論 文 の 内 容 の 要 旨

論文題目　　An Efficient Algorithm for Distributions of Various Feature Values in Bioinformatics Problems
（情報生命科学の諸問題における特徴量分布の効率的計算方法）

氏　　名　　森　遼太

配列アラインメントやRNA高次構造予測、コーディング領域予測、進化系統樹予測、クロマチン状態予測といったように情報生命科学分野には様々な推定問題がある。 これらの問題の多くは、解として可能な全ての候補の中から1つを選択することを目的としており、そのような推定様式は解の点推定と呼ばれる。 点推定の方法には問題毎に様々なものが提案されているが、一般的に解候補の数があまりに膨大なため、たとえ最も確率の高い解を選んだとしてもその確率すら0に近いというようなことが頻繁に生じる [1]。 つまりいかに素晴らしい推定手法を用いようと、そもそも点推定によって得られる解そのものが全く信用に足らないというようなケースが多い。 加えて、扱う問題によっては互いに大きく異なる複数の有力な解候補が存在するような場合もあり[2]、点推定によって解を一つに固定し解候補群の持つ豊富な情報の殆どを破棄してしまう事は、その後の生物学的解析の過程で知見の見逃しや誤謬を招く危険性をはらんでいる。

そこで本研究では、解の分布の分配関数と解ごとに定義される何らかの特徴量がそれぞれ同型の動的計画法で計算可能な場合に、点推定解を介さず解分布全体を考慮した特徴量の分布を高速に求める手法を提案する。 特徴量が有界の整数値を取る場合については既に配列アラインメント問題[3]やRNA二次構造予測問題[4]で効率的なアルゴリズムが提案さ

れているが、本研究はこれが有界実数値を取る場合であっても、[3, 4]と同程度の計算量で目的の分布が計算可能となることを示す。

METHODS

特徴量が整数値の場合、本研究が扱う特徴量の分布は以下のように定式化できる。

$$p_v = \sum_{\theta \in C_v} p(\theta|D) = \sum_s Z_s \delta_{sv} / Z$$

ここで、Cvは特徴量がvとなる解候補の集合で、Dは入力データ、Zは解分布の分配関数、Zsは解のうち特徴量がsとなるもののみを小計した分配関数の部分和、δはクロネッカーのデルタとする。 このような場合は、動的計画法と離散フーリエ変換を組み合わせる事によって効率的に計算する手法が存在する[3, 4]。 一方で、特徴量が有界実数値を取る場合は、一般に候補解それぞれが全く異なった値を取ることになり、上に挙げたように特徴量のそれぞれの値を数え上げるような手法は使うことができない。そこで代替として特徴量の幅を均等にN分割し、そのv番目のビンに対して以下のような値を定める。 なお本稿においては表記を簡潔にするため以降は特徴量の最小値を0とし、ビンのサイズを1と置くが、実際にはこれらに制限はない。

$$p_v = \sum_\theta z_\theta f(s_\theta, v) / Z$$

ここで、z_θは解θのボルツマン因子、s_θは解θの持つ特徴量とする。 また、関数fは特徴量をビンに振り分ける窓関数であり、例えば、

$$f(s_\theta, v) = \begin{cases} 1 & (v - 1/2 \leq s_\theta < v + 1/2) \\ 0 & (otherwise) \end{cases}$$

とすればpvは特徴量がv-1/2以上v+1/2未満となる確率を表すことになる。 ところがこの窓関数は任意に設定できるわけではなく、1. 効率的な計算アルゴリズムが構築可能であることと、2. 各ビンへの重みの総和が1となることの二点が必要であり、上記のような関数の場合は前者の条件を満たすことができない。 そこで、本研究では１，２を満たす関数として以下のような窓関数を提案する。

$$f(s_\theta, v) = \frac{1}{N}\int_{v-\frac{1}{2}}^{v+\frac{1}{2}} \cos\left[\pi\frac{(s_\theta - t)(N-1)}{N}\right]\frac{\sin\left[\pi(s_\theta - t)\right]}{\sin\left[\pi\frac{s_\theta - t}{N}\right]}dt$$

また、この関数はN→∞の極限において、

$$\lim_{N\to\infty} f(s_\theta, v) = \int_0^{\pi-2\pi(s_\theta - v)} \frac{\sin t}{\pi t}dt + \int_0^{\pi+2\pi(s_\theta - v)} \frac{\sin t}{\pi t}dt$$

となる。この積分区間をM等分し、区間ごとの帯の足しあわせとすることで、pvについて以下の様な近似を得る。

$$p_v \approx \sum_\theta z_\theta \sum_{m=0}^{M-1} g(s_\theta, v-\frac{1}{2}+\frac{m}{M})/NM, \quad g(s_\theta, v) = \cos\left[\pi\frac{(s_\theta - v)(N-1)}{N}\right]\frac{\sin\left[\pi(s_\theta - v)\right]}{\sin\left[\pi\frac{s_\theta - v}{N}\right]}$$

加えて詳細については後述するが、実際には上式に対し更に窓関数のサイドローブの影響を除去するような補正を行う。

分布の計算にこの窓関数を採用すれば、多項式時間の分布計算アルゴリズムを構成することができる。 具体的な計算量については、動的計画行列を埋めていく際の特徴量ゲインの計算量が十分小さければ、例えば配列アラインメント問題であれば配列長nに対してN並列可能なO(n^2N)時間+NM並列可能なO(N^2M)時間、RNA二次構造問題であれば配列長nに対してN並列可能なO(n^3N)時間+NM並列可能なO(N^2M)時間となる。

RESULTS

提案手法が期待どおりに動くことを確かめるため、RNA二次構造の自由エネルギーを特徴量とし、その分布を計算するアルゴリズムを実装した。その結果、理論通りの時間・空間計算量によって目的の分布を計算することが可能であるということが確かめられた。

FUTURE WORK

1. サイドローブの影響を最小化するための補正法の検討

現在は単純に分布全体に一律のバイアスが掛かっていると見做し、正規化するというような補正を行っている。 しかしこの方法では実はメインローブ付近で窓関数が落ち込むことによる悪影響を無視できない可能性がある。そこで、ビンを要求解像度より細かめに取って計算した後スムージングのような操作を施すことで、悪影響を軽減する工夫を検討している。

2.より有用な応用先の検討

実装の容易さから今回RNA二次構造のエネルギーの分布を計算するアルゴリズムを構築したが、現在はp値の分布や予測クロマチン状態のクレジビリティリミット計算といったより有用と考えられる応用先の検討及び実装を行っている。

REFERENCES

[1] Do et al. Bioinformatics, 2006 [2] Hung et al. J. Comput. Biol., 2006

[3] Newberg et al. J. Comput. Biol., 2009 [4] Mori et al, BMC genomics, 2014