統計数理(2014)
第 62 巻 第 2 号 257–272

特集「生体高分子の揺らぎとダイナミクス
—シミュレーションと実験の統計解析—」
[研究詳解]

1 分子計測実験から分子状態を識別する統計的データ解析

岡本 憲二†

(受付 2013 年 12 月 31 日；改訂 2014 年 5 月 15 日；採択 7 月 2 日)

要　　旨

近年の生物学においては，蛋白質をはじめとした分子のはたらきを理解することの重要性が高まっている．そのためには，個別分子の振る舞いを直接観察できる単一分子計測技術が有効であり，広く用いられるようになってきている．1 個の分子から得られる信号は，きわめて微弱なため大きな揺らぎを含むが，一方で，有限個の分子状態の間での遷移ダイナミクスとして解釈できるという特徴がある．分子レベルでは，分子の振る舞いも信号生成も確率現象であるため，一見乱雑な信号から分子状態の情報を取り出すために統計的な解析を有効に用いることができる．本稿では，FRET 現象を利用した分子構造変化ダイナミクスの 1 分子計測実験データを対象とした統計的データ解析手法について紹介する．時系列データから状態遷移軌跡を復元する隠れマルコフモデルの概要と，最尤法から変分ベイズ法への転換，さらに時間依存するタイムスタンプ信号への拡張について解説した後，シミュレーションによる評価と実験データへの適用について紹介する．最後に，別の時系列データ解析法である変化点検出法についても触れる．

キーワード：生体分子，1 分子計測，FRET，隠れマルコフモデル，変分ベイズ，状態識別.

1. はじめに：分子から見る生物学

生命はあらゆる階層，さまざまな局面で神秘的と言えるほど複雑で巧妙な振る舞いを見せてくれる．しかし，この世界に存在する限り，物理や化学の法則から逃れられる訳ではない．生命を構成する最小単位は分子であり，あらゆる生命現象は分子の作用によって説明できるはずである．特に蛋白質は，ヒトの場合には数万種類に及ぶほど多種多様であり，それぞれが機能を担っている．どの分子が生命現象のどの場面ではたらくかは，分子生物学が明らかにしてきた．たとえば，ある分子が癌の原因になる，あるいは肥満に関係している，といった説明が一般向けのテレビ番組などで紹介されることも珍しくなくなった．

では分子は，具体的には，どのようなメカニズムで機能しているのだろうか？その問いに答えるためには，分子レベルの挙動を物理の現象として理解する生物物理学のアプローチが必要となる．しかし，細胞を観察する，あるいは試験管で液を混ぜるといった実験では，アボガドロ数個程度の分子の平均としての挙動を見ることになり，多くの重要な情報が埋もれてしまう．分子反応の詳細を知るためには個別の分子を観察することが必要であり，実際，単一の分子か

†独立行政法人理化学研究所 佐甲細胞情報研究室：〒 351–0198 埼玉県和光市広沢 2–1

ら信号を得る計測技術が実現されている．1 分子計測技術としては，たとえば，生体膜を通過する電荷を数えるパッチクランプ実験や，探針の先端に加わる微小な力を検出する原子間力顕微鏡などがあり，光を用いた計測法としても，高輝度の散乱体や蛍光体を標的分子にラベルする手法もある．本稿では，単一の蛍光分子から放出されるフォトンを計測する，蛍光 1 分子計測について紹介する．

蛍光 1 分子計測実験で得られる信号にはいくつか特徴がある．分子レベルでは，蛍光発光も含めて，すべての現象は確率的に現れる．1 つの分子からの信号は微弱なため，確率的な揺らぎの影響が無視できなくなる．そして，1 分子信号は連続変化ではなくステップ的に変化する場合が多く，確率的な状態変化が背後に潜んでいると考えられる．そのような信号から有益な情報を取り出すためには，統計的な方法論を用いた解析手法が有効に使える場面も多い．

本稿では，1 分子実験，特に FRET と呼ばれる現象を利用した実験のデータを有限個の状態に還元するデータ解析法について，筆者の最近の取り組みも含めて紹介したい．

2. 蛍光 1 分子 FRET 計測実験の原理

2.1 蛍光イメージングをベースとした 1 分子実験

生物学の実験では以前から，蛍光イメージングがよく用いられてきた．特定の分子を蛍光色素でラベルすることで，その分布や動きを調べることができる．蛍光色素分子は，特定の波長(色)を持った光を吸収し，少し波長の異なる(波長が長い)光を放出する．実験では，レーザー等の単色の光(励起光)を入射し，波長で光を分離する特殊なフィルタを通すことで蛍光だけを検出器に導き，蛍光信号を得る．1 分子計測の場合も基本的な原理は同じだが，分子単位で見た時には，蛍光強度は連続値ではなく，フォトン単位の吸収・放出現象として考える必要がある．

蛍光 1 分子実験では，フォトン 1 個を判別できる程の，きわめて高感度な検出器を用いる必要があり，主にカメラとシングル・フォトン・カウンティング(SPC)検出器が用いられる．1 分子計測で主に用いられるカメラには EM-CCD タイプと sCMOS タイプがある．1 分子蛍光イメージの例を図 1(A)に示す．実際の蛍光分子のサイズは数 nm 程度であり，光学顕微鏡の分解能(〜数百 nm)よりはるかに小さいため，各分子はぼんやりと広がった輝点として画像化される．SPC 検出器としては光電子増倍管(PMT)とアバランシェ・フォトダイオード(APD)がある．どちらも点検出器のため，顕微鏡の焦点を 1 つの分子に合わせて，その分子からの蛍光のみを検出する．SPC 検出器はフォトンが入射する度に電気パルスを生成し，それぞれが 1 個のフォトンに対応する電気パルスの列を出力する(図 1(B))．

2.2 蛍光フォトン信号

ここで，後述の解析のベースとなる，フォトン信号の統計的な性質について説明しておこう．

一般的に蛍光の時系列信号といえば図 1(E)のような信号がイメージされるだろう．しかし，この信号がつくられる過程を考えてみれば(図 1(D))，多数のフォトンの存在があり(図 1(C))，それらを計数していることになる．マクロに見て蛍光強度が一定であると見なせる状況では，蛍光発光は，ほぼ一定確率で間欠的に生じるポアソン過程と見なすことができる．すなわち，フォトンとフォトンの間の時間間隔 Δt の確率分布は，蛍光強度(単位時間あたりの平均フォトン数)I を用いて指数関数 $p(\Delta t) = Ie^{-I\Delta t}$ で表せる．あるいは，ある時間幅のビンに区切ってその間のフォトンを数える場合には，フォトン数 n の確率分布はポアソン分布 $p(n) = \mu^n e^{-\mu}/n!$ (ただし μ はビンあたりの平均フォトン数)にしたがう．SPC 装置を用いる場合，この n の時系列 $\{n_j\}$ を得る SPC 計測が一般的におこなわれる．あるいは，$\{\Delta t_i\}$ または $\{t_i\}$ をすべて記録することでフォトン 1 個 1 個の検出時刻を記録するタイムスタンプ(TS)計測を実現すること

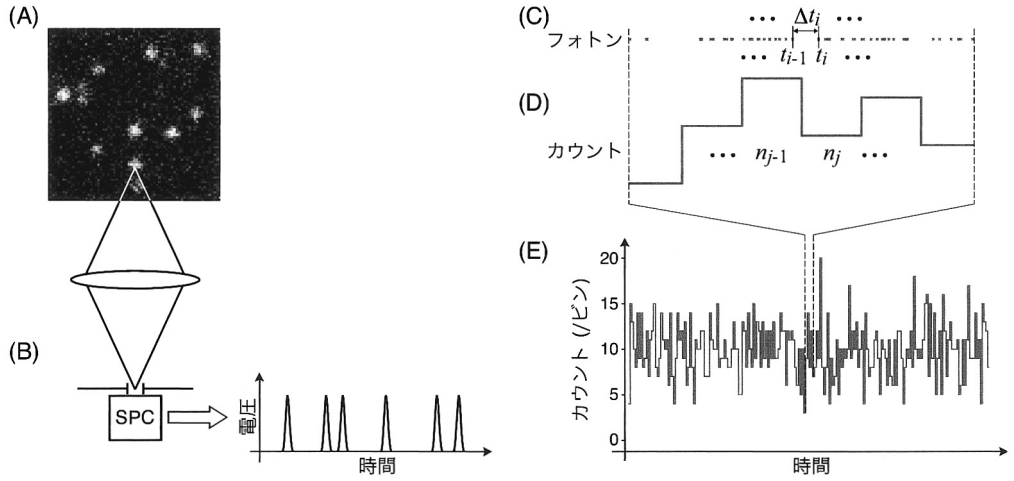

図 1.（A）蛍光色素（AlexaFluor488）でラベルした EGFR C-tail 分子の EM-CCD イメージ．（B）Single photon counting（SPC）検出器は，1 つの分子からの光を検出し，それぞれが 1 フォトンに対応する電気パルスの列を出力する．（C）–（E）フォトン信号の概念．（C）フォトン（×）が間欠的に検出されると，（D）一定時間毎に区切って計数することで（E）いわゆる蛍光時系列信号が得られる．

もできる．TS 計測の場合，フォトン信号が本来持つ統計情報を完全に保存できることになる．

フォトン信号のノイズについても触れておこう．SPC レベルの微弱な信号でフォトン数が限られている場合には，信号の揺らぎが無視できなくなる．この揺らぎはショット・ノイズと呼ばれ，ノイズとは呼ばれるものの外因性ではなく，フォトン放出の確率過程に起因するため，本質的に取り除くことはできない．図 1（E）は平均 10 カウントのフォトン信号をシミュレートしたものであり，外因性のノイズは含まれていないが，ポアソン分布の性質により標準偏差 $\sqrt{10}$ 程度の揺らぎが生じている．SPC 検出器における典型的な外因性ノイズとしては，暗電流がある．本来，フォトン吸収がトリガーとなって生成される電気パルスが，熱によって確率的に生成されてしまう現象を指す．これは冷却によってある程度抑えることはできるが，市販品では上位機種でも数十カウント/秒程度の偽信号が生じる．もう 1 つのノイズ源として，散乱光・迷光など背景光による信号がある．これは光学系を注意深く設計することである程度抑えることはできるが，完全に 0 にするのは難しい．暗電流や背景光に起因する偽カウントは，どちらもポアソン過程で生じる．平均カウントレート x と y のポアソン過程の信号が混じった場合，得られる信号はレート $(x+y)$ のポアソン過程になる．したがって，SPC 装置で得られるパルスのうち，フォトン信号とノイズを区別することは原理的に不可能である．信号はノイズのバイアスを含んでいることを理解し，後の解析・解釈で注意を払う必要がある．

カメラの場合には，フォトン信号はムービーの各フレームで積算されるためポアソン分布となり，SPC 検出器と同様のノイズが生じる．それに加えて，電気的に増幅される段階での揺らぎが加わるため，バックグラウンド信号が常にバイアスとして存在し，その上にフォトン信号が足し合わされた信号となる．そのため，元のポアソン過程の性質はすべては保存されず，ガウス分布として扱うのが容易であり，一般的におこなわれる．

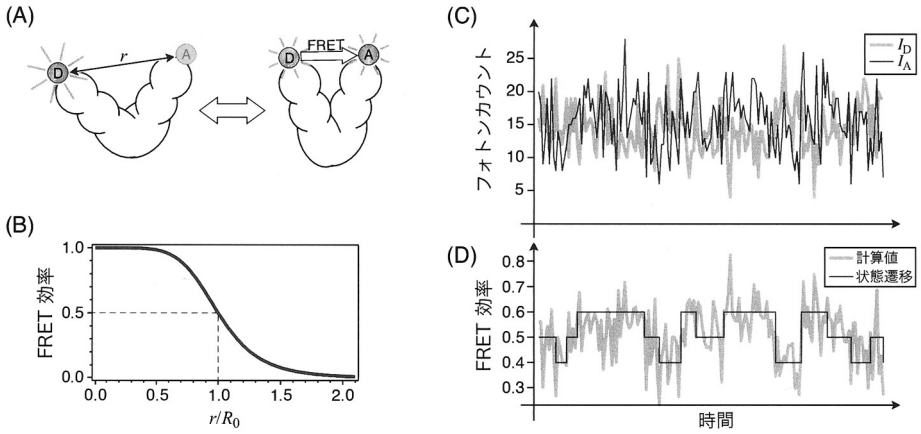

図 2.（A）FRET の概念．ドナー（D）色素とアクセプタ（A）色素がきわめて近接している時に，エネルギー移動が起き，アクセプタが発光する．（B）FRET 効率の色素間距離 r に対する依存性．（C）FRET 計測で得られるドナー（グレー）およびアクセプタ（黒），2 系統の蛍光時系列信号の例（シミュレーション）．（D）（C）から計算により得られた FRET 時系列信号（グレー）．ステップ的に変化して見える信号の背後には，状態遷移ダイナミクス（黒）が隠れている．

2.3 FRET を利用した分子構造計測

生体分子のサイズはたかだか数-数十 nm 程度であり，光学顕微鏡の空間分解能よりもはるかに小さい．そのため，分子の構造変化を直接イメージングすることはほぼ不可能である．そこで，分子の構造情報を蛍光信号として得る手法として，蛍光共鳴エネルギー移動（fluorescence/Förster resonance energy transfer; FRET）と呼ばれる現象が，1 分子計測も含めて，利用されている．以下で，FRET の原理を簡単に説明する．

2 種類の蛍光色素分子が非常に近接している時，一方の色素分子（ドナー）を励起したエネルギーがフォトンの放出を介さずに別の色素分子（アクセプタ）に移動し，アクセプタが発光する場合がある（図 2（A））．このエネルギー移動現象が FRET と呼ばれ（Förster, 1946; Lakowicz, 2006），FRET 効率，すなわち FRET が起きる確率 $E_{\mathrm{FRET}}(r)$ は，色素分子間の距離を r として

$$E_{\mathrm{FRET}}(r) = \frac{1}{1 + (r/R_0)^6} \tag{2.1}$$

と表される．ただし， R_0 は Förster 距離と呼ばれる定数で，$E_{\mathrm{FRET}} = 1/2$ になる距離を表し，色素の種類（組み合わせ）や媒質の誘電率等によって変化するが，典型的には 5–10 nm 程度の値になる．E_{FRET} は r の 6 乗に依存し，図 2（B）を見ても分かるように，R_0 近傍で非常に鋭敏に距離に依存する．実験では，ドナーおよびアクセプタの蛍光強度 I_{D}，I_{A} を別々に計測し，それらの比，すなわち $E_{\mathrm{FRET}} \equiv I_{\mathrm{A}}/(I_{\mathrm{A}} + I_{\mathrm{D}})$ を得る（実際には，実験条件によって生じるずれを補正する必要がある）ことで，蛍光信号から距離情報に換算することができる．

蛋白質分子はしばしば「分子機械」に例えられるように，ドメインと呼ばれるいくつかのブロック構造が連結された形状をしている場合が多く，その機械的な動きが重要な役割を果たしていると考えられている．FRET の性質を利用すれば，分子の 2 カ所に色素をラベルすることで，E_{FRET} の変化から分子構造に関する情報を得ることができる（図 2（A））．そしてこれは蛍光計測であるので，1 分子感度を実現でき，個別分子の構造変化を観察することも可能となる．

2.4 データ解析への要請

さて，1 分子 FRET 実験データの解析について考えよう．例として，1 分子 FRET 実験で得られる蛍光信号(シミュレーション結果)を図 2(C)に示した．微弱な蛍光のため上述のショット・ノイズも無視できず，結果として，蛍光強度比として計算される FRET 効率も大きな揺らぎを含む(図 2(D)のグレー線)．

一方で，(FRET に限らず) 1 分子計測で得られる信号の特徴として，ステップ的な変化を示す場合が多いことが挙げられる．これは通常，分子がとり得る有限個の状態があり，その間を遷移するダイナミクスを反映していると解釈される．すなわち，分子がある状態に留まる間は信号は一定レベルを保ち，ある時に別の状態へと瞬間的に遷移する．この時，もし状態間で信号レベルに十分に大きな差があり，揺らぎが十分に抑えられていれば，目視で明らかに状態遷移を言い当てることが出来る．あるいは，適当なしきい値を設定して機械的に検出することも容易にできるだろう．しかし，図 2(D)のようにそれぞれの信号レベルが接近し，揺らぎも大きい場合には簡単ではない．解決策としては，十分にビンサイズを大きくして積算時間を増やし，揺らぎを抑えるアプローチが考えられる．しかし，トレードオフとして時間分解能が失われることになり，必ずしも望ましくない．

そこで，発想を転換してみよう．われわれが知りたい情報は分子の状態であり，信号を美しく整形することは必ずしも必要ではない．一見乱雑な信号からでも状態遷移の情報を取り出すことができればよいのである．すなわち，時系列データの解析法に求める条件は，実験から得られる時系列データ(図 2(D)のグレー線)から状態遷移軌跡(同 黒線; シミュレーションに用いた FRET 変化曲線)を取り出すことである．この問題は，分子ダイナミクスの詳細(状態数や各状態の信号レベル等)が既知であれば，比較的簡単なフィッティングで解決できるかも知れない．しかし実際の実験では，そのような情報は与えられていないことが多い．そもそも，それが分からないから実験をおこなうのである．したがって，モデル・フリーで実験データのみから状態数を決定し，状態遷移の軌跡を復元できることが求められる．

このような要請に応える解析法として，隠れマルコフモデルや変化点検出法など，いくつもの試みが既になされている．以下では，そのいくつかを紹介する．

3. 隠れマルコフモデル

前節で述べた要請に応えるデータ解析法の 1 つとして，隠れマルコフモデル(hidden Markov model; HMM)がある．1 分子(FRET)実験データに関しても，HMM を用いた解析法が提案されている．

3.1 最尤法による取り扱い

時系列データ $\boldsymbol{X}=\{x_1,x_2,\ldots,x_N\}$ が得られたとき，各点に対応する潜在変数 $\boldsymbol{Z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_N\}$，$\boldsymbol{z}_n=\{z_{n1},z_{n2},\ldots,z_{nK}\}$ を考える．ただし，K はデータ源が取り得る状態数とし，x_n の時点で系が状態 k にある時，$z_{nk}=1$，$z_{nj}=0(j\neq k)$で状態を表す．時系列データは等時間間隔でサンプリングされ，データ源は単純マルコフ過程で状態を変化させるとすると，グラフィカルモデルは図 3(A)のようになり，データと潜在変数の同時確率分布は一般的に

$$p(\boldsymbol{X},\boldsymbol{Z}|\Theta)=p(\boldsymbol{z}_1|\boldsymbol{\pi})\times\prod_{n=2}^{N}p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1},\boldsymbol{A})\times\prod_{m=1}^{N}\prod_{k=1}^{K}p(x_m|\phi_k)^{z_{mk}} \tag{3.1}$$

で表すことができる．第 1 項 $p(\boldsymbol{z}_1|\boldsymbol{\pi})=\prod_{k=1}^{K}\pi_k{}^{z_{1k}}$ は，初期状態 $\boldsymbol{z}_1$ として取り得る状態の確率分布を表し，パラメータ $\boldsymbol{\pi}=\{\pi_k\}$ で表される．第 2 項は，状態遷移確率を表す確率分布で

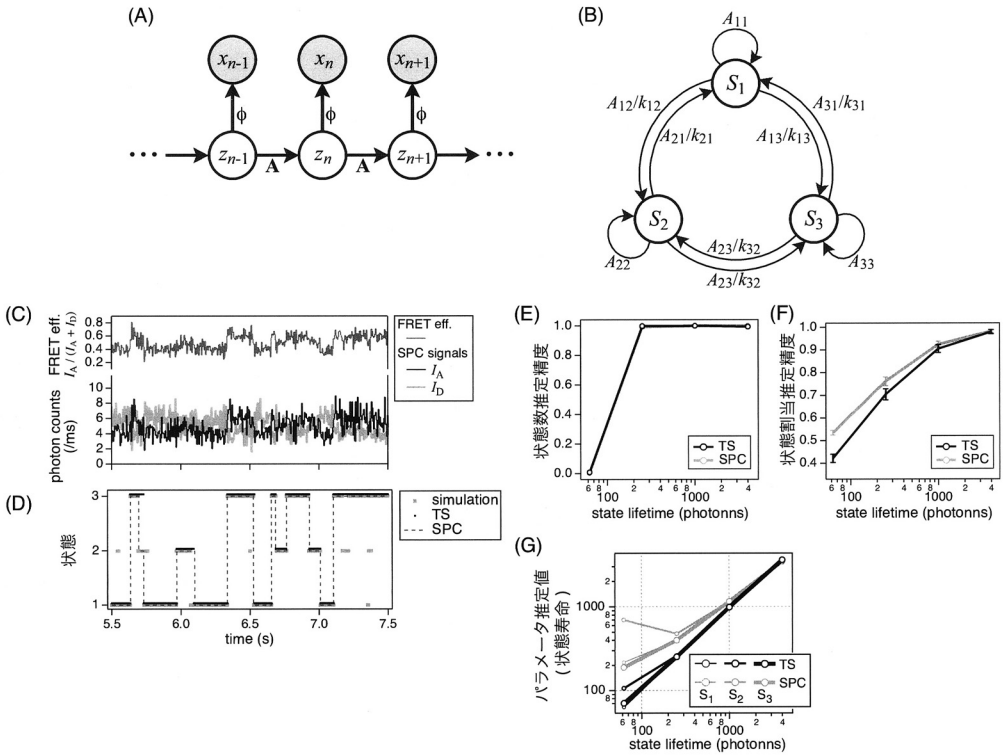

図 3.（A）HMM のグラフィカル・モデル．（B）HMM で用いられる状態遷移確率 $\boldsymbol{A}$（状態遷移速度 $\boldsymbol{k}$）．3 つの状態 S_1，S_2，S_3 の間の遷移確率（速度）を表す．（C）シミュレーションで生成した FRET 計測データ．（D）VB-HMM 法によって（C）から復元された状態遷移軌跡．（E）–（G）シミュレーションによる VB-HMM 解析法の性能評価．（E）状態数推定の正解率．（F）フォトン単位（SPC に関してはビン単位）での状態割り当ての正解率．（G）シミュレーションパラメータの 1 つ，状態寿命の推定結果．［（C）–（G）Reprinted from Biophysical Journal, Vol. 103, K. Okamoto and Y. Sako, ‘‘Variational Bayes Analysis of a Photon-Based Hidden Markov Model for Single-Molecule FRET Trajectories’’, 1315–1324, Copyright（2012）, with permission from Elsevier.］

$p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, \boldsymbol{A}) = \prod_{i=1}^{K}\prod_{j=1}^{K} {A_{ij}}^{z_{n-1,i}z_{nj}}$ で表される．パラメータ $\boldsymbol{A}$ は $K\times K$ 行列で，要素 A_{ij} は i 状態から 1 ステップ後に j 状態に移る確率（$i=j$ の場合，同じ状態に留まる確率も含む）を与える（図 3（B））．第 3 項は潜在変数と実測データを結びつける出力確率で，$p(x_m|\phi_k)$ は k 状態にある分子が信号 x_m を生成する確率を表す．ϕ_k は k 状態についての関数形を定めるパラメータの集合であり，たとえばガウス関数 $p(x_m|\mu_k, {\sigma_k}^2) = (2\pi{\sigma_k}^2)^{-1/2}\exp\{-(x_m-\mu_k)^2/2{\sigma_k}^2\}$ を与える場合，ガウス分布の平均値 μ_k と分散 ${\sigma_k}^2$ がパラメータとなる．Θ は，$\boldsymbol{\pi}$，$\boldsymbol{A}$，$\boldsymbol{\phi}$ を含むすべてのパラメータを表すとする．マルコフ過程であれば，第 1・2 項には一般性があるので，第 3 項の関数形とパラメータ $\{\phi_k\}$ によってモデルを表すことになる．

式（3.1）で表される HMM を EM（expectation maximization）アルゴリズムを用いた最尤法によって解く方法は，すでに確立されている（Bishop, 2006）．詳細は省くが，まず仮のパラメータ集合 Θ^{old} を与え，バウム-ウェルチ（Baum-Welch）法やビタビ（Viterbi）法などにより最適な $\boldsymbol{Z}$ の事後分布を得る E-ステップを計算した後，M-ステップで Θ を更新する．この E-ステップ

と M-ステップとの反復計算を $p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)$ が収束するまで繰り返すことで，状態数 K を仮定した上での $\boldsymbol{Z}$ 分布とパラメータ Θ の最適値を得る．同様の計算を状態数 K を変えながら繰り返して最適な状態数を決める，すなわち，状態数の異なるモデルの中から最適なモデルを選択する必要がある．しかしこの時，K 毎に最適化された $p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)$ の値を直接比較することはできない．最尤法では尤度のみが評価され，状態数が増えることに対するペナルティがないため，状態数を増やせば増やすほど，ノイズを含めたデータの揺らぎや外れ値に対応できる過学習が起きるためである．したがって，赤池情報量規準（Akaike's Information Criterion; AIC）やベイズ情報量規準（Bayesian Information Criterion; BIC）などの検定を別途用い，状態数の増加に対するペナルティを課して最適な状態数を決定する必要がある．

最尤法を用いた HMM の 1 分子 FRET 実験データへの応用に関しては，蛍光強度にガウス分布やポアソン分布，FRET に関してはガウス分布やベータ分布を仮定したモデルを用い，状態数推定に AIC や BIC を用いる方法が報告されている（McKinney et al., 2006; Liu et al., 2010）．

3.2 変分ベイズ法による取り扱い

最尤法の解法は，尤度関数のパラメータでの微分に基づいている．そのため前述のように過学習が起き，状態数を制約する手続きが別途必要になる．一方，尤度関数の代わりに，パラメータ分布を周辺化によって組み込んだモデルエビデンス $p(\boldsymbol{X}) = \int p(\Theta)p(\boldsymbol{X}|\Theta)d\Theta$ を評価する方法がある．パラメータは分布関数で表され，その汎関数であるエビデンスの汎関数微分（変分）に基づいてモデルを解くことから，変分ベイズ（variational Bayes; VB）法と呼ばれる．

この場合も，われわれの目的は $p(\boldsymbol{Z}, \Theta|\boldsymbol{X})$ の最適解を求めることであり，その近似解 $q(\boldsymbol{Z}, \Theta)$ を導入するとエビデンスの対数は，

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{3.2}$$

と変形できる．ただし，

$$\mathcal{L}(q) = \sum_{\boldsymbol{Z}} \int q(\boldsymbol{Z}, \Theta) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}, \Theta)}{q(\boldsymbol{Z}, \Theta)} \right\} d\Theta \tag{3.3}$$

$$\mathrm{KL}(q||p) = -\sum_{\boldsymbol{Z}} \int q(\boldsymbol{Z}, \Theta) \ln \left\{ \frac{p(\boldsymbol{Z}, \Theta|\boldsymbol{X})}{q(\boldsymbol{Z}, \Theta)} \right\} d\Theta \tag{3.4}$$

とする．$\mathrm{KL}(q||p)$ はカルバック-ライブラー（Kullback-Leibler）ダイバージェンスであり，2 つの分布関数 p, q の相似の程度を表す汎関数である．$\mathrm{KL}(q||p)$ は常に正の値を取り，p と q が一致する時にのみ 0 になる．したがって，$\mathcal{L}(q) \leq \ln p(\boldsymbol{X})$ であり，$\mathcal{L}(q)$ は $\ln p(\boldsymbol{X})$ の下界を表すことから，変分下限とも呼ばれる．つまり，$\mathcal{L}(q)$ を最大化する $q(\boldsymbol{Z}, \Theta)$ を求めることが目的となる．

ここで，

$$-\mathcal{L}(q) = -\sum_{\boldsymbol{Z}} \int q(\boldsymbol{Z}, \Theta) \ln p(\boldsymbol{X}, \boldsymbol{Z}, \Theta) d\Theta + \sum_{\boldsymbol{Z}} \int q(\boldsymbol{Z}, \Theta) \ln q(\boldsymbol{Z}, \Theta) d\Theta \tag{3.5}$$

において，$U \equiv -\ln p(\boldsymbol{X}, \boldsymbol{Z}, \Theta)$ と定義し，第 2 項は情報エントロピーの表式であることに注目して S で表すことにすると，

$$-\mathcal{L}(q) = \langle U \rangle - TS \tag{3.6}$$

と表すことが出来る．ただし $\langle U \rangle$ は U の期待値を表し，T は表式をそろえるための意味の無い単位温度とする．式 (3.6) は熱力学におけるヘルムホルツ自由エネルギーと同じ表式になっている．すなわち，$\mathcal{L}(q)$ を最大化することは系の自由エネルギーを最小化することに等しい．第

1 項は，最尤法と同様に，モデルを観測データに一致させることで内部エネルギーを最小化させる効果を表す．それと同時に，第 2 項ではエントロピーを最大化させることを求めており，最尤法で見られるような過学習を抑制する．つまり，AIC や BIC でおこなわれる状態数増加へのペナルティが暗に含まれていることを意味している．

実際に変分ベイズの計算をおこなうには，物理学で用いられる平均場近似と同様の近似を用いる．潜在変数とパラメータをまとめて $\boldsymbol{Z}$ で表すとした場合，$\boldsymbol{Z}$ をいくつかの互いに独立なグループ $\boldsymbol{Z}_i$ $(i = 1, \ldots, M)$ に分解することができる，すなわち $q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i)$ であるとすると，

$$\mathcal{L}(q) = \int q_j \ln \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) d\boldsymbol{Z}_j - \int q_j \ln q_j d\boldsymbol{Z}_j + \text{const.} \tag{3.7}$$

と変形することができる．ただし，$\mathbb{E}_{i \neq j}[\cdots]$ は $i \neq j$ のすべての $\boldsymbol{Z}_j$ による期待値を表すとして，$\tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const.}$ とする．$\{q_{i \neq j}\}$ を固定化した上で $\mathcal{L}(q)$ が最大化するのは，式(3.7)が負の KL-ダイバージェンスの関係であることに注意すると，$q_i(\boldsymbol{Z}_j) \cong \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ の時なので，最適解 $q_j^*(\boldsymbol{Z}_j)$ は，

$$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const.} \tag{3.8}$$

で与えられる．これを $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M$ について交互に計算する．具体的な計算法としては，最尤法と同様の E-M ステップの反復計算を流用することができる．E-ステップは潜在変数 $\boldsymbol{Z}$ の計算に相当し，最尤法の場合と同様にバウム-ウェルチ法やビタビ法などのアルゴリズムを用いることができる．M-ステップでは，その他のパラメータについて式(3.8)を用いて計算する．各ステップの計算が終わると，変分下限 $\mathcal{L}(q)$ を計算して収束の度合いをチェックする．式(3.1)で表される一般的な HMM の場合には，

$$\begin{aligned} \mathcal{L}(q) = {} & \mathbb{E}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)] + \mathbb{E}[\ln p(\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{A})] + \mathbb{E}[\ln p(\boldsymbol{\phi})] \\ & - \mathbb{E}[\ln q(\boldsymbol{Z})] - \mathbb{E}[\ln q(\boldsymbol{\pi})] - \mathbb{E}[\ln q(\boldsymbol{A})] - \mathbb{E}[\ln q(\boldsymbol{\phi})] \end{aligned} \tag{3.9}$$

を計算すればよい(ただし，$\mathbb{E}$ の添え字は省略した)．この計算を状態数 K を変えて繰り返し，収束した $\mathcal{L}(q)$ を直接比較することで，最適な状態数モデルを選択することができる．

変分ベイズ法ではパラメータを周辺化する必要があるため，モデルを決める際には，出力関数だけでなく事前分布も与えなければならない．事前分布の関数としては，特別な意図がない限り，典型的に使えるものもある．たとえば $\boldsymbol{\pi}$ や $\boldsymbol{A}$(の i 列)のように合計が 1 になるよう制約される確率分布の場合，事前分布関数にはディリクレ分布を用いることができる．ガウス関数を表すパラメータ(平均，分散)については，ガウス-ガンマ混合分布で表すことが出来る(Bronson et al., 2009)．パラメータの事後分布関数に関しては特に与える必要はなく，式(3.8)を計算することで自動的に q が求まるので，それらを用いて各パラメータの期待値を計算することができる．

変分ベイズ-隠れマルコフモデル(VB-HMM)法の 1 分子 FRET 実験データへの応用として，FRET(ガウス分布)の時系列データに対する適用が報告されている(Bronson et al., 2009)．

3.3 フォトン信号への拡張

ここまで紹介した解析法では，SPC 計測あるいはカメラによるイメージング計測が仮定され，一定時間ビン毎の蛍光信号，あるいは，FRET 信号が取り扱われてきた．しかし，2.2 節で述べたように，SPC 検出器を用いれば TS 計測を実現でき，信号としては最大の情報量を得ることができる．そこで，TS-FRET 信号を取り扱うための VB-HMM 法の拡張をおこなった(Okamoto and Sako, 2012)．

TS 信号では，データ列 $\{x_n\}$ を，$n-1$ 番目のフォトンから n 番目のフォトンまでの経過時間と定義する(図 1(C))．ここで，データ点間の時間間隔が可変であることに注意する必要がある．通常の HMM では状態遷移確率を定数行列 $\boldsymbol{A}$ で表すが，系が本来特性として持っているのは状態遷移「レート」であり，時間間隔一定の仮定によって確率で表すことが許されていたのである．その仮定が成り立たない以上，遷移確率は時間間隔の関数として書き表さなければならない．そこで，状態遷移確率関数を遷移レートに基づいた形に書き直す．物理・化学現象に直結するレートを計算パラメータとして直接取り扱えるという意味でも，より直感的とも言える．

遷移確率 $\boldsymbol{A}$ と同様に，各状態間の遷移速度 $\{k_{ij}\}$ $(i \neq j)$ を定義する(図 3(B))．この場合，同じ状態に留まる $i=j$ の場合を考える必要はない．さらに，i 状態の減衰速度 $\kappa_{ii} \equiv \sum_{j\neq i}^{K} k_{ij}$ と，i 状態から遷移が起きた時に j 状態に移る確率 $\kappa_{ij} \equiv k_{ij}/\sum_{j\neq i}^{K} k_{ij} = k_{ij}/\kappa_{ii}$ $(i \neq j)$ とを定義する．κ_{ij} は i を除いた j での総和が 1 となる．いま時間間隔は x_n で表されるため，この $\boldsymbol{\kappa}$ を用いて遷移確率関数を書き直すと

$$p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, x_n, \boldsymbol{\kappa}) = \prod_{i}^{K}\left\{\exp(-\kappa_{ii}x_n)^{z_{n-1,i}z_{ni}} \times \prod_{j\neq i}^{K}[\kappa_{ij}\{1-\exp(-\kappa_{ii}x_n)\}]^{z_{n-1,i}z_{nj}}\right\} \tag{3.10}$$

となる．しかし，このままではバウム-ウェルチ法やビタビ法などのアルゴリズムを直接用いることができない．そこで，$p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, \Theta)$ の形に変形することを考える．TS 信号の場合，出力確率は蛍光強度 I をパラメータとして

$$p(x_n|I_i) = I_i \exp(-I_i x_n) \tag{3.11}$$

で与えられるので，x_{n-1} に関する周辺化をおこなう事で，

$$\begin{aligned} p(z_{nj}|z_{n-1,i}, \boldsymbol{\kappa}, \boldsymbol{I}) &= \int p(z_{nj}|z_{n-1,i}, x_{n-1}, \boldsymbol{\kappa}) p(x_{n-1}|\boldsymbol{I}) dx_{n-1} \\ &= \begin{cases} \dfrac{I_i}{\kappa_{ii}+I_i} & (i=j) \\ \dfrac{\kappa_{ii}\kappa_{ij}}{\kappa_{ii}+I_i} & (i\neq j) \end{cases} \end{aligned} \tag{3.12}$$

を得る．これにより，$\boldsymbol{X}$，$\boldsymbol{Z}$，Θ についての同時分布関数

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{Z}, \Theta) = p(\Theta) \times \prod_{i}^{K} \pi^{z_{1i}} \times \prod_{n=2}^{N}\prod_{i}^{K}\left\{\left(\frac{I_i}{\kappa_{ii}+I_i}\right)^{z_{n-1,i}z_{ni}} \times \prod_{j\neq i}^{K}\left(\frac{\kappa_{ii}\kappa_{ij}}{\kappa_{ii}+I_i}\right)^{z_{n-1,i}z_{nj}}\right\} \\ \times \prod_{m=1}^{N}\prod_{i}^{K}\{I_i \exp(-I_i x_m)\}^{z_{mi}} \end{aligned} \tag{3.13}$$

を得ることができた．これを解くには，3.2 節で述べたように，平均場近似により $\boldsymbol{Z}$ の最適分布を得る E-ステップと，パラメータの分布関数を更新する M-ステップを交互に反復計算する．事前分布としては，I_i に関してはガンマ分布，κ_{ii} に関しては尺度不変事前分布 $p(\kappa_{ii}) \propto {\kappa_{ii}}^{-1}$，$\kappa_{ij}(i\neq j)$に関してはディリクレ分布が使える．

ここまでで TS 形式の蛍光信号については取り扱えるようになった．次にこれを FRET 信号へ拡張したい．

FRET 信号では，ドナーとアクセプタのフォトンは別々の検出器で検出され，2 系列の信号として得られる．しかしここでは，両者をまとめて 1 つのフォトン列とみなし，各フォトンの色を区別することにする．そのためデータを拡張し，$\boldsymbol{x}_n = \{\Delta t_n, \rho_n\}$ とする．ただし，ドナーの場合 $\rho=0$，アクセプタの場合 $\rho=1$ と定義する．また，新たなパラメータ，FRET 効率 E を導入する．すると，式(3.13)の同時分布のうち，初期分布確率と遷移確率関数は変更する必

要がなく，出力関数を

$$p(\boldsymbol{x}_m|\boldsymbol{z}_m,\boldsymbol{I},\boldsymbol{E}) = p(\rho_m|\boldsymbol{z}_m,\boldsymbol{E}) \times p(\Delta t_m|\boldsymbol{z}_m,\boldsymbol{I}) = \prod_i^K \{{E_i}^{\rho_m}(1-E_i)^{1-\rho_m} \times I_i \exp(-I_i \Delta t_m)\}^{z_{mi}} \tag{3.14}$$

と書き換えるだけで，TS-FRET 信号のモデルを表すことができる．新たに E に関する事前分布を用意する必要があるが，0–1 の間で値をとる分布としてベータ分布を用いることができる．

フォトン信号のもう 1 つの形式である SPC 信号についても，VB-HMM 法での取り扱いを可能にした．この場合は x_n の時間間隔は一定であるため，状態遷移行列 $\boldsymbol{A}$ を用いた式(3.1)が使える．また，出力関数はポアソン分布関数に基づき，$\boldsymbol{\mu}$ を用いて

$$p(x_m|\boldsymbol{z}_m,\boldsymbol{\mu}) = \prod_{i=1}^K \left\{ \frac{{\mu_i}^{x_m}}{x_m!} \exp(-\mu_i) \right\}^{z_{mi}} \tag{3.15}$$

と書くことができる．

これを FRET 信号に拡張する場合も，TS 信号とは異なる取り扱いをする．2 系列の蛍光信号を，ドナー・アクセプタそれぞれのビンあたりのカウント数 d_n, a_n として $\boldsymbol{x}_n = \{d_n, a_n\}$ に置き換える．出力確率は，ドナー，アクセプタそれぞれのフォトンについてポアソン分布を求めることで

$$p(\boldsymbol{x}_m|\boldsymbol{z}_m,\boldsymbol{\mu},\boldsymbol{E}) = \prod_{i=1}^K \left\{ \frac{(1-E_i)^{d_m}{E_i}^{a_m}}{d_m! a_m!} \times {\mu_i}^{d_m+a_m} \exp\big(-\mu_i\big) \right\}^{z_{mi}} \tag{3.16}$$

と表す．事前分布は μ についてはガンマ分布，E についてはベータ分布を用いることができる．

本節の立式の詳細は，参考文献(Okamoto and Sako, 2012)を参照されたい．

実際の実験では，多数の 1 分子についてそれぞれ時系列データを得ることになる．基本的に分子は共通の振る舞いをすると仮定すると，パラメータをグローバルに評価することでより精度を高められる可能性がある．あるいは，分子の振る舞いに関する知見がいくらか蓄積されれば，パラメータの事前分布関数の形状を規定するハイパーパラメータを調整することも可能だろう．ただし，式(3.14)–(3.16)中の I, E は単純に蛍光信号の和および比で表される見かけの値であり，物理的なパラメータとして扱うためには，実験条件に合わせて補正した表式に改める必要があるが，式が複雑化するため，現状では計算可能な状態に解けていない．

3.4 VB-HMM-TS-FRET 解析法のシミュレーションによる評価

VB-HMM を用いた TS-FRET 解析法の有効性を検証するため，シミュレーション生成したデータを用いてテストした．生成したデータの例を図 3(C)に示す．図では SPC 形式の信号として表しているが，元のデータは TS 信号として生成しており，TS 解析と SPC 解析で結果を比較することができる．この例の場合，元のデータの 3 状態を正しく推定することができ，図 3(D)のように状態遷移軌跡を復元することができた．滞在時間の短い状態変化は取りこぼしているケースもあるが，概ね正しく軌跡を復元することに成功している．

また，データの生成パラメータを変化させながら，各条件で同様の信号生成-解析のプロセスを 1,000 回ずつ繰り返し，その結果を統計として評価した(図 3(E)–(G))．いずれの結果からも，解析の精度が状態あたりの滞在時間(フォトン数)に強く依存していることが分かる．図 3(E)は状態数推定の正解率を示し，各状態にフォトンが 200 個程度あれば状態数をほぼ正確に推定できている．(F)はフォトン単位(SPC の場合ビン単位)での状態割り当ての正解率を示し，

90%以上の正解率を保つためには各状態あたり 1,000 個程度のフォトンが必要であることが分かる．パラメータの推定精度に関しては，状態割り当て正解率が 90%以上であれば，ほぼ正確に推定ができた．

同様のデータを VB-HMM-SPC-FRET 法でも解析をおこなった．その結果は，ビン幅をどのように設定するかに大きく依存するが，適切なビン幅を設定することができれば，TS 解析と遜色ない結果が得られることが分かった．ただし(G)に示すように，状態寿命の推定結果に関しては，TS 解析の方がより正確で，ほぼ完璧な推定ができた．

結論としては，TS 解析とビン幅を適切に設定した SPC 解析とでは，ほぼ同等の性能が得られると考えられる．ただし，実際の実験では，未知の現象を取り扱う必要があり，事前に最適なビン幅を知ることは難しい場合が多い．また，計測の最中にフォトンレートが大きく変化するケースも考えられる．TS 計測の場合，事前にビンを設定する必要がなく，実験中にレートが大きく変わるケースへも対応でき，常に最大の情報量を得ることができるメリットは大きいと考えられる．

3.5 1 分子実験データへの適用例：ホリデー・ジャンクション DNA の FRET 計測

最後に，VB-HMM-TS-FRET 解析法を実際の実験データに適用した例について紹介する．

DNA は通常，2 本でらせん構造を構成して安定に存在している事がよく知られている．しかし，4 本の DNA が十字形を構成するホリデー・ジャンクション(Holliday junction; HJ)と呼ばれる構造が形成される場合がある．HJ は細胞の減数分裂過程などで見られ，父系由来と母系由来の遺伝情報を一部組み換える相同組み換えに関与する．HJ には，腕が伸び縮みするブランチ・マイグレーションと呼ばれる現象がある．対になる腕に FRET ラベルを施し，この伸び縮み運動を 1 分子 FRET 計測によって観察した(図 4(A))．伸び縮みは塩基単位で起きるので，離散的な構造変化が観察でき，また塩基配列を選ぶことでとり得る状態数を決めることもできる．

得られた実験データの例を図 4(B)に示す．蛍光信号(上段)からトータル蛍光強度 I_c(中段，グレー)および FRET (下段，グレー)を計算により求めた．I_c は 2 つの蛍光信号の和に実験条件を考慮した補正を施した値で，この値が一定であれば，蛍光信号の揺らぎが FRET 変化に由来するものであると見なすことができる．FRET 信号はステップ的な時間変化を示しているが，揺らぎも大きい．この TS 信号に VB-HMM-TS-FRET 解析を適用し，I_c(中段，黒)および FRET(下段，黒)の軌跡を復元した．得られた FRET 軌跡は，塩基配列から期待される通り 3 状態間の遷移を示した．図 4(C)に FRET のヒストグラムを示す．実験データからの計算値(グレー)では分布が広がり状態を識別できないのに対して，VB-HMM 解析結果の状態遷移軌跡(黒)では 3 状態が明確に区別できている．

4. 変化点検出法

HMM の他に，1 分子蛍光の時系列信号を取り扱う解析方法としては，変化点検出(change point detection; CPD)法も提案されている．分子のそれぞれの状態が固有の確率関数にしたがって信号を生成するとすると，状態毎に信号の統計的性質(平均や分散，分布の形状など)は異なる．CPD 法では，その名の通り，データの性質が変化する点を検出することで，状態遷移を検出する．以下で，TS 形式の(1ch の)蛍光強度信号についての CPD 法について簡単に紹介する．

4.1 対数尤度比検定による CPD

時間 T の間にフォトン N 個が検出されたデータ区間に対し，k 番目のフォトン p_k で状態遷移が起きたかどうかを評価したい(図 5(A)–(C))．もし p_k が変化点であれば，その前後で平均

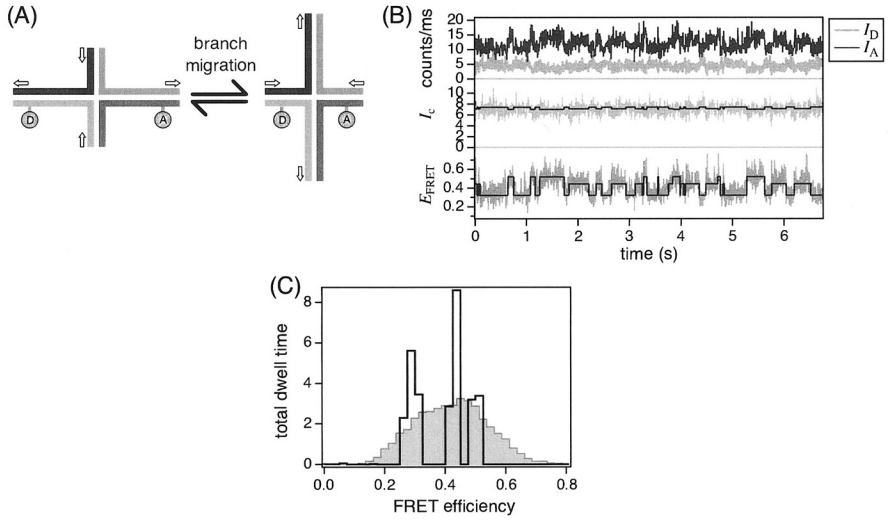

図 4. VB-HMM 法の 1 分子 FRET 実験データへの適用例．(A)DNA によって作られる十字構造であるホリデー・ジャンクション．左右の腕の伸び縮み運動を FRET で計測する．(B)実験データの例．上段は蛍光信号の測定データ．中段・下段のグレーは実験データからの計算値，黒は VB-HMM 法による解析結果．(C)FRET 分布．実験データ(グレー)と VB-HMM 法による解析結果(黒)．[Reprinted from Biophysical Journal, Vol. 103, K. Okamoto and Y. Sako, ‘‘Variational Bayes Analysis of a Photon-Based Hidden Markov Model for Single-Molecule FRET Trajectories’’, 1315–1324, Copyright(2012), with permission from Elsevier.]

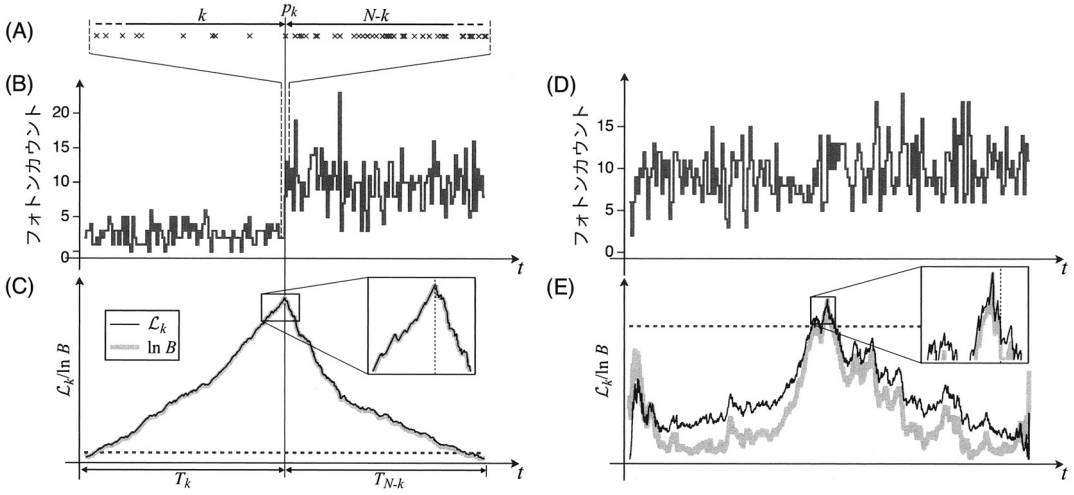

図 5. (A)–(C)TS 信号の変化点検出法の概念．あるデータ区間のうち k 番目のフォトン p_k で信号が変化するか評価する．(B)(D)シミュレーションで生成した，データ区間の中央で信号が変化する蛍光時系列信号．I_1/I_2 =(B)0.2,(D)0.9．(C)(E)対数尤度比 $\mathcal{L}_k$ およびベイズ因子 B の計算結果．値がしきい値(破線; $\mathcal{L}_k$ についての理論値)を超えれば変化点があると見なす．点線は実際の変化点の位置．

値の異なる確率分布関数が仮定されるので，区間全体の尤度はポアソン分布に基づいて

$$L_{\mathrm{A}} = \frac{(\hat{I}_1 T_k)^k e^{-\hat{I}_1 T_k}}{k!} \cdot \frac{(\hat{I}_2 T_{N-k})^{N-k} e^{-\hat{I}_2 T_{N-k}}}{(N-k)!} \tag{4.1}$$

として得られる．ただし，$\hat{I}_1 = k/T_k$，$\hat{I}_2 = (N-k)/T_{N-k}$，$T_k$ はデータ始点から p_k までの時間，$T_{N-k}(= T - T_k)$ は p_k からデータ終点までの時間とする．逆に，p_k が変化点でない(この区間内に変化点が存在しない)場合を帰無仮説とすると，尤度は

$$L_{\mathrm{o}} = \frac{(\hat{I}_0 T_k)^k e^{-\hat{I}_0 T_k}}{k!} \cdot \frac{(\hat{I}_0 T_{N-k})^{N-k} e^{-\hat{I}_0 T_{N-k}}}{(N-k)!} \tag{4.2}$$

となる(ただし $\hat{I}_0 = N/T$)．これらの尤度の比の対数 $\mathcal{L}_k = \ln(L_{\mathrm{A}}/L_{\mathrm{o}})$ を評価して有意に高ければ変化点が存在すると見なすことができ，区間内で最も大きい $\mathcal{L}_k$ を示すフォトンを変化点と見なす(図 5(C))．変化点が検出されれば，区切られたそれぞれの区間に対して同様の評価をおこなう．このプロセスを再帰的に繰り返すことで，データに含まれる変化点すべてを検出する(Watkins and Yang, 2005)．

ポアソン過程の CPD 法は古くから研究されており，たとえば炭鉱事故の発生頻度の分析も試みられている(Cox and Lewis, 1966)．Watkins らもそれらの成果を活用し，しきい値の設定や，区間端での補正，信頼区間の評価などを TS 信号に応用することに成功している．

4.2 ベイズ因子による CPD

対数尤度比の代わりに，ベイズ因子を用いて変化点の判定をおこなう手法も提案されている(Ensign and Pande, 2010)．式(4.1)–(4.2)と同様に，変化点がある場合と無い場合で周辺尤度を計算し，両者の比であるベイズ因子

$$B = \frac{\int p(\Theta) p(\boldsymbol{X}|\Theta, H_2) d\Theta}{\int p(\Theta) p(\boldsymbol{X}|\Theta, H_1) d\Theta} = \frac{p(\boldsymbol{X}|H_2)}{p(\boldsymbol{X}|H_1)} \tag{4.3}$$

を求める．ただし，H_1, H_2 はそれぞれ帰無仮説，対立仮説を表すモデルとする．尤度関数 $p(\boldsymbol{X}|\Theta, H_{1,2})$ に関しては式(4.1)–(4.2)と同様の関数を用いることができるが，その他に事前分布関数を与える必要がある．Ensign らは TS 強度信号を仮定し，事前分布の影響を抑えるために分散の大きい正規分布を与えて，ベイズ因子を求めることに成功した．

4.3 シミュレーションによる評価

両者の有効性を検証するため，シミュレーション生成したデータの解析をおこなった．区間の中央で強度が変化する TS 信号を生成し(図 5(A))，各フォトンについて $\mathcal{L}_k$ および B を計算した(図 5(C))．強度変化が十分に大きい場合($I_1/I_2 = 0.2$; 図 5(B))では，$\mathcal{L}_k$ はしきい値(破線; 理論値)を大きく越え，ほぼ実際の変化点(点線)の位置で最大値をとっている．ベイズ因子も対数を取ると，(絶対値は異なるが)ほぼ同様の分布を示している．強度変化がわずかな場合には変化点を検出できないケースも増えるが，図 5(D)の例($I_1/I_2 = 0.9$)ではしきい値を越える $\mathcal{L}_k$ が検出されている．

このように，一見，有効に見える CPD 法だが，筆者が試みたところ，実用上は問題があるように感じられた．まず，対数尤度比にせよベイズ因子にせよ，それ自体は任意のモデルに関して導出することが比較的簡単にできる．しかし，実際に判定をおこなうためのしきい値を決定する理論が(過去の蓄積があるポアソン過程の対数尤度比の場合を除き)欠けている．そのため，容易に他のモデル(例えば 2ch の FRET 信号)に応用することができない．対照実験で経験的に決められる場合もあるかもしれないが，できれば理論的に裏付けられる方が望ましい．ま

た，対数尤度比を用いた CPD 法に関して，前述の VB-HMM 法と同様のシミュレーションをおこなって性能を比較したところ，VB-HMM ほどの精度を得ることができなかった(Okamoto and Sako, 2012)．VB-HMM 法では常に全体のフォトン分布を使って最適化をおこなうのに対して，CPD 法では再帰的に区間を区切ることになり，変化点周辺の限られたフォトンだけから評価をおこなうことがエラー率を高める結果につながっているのかもしれない．

5. おわりに

近年では，生物学の実験においても統計的手法や情報理論に基づいたデータ解析が用いられている．本稿では 1 分子 FRET 計測における時系列データの取り扱いについて紹介したが，特に 1 分子実験では，分子レベルの確率過程を顕わに取り扱うことになるため，統計的な解析が馴染みやすい．他にもたとえば，細胞膜に埋まって拡散運動する蛋白質では，拡散速度の異なる状態間を切り替わるダイナミクスに HMM が用いられ解析されている(Chung et al., 2010)．また，S/N 比を容易に大きくできない 1 分子画像において，分子の位置や動きを検出する画像解析でも，さまざまな手法が提案されている(たとえば Jaqaman et al., 2008)．しかし，当然，成功した例ばかりではなく，解決すべき課題も残されている．本稿で紹介した内容でも，たとえば，3.3 節において遷移確率として時間依存した関数形 $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, x_{n-1}, \Theta)$ を直接取り扱うことができれば，より精度を高めることができるかもしれない．また，CPD 法において任意の尤度関数に関してしきい値の求める方法論が確立できれば，さまざまな計測に応用することが可能になる．それ以外にも，これまでは注目されていないが，適切なデータ解析を適用することで重要な情報が得られる実験データもあるだろう．従来は，生物学の実験者がデータ解析法を勉強して応用する場合が主だったと思われるが，そのアプローチにも限界はある．数理解析の専門家であれば解決できる問題も数多くあるに違いない．本稿の読者の中から，生物学実験やそのデータ解析に興味を持って取り組んでいただける方が現れれば幸いに思う．

参 考 文 献

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez, R. L., Jr. and Wiggins, C. H. (2009). Learning rates and states from biophysical time series: A Bayesian approach to model selection and single-molecule FRET data, *Biophysical Journal*, **97**, 3196–3205.

Chung, I., Akita, R., Vandlen, R., Toomre, D., Schlessinger, J. and Mellman, I. (2010). Spatial control of EGF receptor activation by reversible dimerization on living cells, *Nature*, **464**, 783–787.

Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events*, Methuen, London.

Ensign, D. L. and Pande, V. S. (2010). Bayesian detection of intensity changes in single molecule and molecular dynamics trajectories, *The Journal of Physical Chemistry B*, **114**, 280–292.

Förster, Th. (1946). Energiewanderung und Fluoreszenz, *Die Naturwissenschaften*, **33**, 166–175.

Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S. L. and Danuser, G. (2008). Robust single-particle tracking in live-cell time-lapse sequences, *Nature Methods*, **5**, 695–702.

Lakowicz, J. R. (2006). *Principles of Fluorescence Spectroscopy*, 3rd ed., Springer, New York.

Liu, Y., Park, J., Dahmen, K. A., Chemla, Y. R. and Ha, T. (2010). A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis, *The Journal of Physical Chemistry B*, **114**, 5386–5403.

McKinney, S. A., Joo, C. and Ha, T. (2006). Analysis of single-molecule FRET trajectories using hidden Markov modeling, *Biophysical Journal*, **91**, 1941–1951.

Okamoto, K. and Sako, Y. (2012). Variational bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories, *Biophysical Journal*, **103**, 1315–1324.

Watkins, L. P. and Yang, H. (2005). Detection of intensity change points in time-resolved single-molecule measurements, *The Journal of Physical Chemistry B*, **109**, 617–628.

Discrimination of Molecular States in Single-molecule Experimental Data by Statistical Data Analysis

Kenji Okamoto

Cellular Informatics Laboratory, RIKEN

In recent biology study, it is becoming increasingly important to understand the behavior and functions of biomolecules such as proteins. Single-molecule measurement enables us to observe individual molecules and has been widely used to investigate molecular dynamics directly. One of major problems of single-molecule experimental data is weakness of the signal obtained from only a single molecule, leading to the significant fluctuation. On the other hand, its time series can be interpreted as consecutive transitions among a limited number of molecular states in many cases. Since both of those features are stochastic processes, statistical data analysis is suitable to handle such signals. To date, a number of methods have been developed to discriminate molecular states buried in apparently noisy signals. This article introduces some statistical methods to treat experimental data of single-molecule FRET (smFRET) measurements, which can examine the structural dynamics of biomolecules. The hidden Markov model (HMM) reproduces a state transition trajectory from a time series of smFRET data. Maximum likelihood estimation (MLE) or variational Bayes (VB) inference is employed to solve the HMM. The VB-HMM is modified to treat time-dependent time-stamp signals and some examples of applications to simulated signals and experimental data are shown. Finally, I would like to mention another method to analyze time series data: change point detection.

Key words: Biomolecules, single-molecule measurement, FRET, hidden Markov model, variational Bayes, state discrimination.