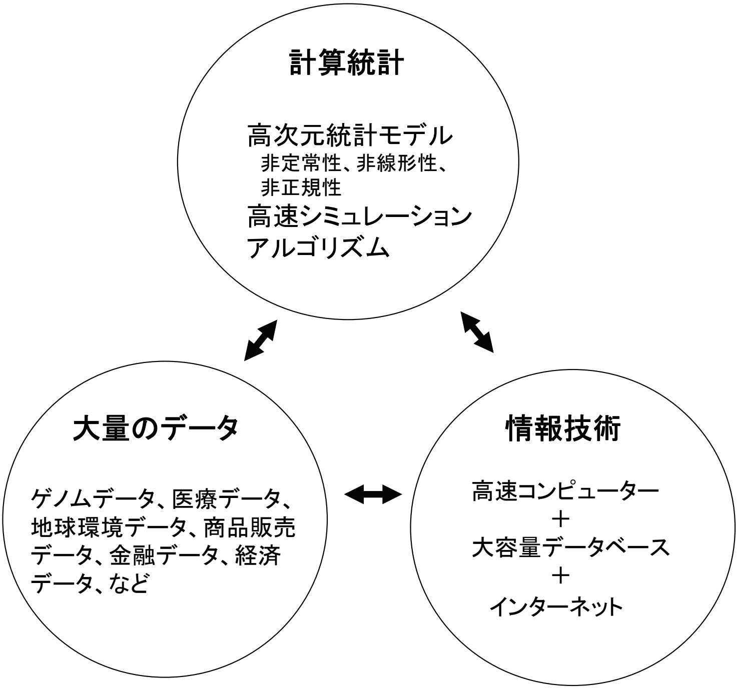
1

計算統計入門

手塚 集


計算統計
高次元統計モデル
非定常性、非線形性、
非正規性
高速シミュレーション
アルゴリズム
大量のデータ
ゲノムデータ、医療データ、
地球環境データ、商品販売
データ、金融データ、経済
データ、など
情報技術
高速コンピューター
＋
大容量データベース
＋
インターネット

第 1 章

ビュッフォンの麺

確率的な考え方を計算に持ち込んだ最初の例としてよく知られているのが，「ビュッフォンの針」である．ビュッフォンは 18 世紀フランスの貴族（侯爵）で博物学者として多彩な仕事を残した人物だ．その仕事の 1 つがこの話である．具体的には，確率的なアプローチを使って π の計算をするというものだ．幅が一定の板を敷き詰めた床の上に針をでたらめに落とすことを考えよう．ビュッフォンは，その針が板の境界線と交わった回数を，針を落とした回数で割った比から π が計算できることを見つけたのである．簡単のため，板の幅（平行線の間隔）を 1 とし，針の長さを ℓ としよう．ここで，$\ell < 1$ を仮定する．するとでたらめに針を落として，それが平行線と交わる確率 p は，次のように考えれば求まる．図 1.1 に示すように，針の中心の座標を x，針の平行線に対する角度を θ とおくことで，針が床の上に落ちたときの状態を表すことができる．ここでは，針をでたらめに落とすことを考えているので，x は区間 $[0,1)$ 内の一様分布，θ は区間 $[0,\pi)$ 内の一様分布に従い，互いに独立である．また，確率変数 χ を次のように定義しよう．

$$\chi(x,\theta) = \begin{cases} 1, & \text{針が境界線に交わった場合}, \\ 0, & \text{その他}. \end{cases}$$

すると，確率変数 χ の期待値は

$$\mathrm{E}(\chi) = 1 \times p + 0 \times (1-p) = p$$

となって交わる確率に一致することがわかる．

まず，問題を単純化して針が平行線に対して常に垂直に落ちると仮定しよう．この場合は自由度が x だけになるので，交わる確率は長さ ℓ に一致する．つま

2
代数生物学

吉田 寛

記号計算 (Symbolic Computation) とは，「式をそのまま計算する」ことである．多項式という代数構造上での計算という意味で，代数計算 (Algebraic Computation) とも呼ばれる．記号計算の計算過程においては，すべての計算が正確に行われ，浮動小数点演算のような近似的数値計算は用いられない．たとえば， $x_0 = \sqrt{2}$ なる数を，浮動小数点でいったん表してしまうと，途中で $x_0^2 - 1$ を計算しても 厳密に 1 になることは保証されない．それとは対照的に，記号計算では，いつでも 1 になることが保証されている．われわれが藁半紙の上で，式を変形して別の有用な式を導出したりすることを，記号計算が代わりに，大規模かつ厳密に行うものと考えてもよいかもしれない．事実，記号計算は一時代前に，天文学分野で現れた手に負えない膨大な数式の処理において，その有用性を知らしめたのであった．

いま，なぜ記号計算なのか．それは，学問的に意義深く，かつ，面白くて大きな問題を，記号計算で解くには，ギガバイト (Giga Byte) オーダーのメモリが必要となるからである．そのような問題は，20 世紀においては，とんでもなく高くついたであろう．逆をいえば，これからの計算機は，標準でギガバイト，さらには，テラバイトオーダーのメモリを積むであろうし，そのような計算機を使うと，どんどん面白くて大きな問題が解けると期待される．

第１章

多細胞系の形式言語による理解と記号計算による関係式の導出

多細胞生物の発生は，ダイナミカルなものであり，１つの細胞からさまざまな細胞タイプが生じる．本章では，このような系を形式言語という枠組みで理解しようと試みる．まず最初に，単細胞から多細胞系への進化過程を比較的簡単に追うことのできる緑藻について形式言語理論の適用を試みる．これは，形式言語理論において比較的標準的手法である階層分けの手法に基づいている．その次に，ゴキブリの脚の挿入再生能力を模した１次元鎖状細胞群のモデルを，形式言語理論の一分野であるリンデンマイヤ・システムをもとに解析する．このモデル上で，限量記号消去法 (Quantifier elimination) なる代数的手法を用いることにより，細胞タイプ多様性の厳密な関係式を求めた．

本章は，以下のような構成である．

1. 代数生物学のはじまり
2. 多細胞系の理解に向けて I:
形式言語によるクラミドモナスからボルボックスへの形態進化の理解
3. 多細胞生物の理解に向けて II:
限量記号消去法による多細胞系の多様性の理解

最初に，「代数生物学」という分野が，どのような意図をもって創設されたのかを紹介した後，緑藻類のクラミドモナスからボルボックスをとりあげる．そして，これをどのように形式言語の俎上に載せるかを見る．そこでは，通常文字列を取り扱うことの多い形式言語を，文字列のような「並び」が意味のない化学物質に対して適用した．緑藻が生成しうる化学物質（細胞外基質）の複雑度の階層の概念により緑藻の進化過程に，ある説明を与えた．次に，形式言語の