデルになる。したがって、この制約と、

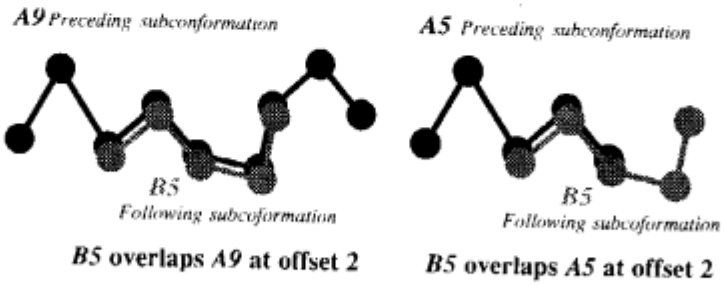


図 4: 局所構造の重なり具合

一次制約を両者程良く確率的に満たすような局所構造の組み合わせを探索すれば、立体構造予測は出来てしまう。勿論、この状態では予測された立体構造は局所構造の分類記号の組み合わ

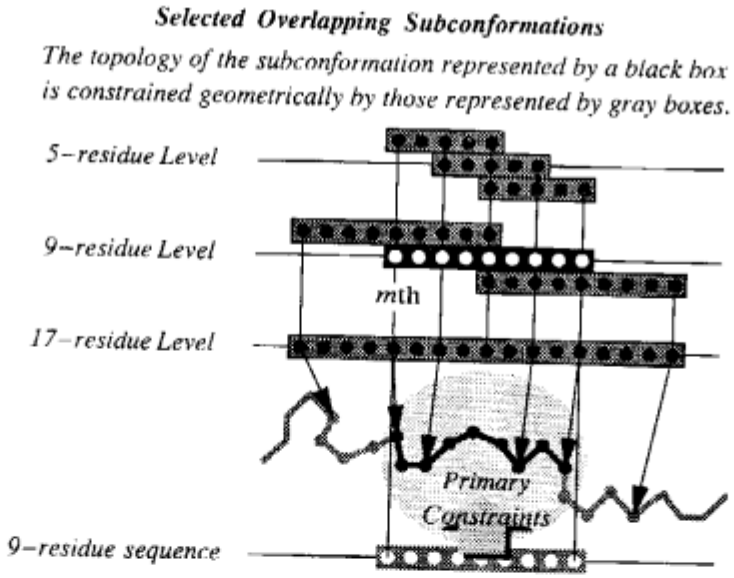


図 5: 確率的制約推論の制約関係を示したもの

せによって記号記述されているから、これをもとの座標表現に戻す必要があるけれど、これはさっき書いたように数学的に可能だ。なんかこうして方針がたつと、できそうな気がしてきた。

9　おわりに

紙面の都合もあってイキナリ終わってしまうけれど、統計をベースにした確率的制約推論などの AI 的な方法でタンパク質立体構造は予測できると思っている。二次構造予測が限界に達していることも多くの研究者にとって明らかになってきた。革命的な方法論が提案されるのは

これからかなのだろう。それだけ情報科学の関係者も生物学の知識を蓄え始めたということでもある。

かつては夢だったタンパク質立体構造予測問題が夢ではなく計画に変わった。計画は実行すれば実現する。「夢 計画 実行」は実現のための大事なステップだ。宝くじは買っても当たる可能性は凄く低いけれど、買わなければ当たらない。同じように、大きな夢は実現が不可能に思えるかも知れないが、夢を持ち続ければ実現するかもしれない。それに、大きな夢は研究者の視野を大きくして、本当に必要な研究が何であるかを分からせてくれる。そして、夢はやがて計画となり、計画は困難は伴うかも知れないが、実行すれば実現できる。結果が悪ければ、もう一度やり直せば良い。研究開発に関わらず何事もそうではないか。

参考文献

[Watson 68] Watson, J.D., *The Double Helix*, 1968, Weidenfeld and Nicolson Ltd., London.

[Watson S43] ジェームズ D. ワトソン著, 江上不二夫, 中村桂子 訳 "二重らせん" 昭和 43 年, タイムライフインターナショナル.

[毛利 92] 毛利秀雄, "ネオテニー考", in *UP 239 号 (9 月号)*, 1992, pp. 8-12, 東京大学出版会

[Cohen 等 82] Cohen, F.E., M.J.E. Sternberg, W.R. Taylor, *"Analysis and prediction of the packing of α-helices against a β sheet in the tertiary structure of globular proteins"* in *J. Mol. Biol. 156*, 1982, pp. 821-862.

[Fasman 89] Fasman, G.D.(editor), *Prediction of Protein Structure and the Principles of Protein Conformation*, 1989, New York: Plenum Publishing Corporation.

[Branden と Tooze 91] Branden, C.,and J. Tooze, *Introduction to Protein Structure*, 1991, New York: Garland Publishing, Inc.

[Chou 74] Chou, P. Y. and Fasman, G.D. *"Prediction of protein conformation"* in *Biochemistry 13*, 1974, pp. 222-244.

[Onizuka 等 93] Onizuka, K., K. Asai, M. Ishikawa and S.T.C. Wong, *"A Multi-Level Description Scheme of Protein Conformation"*, submitted to *The First International Conference on Intelligent Systems for Molecular Biology*, 1993.

[Miller 等 93] Miller, R.T., R.J. Douthart and A.K.Dunker, *"An Alphabet of Amino Acid Conformations in Protein"*, in *Proc. of the 26th HICSS*, 1993, pp. 689-698.

[Combes 等 89] Combes, J.M. et al, *Wavelets, Time-Frequency Method and Phase Space*, 1987, Springer-Verlag.

[Onizuka, Asai と Ishikawa 93] Onizuka, K., K. Asai, and M. Ishikawa, *"A Scheme for Protein Tertiary Structure Prediction Based on Stochastic Reasoning"*, submitted to *WORKSHOP "ARTIFICIAL INTELLIGENCE and the GENOME" at IJCAI*, 1993.

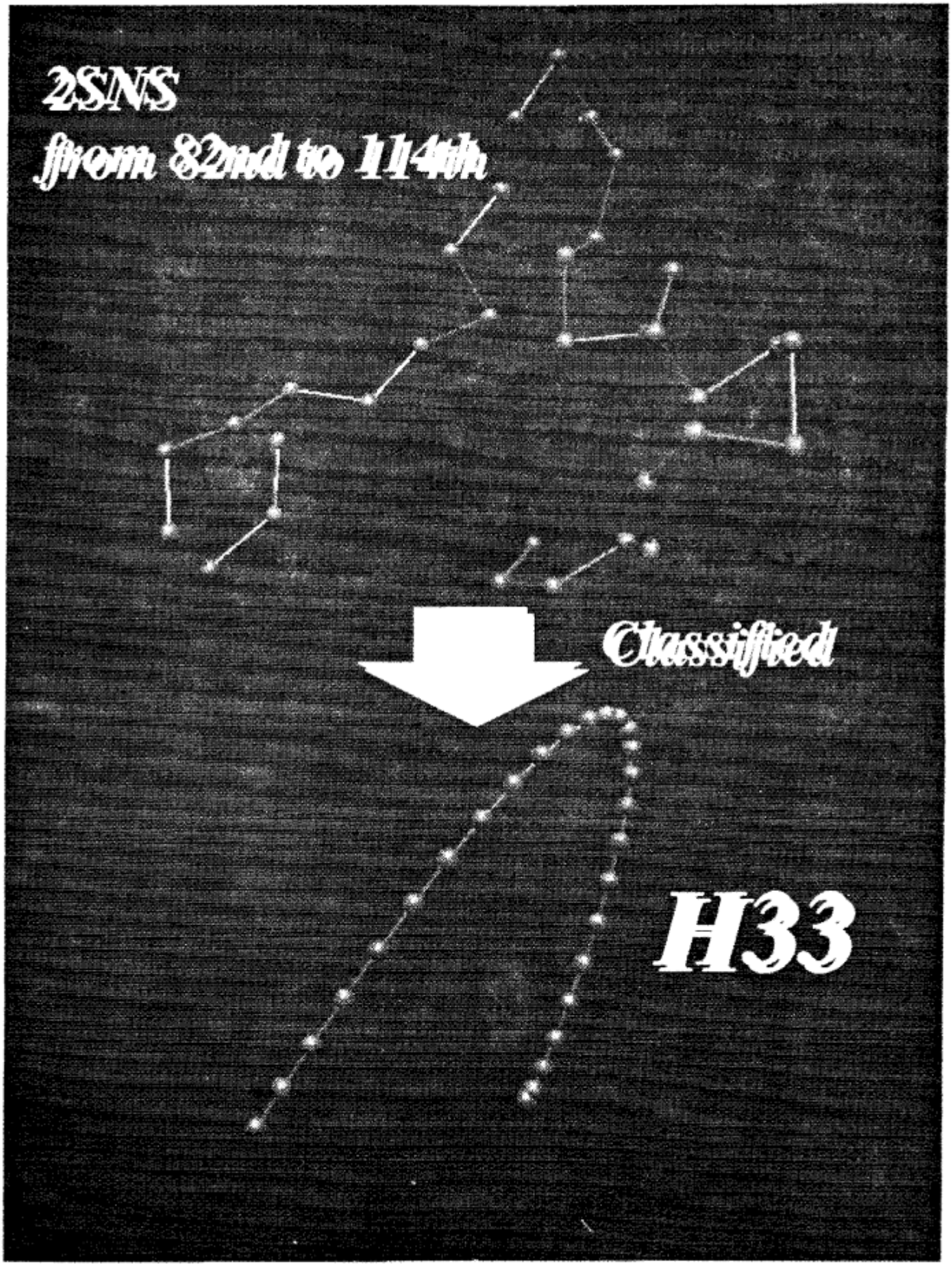
10 口絵の説明

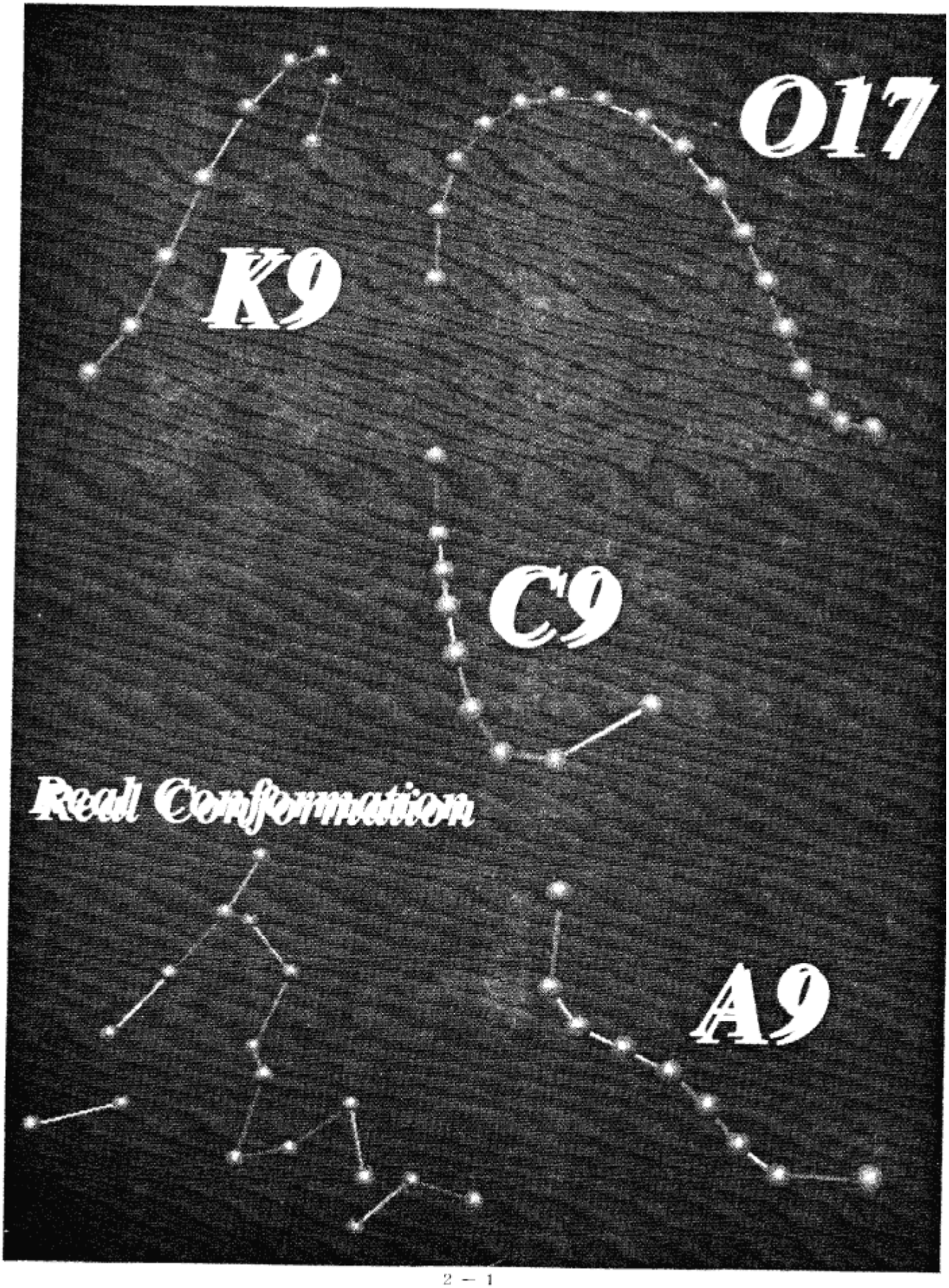
口絵 1: PDB にエントリー名 2SNS で登録されているタンパク質の 82 番目の残基から 114 番目の残基までの 33 残基からなる部分構造は、今回紹介した部分構造分類では *H33* と分類される。この *H33* を逆展開で復元すると、下のようにのっぺりとした構造になる。大きな構造の分類は、この程度のいい加減さで十分だ。

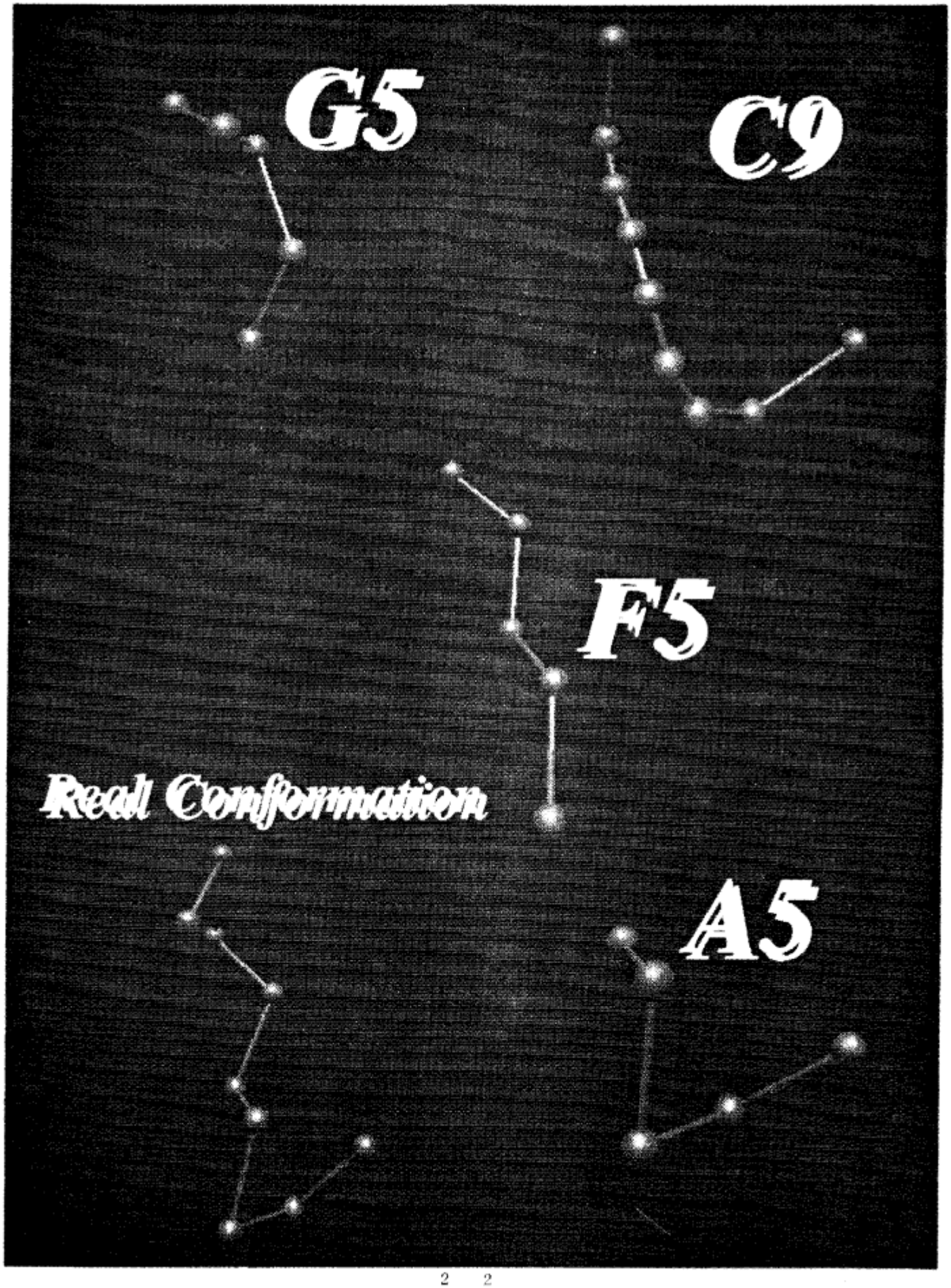
口絵 2: *Real Conformation* というのは、実際の構造。2-1 ではこの 17 残基の構造が *O17* と分類されたことを示し、さらに、その最初の 9 残基が *K9*, 中央部分が *C9*、最後の部分が *A9* と分類されている。同じく、2-2 では、*C9* と分類された構造の最初の 5 残基が *G5*、中央の 5 残基が *F5*、最後の 5 残基が *A5* と分類されていることが分かる。*G5* の 2 番目の残基の位置が *Real Conformation* と比べると、ちょっと変な位置にあるが、これは、量子化したときの誤差である。

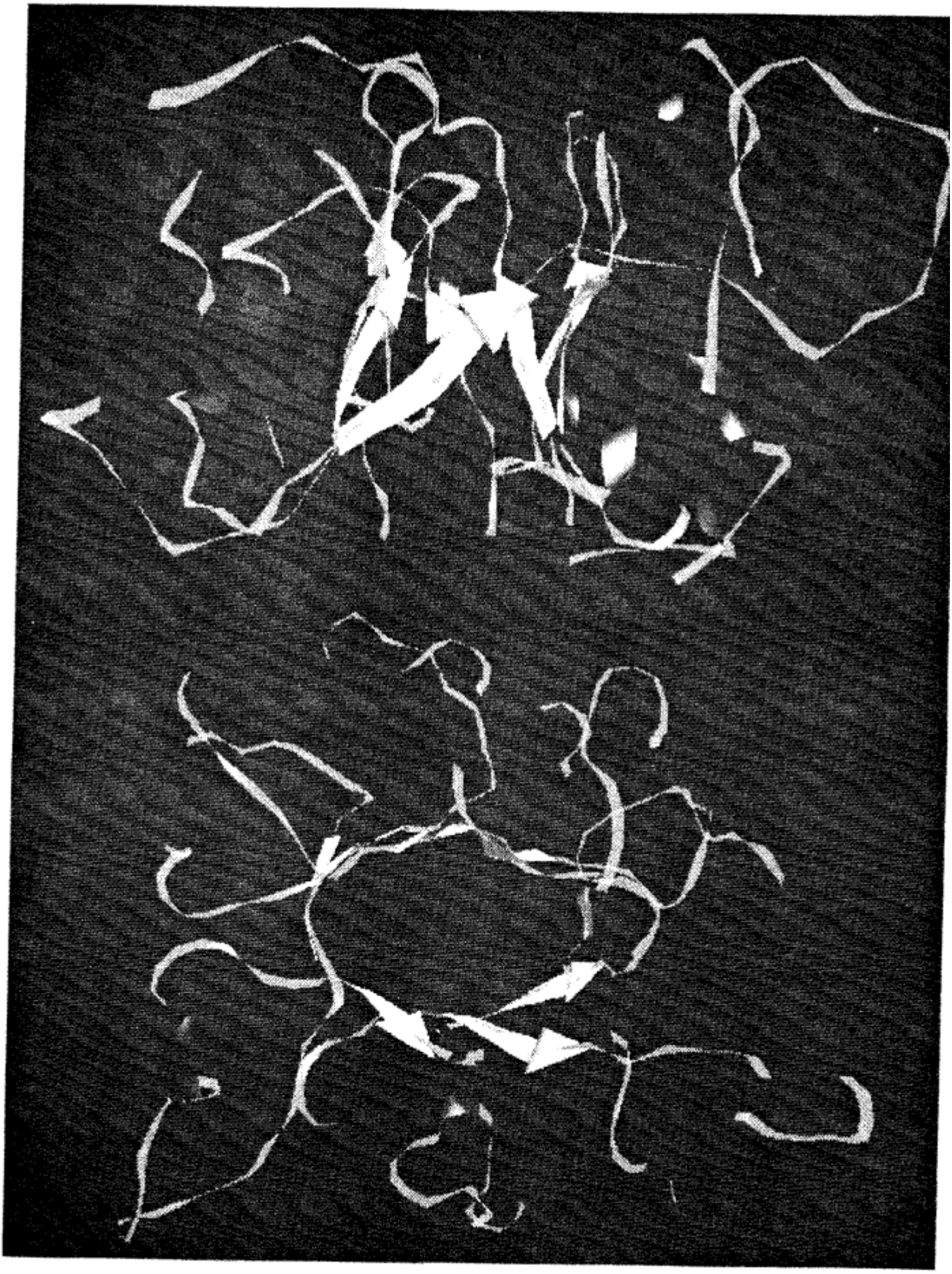
口絵 3: TIM(これは PDB エントリー 6TIM の B 鎖) の立体構造を、リボンと二次構造の組み合わせで表現している。上が横から見たところ、下が、上から見たところ。中央部分では、8 個の strand(黄色い矢印型のもの) が円柱のように並び、その周囲を helix(緑の螺旋) が取り巻いている。これを TIM バレルと呼ぶ。人工的にもっと美しい TIM バレルを作る研究もある。

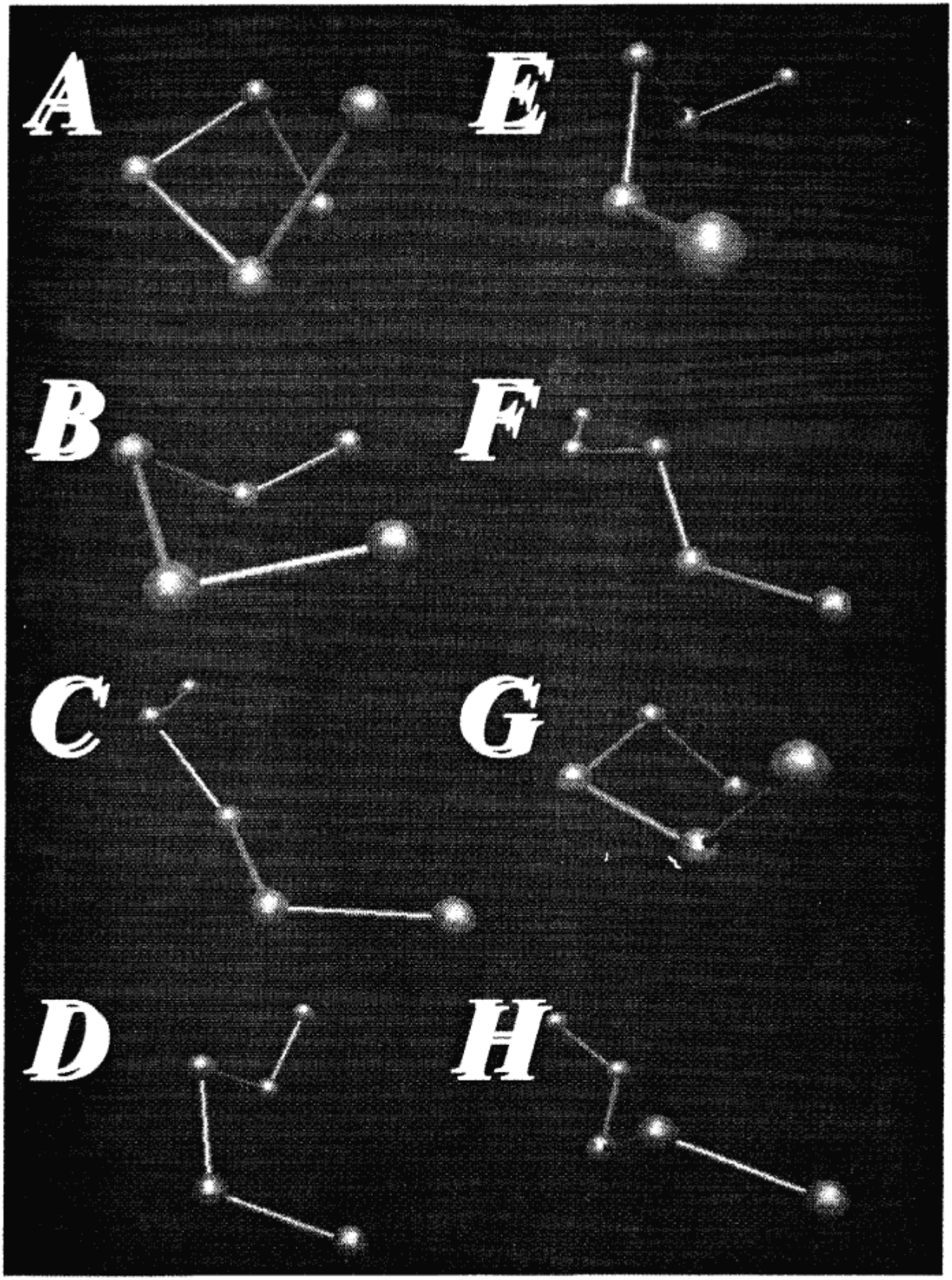
口絵 4: 今回の部分構造分類の例。5 残基からなる構造を分類したもの。手前に来ているのが N 末端側の残基。*A* が helix、*F*, *P* が strand である。そのほか helix の始まりや終わりの部分なども分類されている (*J*, *K*, *M*, *N* など)。

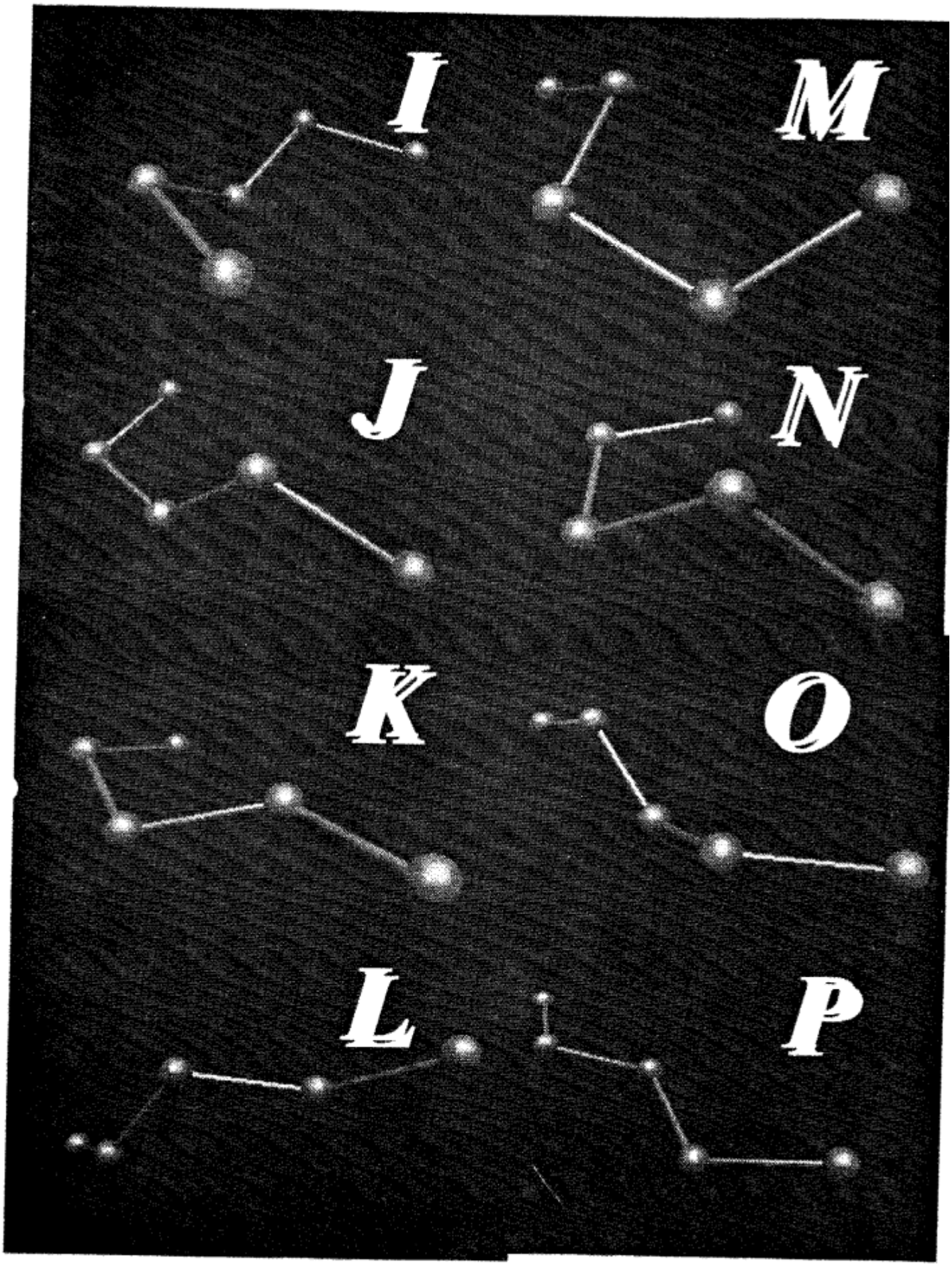

2SNS
from 82nd to 114th
Classified
H33


017
K9
C9
Real Conformation
A9


G5
C9
F5
Real Conformation
A5




A
E
B
F
C
G
D
H


I
M
J
N
K
O
L
P