数理解析研究所講究録
第1799巻 2012年 60-64

海草全単射の漸減構築

酒井義文*

概要

すべての整数対 (i,j) を頂点とし，すべての頂点 (i,j) に対する $(i-1,j)$，(i,j) 間および $(i,j-1)$，(i,j) 間と，いくつかの頂点 (i,j) に対する $(i-1,j-1)$，(i,j) 間にのみ辺をもつ格子グラフにおいて，$e\geq g$ かつ $f\leq h$ である頂点 (e,f) から頂点 (g,h) への交差のない任意の2つの最短経路 P，Q によって挟まれた領域として定義される部分格子グラフを G とする．海藻全単射は，そのような G に対して P の辺の集合から Q の辺の集合への全単射として定義され，P，Q 上の任意の2つの頂点の間の G における最短経路がもつ辺の個数を線形時間で求めることに利用できることが知られている．G に単位格子を一つ追加することで得られる部分格子グラフに対する海藻全単射は，G の海藻全単射から定数時間で容易に求めることができる．本稿では，G から単位格子を一つ取り除くことで得られる部分格子グラフに対する海藻全単射を G の海藻全単射から求める方法について考える．

1 はじめに

すべての整数対 (i,j) を頂点とし，すべての頂点 (i,j) に対する $(i-1,j)$，(i,j) 間および $(i,j-1)$，(i,j) 間と，いくつかの頂点 (i,j) に対する $(i-1,j-1)$，(i,j) 間にのみ辺をもつ格子グラフにおける2頂点間の最短経路の長さ（その経路上の辺の個数）を求める問題は，文字列比較において多くの応用をもつ．たとえば，A，B を長さがそれぞれ m，n の文字列とすると，A の i 番目の文字と B の j 番目の文字が等しいすべての頂点 (i,j) に対してのみ $(i-1,j-1)$，(i,j) 間の辺（以降，斜めの辺という）をもつ格子グラフ（以降，A と B の一致記号対グラフという）において，頂点 $(0,0)$ から頂点 (m,n) への最短経路がもつ辺の個数が l ならば，A と B の最長共通部分列 (longest common subsequence) の長さは，$m+n-l$ である．同様に，A と B の部分文字列 (substring) の最長共通部分列も，部分文字列の開始位置により定まる頂点から終了位置により定まる頂点への最短経路がもつ辺の個数から求まる．また，ギャップ開始ペナルティを伴わない整列 (alignment) の最大スコアについても，記号対のスコア設定に基づいて A と B により定まる適切な格子グラフを用いることで同様に得ることができる [2].

Tiskin[3] は，Alves ら [1] による A の各部分文字列と B 全体の最長共通部分列の長さの間に成立する性質に関する観察を一般化することによって，$m+n$ 個の添え字対が，一方の文字列の任意の部分文字列ともう一方の文字列全体や，一方の文字列の任意の接頭辞ともう一方の任意の接尾辞の最長共通部分列の長さを簡潔に表現することを示し，文字列比較に関する問題を高速に解く数多くのアルゴリズムを提案した．この $m+n$ 個の添え字対は，A と B の一致記号対グラフにおいて，頂点 $(m,0)$ から $(0,0)$ を経由して $(0,n)$ へと至る最短経路 P 上の辺の集合から，頂点 $(m,0)$ から (m,n) を経由して $(0,n)$ へと至る最短経路 Q 上の辺の集合への全単射として表すことができ，P，Q に挟まれる領域の斜めの辺以外のすべての辺からなる集合に対して海草図とよばれる $m+n$ 個の部分集合による分割を求めることで得られる．この全単斜を P，Q に挟まれる領域に対する海草全単射とよぶことにする．

*東北大学大学院農学研究科

海草全単射の定義を，斜めの辺が存在する位置に関して制約のない一般の格子グラフにおける頂点 $(m,0)$ から頂点 $(0,n)$ への交差のない任意の2つの最短経路 P，Q によって挟まれた領域として与えられる部分格子グラフ（以降，右上がり領域グラフという）G に対して拡張すると，G の海草全単射を P，Q 上の任意の2つの頂点間の最短経路がもつ辺の個数を線形時間で求めることに利用できる [2].

右上がり領域グラフ G に単位格子を一つ追加することで得られる右上がり領域グラフに対する海藻全単射は，G の海藻全単射のみから定数時間で容易に求めることができる．本稿では，G から単位格子を一つ取り除くことで得られる右上がり領域グラフに対する海藻全単射を，G の海藻全単射から定数時間で求めるためには，取り除かれる単位格子グラフと対をなすある単位格子が G に含まれるか否かに関する情報を用いれば十分であることを示す．また，右上がり領域グラフに逐次的に単位格子を追加したり削除する過程において，ある単位格子を追加する際に，それと対をなす単位格子の位置を $O(\log(m+n))$ 時間で決定する $O(mn)$ 領域アルゴリズムを提案する．

2 準備

すべての整数対 (i,j) を頂点とし，すべての頂点 (i,j) に対する $(i-1,j)$，(i,j) 間および $(i,j-1)$，(i,j) 間と，いくつかの頂点 (i,j) に対する $(i-1,j-1)$，(i,j) 間にのみ辺をもつ格子グラフを考える．$(i-1,j)$，(i,j) 間，$(i,j-1)$，(i,j) 間，$(i-1,j-1)$，(i,j) 間の辺をそれぞれ，縦の辺，横の辺，斜めの辺という．頂点 $(i-1,j-1)$，$(i-1,j)$，$(i,j-1)$，(i,j) によって誘導される部分格子グラフを単位格子とよび，$u(i,j)$ で表す．$(i-1,j-1)$，$(i,j-1)$ 間の辺，$(i-1,j-1)$，$(i-1,j)$ 間の辺，$(i,j-1)$，(i,j) 間の辺，$(i-1,j)$，(i,j) 間の辺および，もし存在するならば，$(i-1,j-1)$，(i,j) 間の辺をそれぞれ，$u^l(i,j)$，$u^t(i,j)$，$u^b(i,j)$，$u^r(i,j)$，$u^d(i,j)$ で表す．$u^d(i,j)$ をもつ $u(i,j)$ を閉単位格子といい，$u^d(i,j)$ をもたない $u(i,j)$ を開単位格子という．

m，n を任意の正整数とする．したがって，$(m,0)$ から $(0,n)$ への最短経路は m 個の縦の辺と n 個の横の辺からなる．P，Q を，第 k 番目の頂点がそれぞれ (e,f)，(g,h) のとき，$e \leq g$ かつ $f \leq h$ であるような $(m,0)$ から $(0,n)$ への任意の最短経路とする．P，Q の第 k 番目の辺をそれぞれ p_k，q_k で表す．したがって，$p_1,\ldots,p_k$ に含まれる横の辺の個数が $q_1,\ldots,q_k$ に含まれる横の辺の個数よりも大きくなることはない．G を，P，Q およびそれらに挟まれる領域からなる部分格子グラフとし，このようなグラフを，P，Q に対する右上がり領域グラフという．

G の海草図は，G のすべての縦の辺と横の辺からなる集合における $m+n$ 個の海草とよばれる列による分割である．P と Q が共有する辺は，それぞれ1個の海草を構成する．一方，任意の海草 S，T とそれらが出会う任意の単位格子 $u(i,j)$ に対して，以下の条件が成立する．ただし，$u^l(i,j)$，$u^t(i,j)$ がそれぞれ S，T の辺であるとき，S，T は $u(i,j)$ において出会うといい，S における $u^l(i,j)$ の次の辺が $u^r(i,j)$ かつ T における $u^t(i,j)$ の次の辺が $u^b(i,j)$ ならば，S，T は $u(i,j)$ において交差するという．

1. S，T が $u(i,j)$ において交差しないならば，S における $u^l(i,j)$ の次の辺は $u^b(i,j)$ であり，T における $u^t(i,j)$ の次の辺は $u^r(i,j)$ である．

2. S，T が $u(i,j)$ において交差するための必要十分条件は，$u(i,j)$ が開単位格子かつ S，T が $u(i,j)$ 以外の単位格子において交差しないことである．

言い換えると，どの海草の対も，開単位格子において一度も出会うことがないか，そうでなければ，ちょうど1個の開単位格子において交差する．一般に，海草の対は複数個の開単位格子において出会い，それらのどの単位格子において交差してもよいため，海草図は必ずしも一つに定まらない．しかし，各海草には先頭の辺と末尾の辺として，P からの辺と Q からの辺がそれぞれちょうど1個ずつ含まれ，P の各辺 p_k に対応する Q の辺 $q_{\beta(k)}$ は，どの海草図を選ぶかに依存することなしに一意に定まる．このような辺 p_k，$q_{\beta(k)}$ によって定義される P 上の辺を表す添

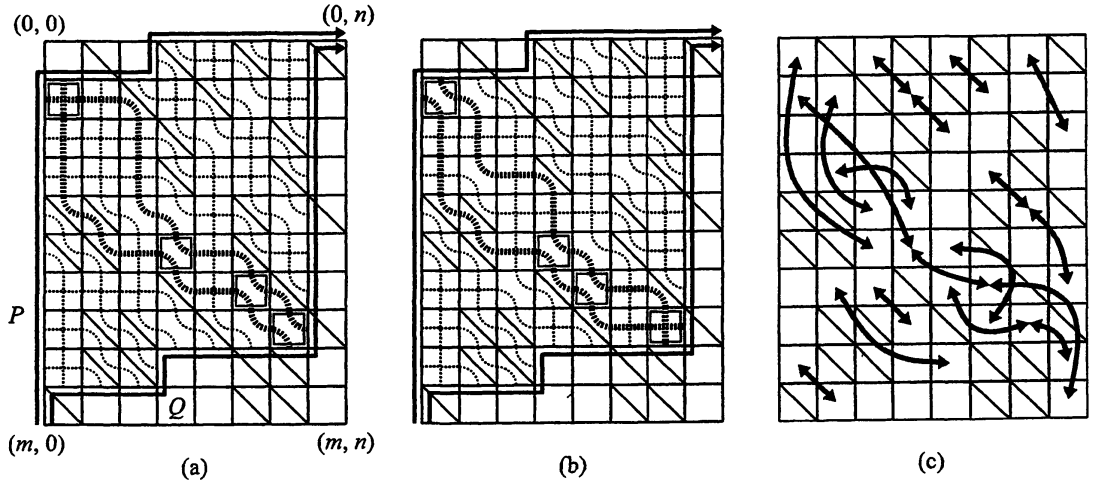

図 1: 格子グラフと，$m = 10$，$n = 8$ の場合の P，Q に対する右上がり領域グラフの例．(a)，(b) はそれぞれ右上がり領域グラフの前交差海藻図と後交差海藻図である．辺を結ぶ各曲線は 2 個以上の辺からなる海藻を表す．(c) は格子グラフにおいて $u(e,f) \longleftrightarrow u(g,h)$ であることを $u(e,f)$ と $u(g,h)$ を結ぶ矢印で表す．

え字の集合から Q 上の辺を表す添え字の集合への全単射 β を，G の海草全単射という．図 1 の (a)，(b) は，同じ右上がり領域グラフの 2 つの海藻図の例である．たとえば，(a)，(b) のどちらの場合も，太い曲線で表される p_9 と q_{10} を結ぶ海藻と p_{10} と q_9 を結ぶ海藻は二重線で示される 4 個の開単位格子において出会うが，交差する開単位格子は互いに異なる．

海藻図において海藻が交差する開単位格子を交差単位格子，そうでない開単位格子を接近単位格子という．ある海藻の対が接近単位格子 $u(i,j)$ と交差単位格子 $u(i',j')$ において出会うとき，$u(i,j)$ が $u(i',j')$ よりも P に近いならば，$u(i,j)$ を前接近単位格子といい，逆に Q に近いならば，$u(i,j)$ を後接近単位格子という．前接近単位格子が存在しない海草図を前交差海草図といい，後接近単位格子が存在しない海草図を後交差海草図という．どちらも一意であることを帰納法によって示すことができる．図 1 の (a)，(b) は，それぞれ前交差海藻図，後交差海藻図である．

海藻図における辺 r をもつ海藻を r 海藻という．$1 \le e \le m$，$1 \le f \le n$，$1 \le g \le m$，$1 \le h \le n$ である任意の開単位格子 $u(e,f)$，$u(g,h)$ に対して，$u(e,f) \longleftrightarrow u(g,h)$ で，辺 $u^l(e,f)$，$u^t(e,f)$ をもつ P と辺 $u^b(g,h)$，$u^r(g,h)$ をもつ Q が存在して，P と Q に対する右上がり領域グラフの前交差海藻図または後交差海藻図において，$u^l(e,f)$ 海藻と $u^t(e,f)$ 海藻が出会う開単位格子が $u(e,f)$ と $u(g,h)$ のみであることを表す．図 1 の (a)，(b) における格子グラフの $u(e,f) \longleftrightarrow u(g,h)$ である開単位格子 $u(e,f)$，$u(g,h)$ の対は (c) に示すとおりである．

3 アルゴリズム

$G_0, G_1, \ldots, G_z$ を，G_0 が単位格子を 1 個ももたず，$y \ge 1$ である G_y が G_{y-1} に $u(i_y, j_y)$ を追加するか G_{y-1} から $u(i_y, j_y)$ を削除することで得られるような任意の右上がり領域グラフの列とする．G_y は P_y，Q_y に対する右上がり領域グラフであるとする．β_y を，G_y の海藻全単射とする．したがって，β_0 は恒等写像である．$\tilde{G}_y$ で，$1 \le x \le y$ かつ G_x が G_{x-1} に $u(i_x, j_x)$ を追加することで得られるすべての x に対する $u(i_x, j_x)$ からなる右上がり領域グラフを表す．以下に，$\tilde{G}_z$ に含まれる $u(e,f) \longleftrightarrow u(g,h)$ である開単位格子 $u(e,f)$，$u(g,h)$ の対をそれぞれ

$O(\log(m+n))$ 時間で求め，これを用いて，逐次的に各 β_{y-1} を β_y に定数時間で更新する $O(mn)$ 領域アルゴリズムを提案する．

アルゴリズムは次の定理に基づく．

定理 1 G を P, Q に対する右上がり領域グラフとし，$u(e,f)$ を任意の開単位格子とする．P が辺 $u^l(e,f)$, $u^t(e,f)$ をもつならば，$u(e,f)$ が前接近単位格子である海藻図を G がもつための必要条件は，$u(e,f) \longleftrightarrow u(g,h)$ である $u(g,h)$ が G に含まれることである．同様に，$u(g,h)$ を任意の開単位格子とすると，Q が辺 $u^b(g,h)$, $u^r(g,h)$ をもつならば，$u(g,h)$ が後接近単位格子である海藻図を G がもつための必要十分条件は，$u(e,f) \longleftrightarrow u(g,h)$ である $u(e,f)$ が G に含まれることである．

証明 辺 $u^l(e,f)$, $u^t(e,f)$ をもつ P_* と辺 $u^b(g,h)$, $u^r(g,h)$ をもつ Q_* に対する右上がり領域グラフ G_* の前交差海草図において，$u^l(e,f)$ 海藻と $u^t(e,f)$ 海藻が出会う開単位格子は $u(e,f)$ と $u(g,h)$ のみであるとする．P は辺 $u^l(e,f)$, $u^t(e,f)$ をもつとする．

はじめに，$u(g,h)$ が G に含まれない場合について示す．G の前交差海草図における $u^l(e,f)$ 海藻，$u^t(e,f)$ 海藻がそれぞれ G_* の前交差海草図における $u^l(e,f)$ 海藻，$u^t(e,f)$ 海藻の接頭辞であることは容易に示せる．したがって，G の前交差海草図における $u^l(e,f)$ 海藻と $u^t(e,f)$ 海藻が出会う開単位格子は交差単位格子である $u(e,f)$ のみである．D を G の任意の海草図とする．G の前交差海草図は，D を海藻図の初期値として，ある海藻の対 S, T が前接近単位格子 $u(i,j)$ において出会うとき，S と T が $u(i,j)$ において交差するように海藻図を変更することを繰り返すことによって得られる．D は逆の手順により G の前交差海草図から得られるから，この過程において $u^l(e,f)$ 海藻と $u^t(e,f)$ 海藻が出会う後接近単位格子が生じないことを示せばよい．

$\mathcal{S}$ を，海草図における $u^l(e,f)$ 海草および P において $u^t(e,f)$ よりも後に現れるすべての辺 p に対する p 海草からなる集合とし，同様に $\mathcal{T}$ を，$u^t(e,f)$ 海草および P において $u^l(e,f)$ よりも前に現れるすべての辺 p に対する p 海草からなる集合とする．$u(i,j)$ が接近単位格子であり，$\mathcal{S}$ 中の海藻 S と $\mathcal{T}$ 中の海藻 T がそれぞれ辺 $u^l(i,j)$, $u^t(i,j)$ をもつとき，S と T は順接近するということにする．この定義より，任意の時刻において，どの $\mathcal{S}$ 中の海藻と $\mathcal{T}$ 中の海藻も順接近しないことを帰納法によって示せば十分である．最初の時刻において，$\mathcal{S}$, $\mathcal{T}$ 中の海藻はすべて前交差海藻図のものであるから，どの $\mathcal{S}$ 中の海藻 S と $\mathcal{T}$ 中の海藻 T も順接近しない．なぜなら，S, T がそれぞれ $u^l(e,f)$ 海藻と $u^t(e,f)$ 海藻ならば，S と T が順接近するためには後接近単位格子で出会わなければならず，一方，そうでないならば，前接近単位格子で出会わなければならないからである．ある時刻において，どの $\mathcal{S}$ 中の海藻と $\mathcal{T}$ 中の海藻も順接近しないと仮定する．次の時刻において，$\mathcal{S}$ 中の海藻 S と $\mathcal{T}$ 中の海藻 T が順接近するためには，$\mathcal{S}$ 中の海藻 S' が存在して，S' は T と前接近単位格子において出会い，かつ，S と後接近単位格子において出会うか，そうでなければ，$\mathcal{T}$ 中の海藻 T' が存在して，T' は S と前接近単位格子において出会い，かつ，T と後接近単位格子において出会う必要がある．しかし，このことは帰納法の仮定に矛盾する．したがって，どの $\mathcal{S}$ 中の海藻と $\mathcal{T}$ 中の海藻も順接近しない．以上より，G の任意の海藻図 D において，$u^l(e,f)$ 海藻と $u^t(e,f)$ 海藻が出会う開単位格子は交差単位格子である $u(e,f)$ のみである

次に，$u(g,h)$ が G に含まれる場合について示す．P と辺 $u^l(g,h)$, $u^t(g,h)$ をもつ Q' に対する右上がり領域グラフ G' が G の部分グラフであるとする．G' の前交差海草図は G_* の前交差海草図と同じ $u^l(e,f)$ 海藻と $u^t(e,f)$ 海藻をもつから，この海藻図において，$u^l(e,f)$ 海草と $u^t(e,f)$ 海草は交差単位格子 $u(e,f)$ と後交差単位格子 $u(g,h)$ でのみ出会う．この海藻図の $u^l(e,f)$ 海草と $u^t(e,f)$ 海草が交差する開単位格子を $u(g,h)$ になるように変更することで得られる海藻図を D' とする．したがって，D' において $u(e,f)$ は前接近単位格子である．D' の各海草を接頭辞とする海草をもつ G の海草図は容易に構築できる．この海草図において $u(e,f)$ は前接近単位格子である．

定理の他の場合についても，同様に示せる．□

提案するアルゴリズムは，それまでに見つかった $u(e,f) \longleftrightarrow u(g,h)$ である開単位格子 $u(e,f)$, $u(g,h)$ の対を，$1 \leq i \leq m$, $1 \leq j \leq n$ である各単位格子 $u(i,j)$ を頂点としてもつグラフ U に $u(e,f)$, $u(g,h)$ を結ぶ辺を追加することによって保持する．また，$\tilde{G}_y$ の前交差海藻図と後交差海藻図および U を用いて，G_{y-1} の海藻全単射 β_{y-1} から G_y の感想全単射 β_y への定数時間更新および $u(e,f) \longleftrightarrow u(g,h)$ である開単位格子 $u(e,f)$, $u(g,h)$ の $O(\log(m+n))$ 時間探索を行う．$\tilde{G}_y$ の前交差海藻図と後交差海藻図の各海藻は，辺 $u^b(i,j)$ または $u^r(i,j)$ を第 $i+j$ 番目の位置にある要素として格納する高さ $O(\log(m+n))$ の二分探索木として表される．$\tilde{G}_0$ の各海藻図における各海藻は 1 個の辺のみからなるため，それを表す二分探索木は $O(\log(m+n))$ 時間で構築できる．

$1 \leq y \leq z$ である各添え字 y の昇順に，アルゴリズムは以下のように動作する．P_{y-1}, P_y がそれぞれ辺 $u^l(i,j)$, $u^t(i,j)$, P_y が辺 $u^b(i,j)$, $u^r(i,j)$ をもつ（G_{y-1} の P 側から単位格子 $u(i,j)$ が削除される）ならば，U が $u(i,j) \longleftrightarrow u(g,h)$ を表す辺をもつか否かを調べる．もしもそのような $u(g,h)$ が G_y に含まれるならば，定理 1 より，β_y は β_{y-1} と等しいため更新する際に変更は必要ない．そうでない場合は，$u^b(i,j)$ が P の第 k 番目の辺であるとき，β_{y-1} の k, $k+1$ に対する値を互いに入れ替える変更のみで β_y が得られる．G_{y-1} の Q 側の単位格子が削除される場合も，同様の方針で β_{y-1} を β_y に更新できる．どちらの場合も $\tilde{G}_y$ は $\tilde{G}_{y-1}$ に等しいから，前交差海藻図と後交差海藻図の更新は必要ない．

P_{y-1}, P_y がそれぞれ辺 $u_r(i,j)$, $u_b(i,j)$, 辺 $u_l(i,j)$, $u^t(i,j)$ をもつ（G_{y-1} の P 側に単位格子 $u(i,j)$ が追加される）場合は，以下のように動作する．$u^b(i,j)$ が P の第 k 番目の辺であるとき，$\beta_{y-1}(k) > \beta_{y-1}(k+1)$ ならば，β_y は β_{y-1} に等しい．そうでないならば，β_{y-1} の k, $k+1$ に対する値を互いに入れ替える変更のみで β_y が得られる．$u(i,j)$ が $\tilde{G}_{y-1}$ に含まれない場合は，以下のように U と海藻図を更新する．$\tilde{G}_{y-1}$ の後交差海藻図は，$u^b(i,j)$ 海藻と $u^r(i,j)$ 海藻の先頭にそれぞれ，$\beta_{y-1}(k) > \beta_{y-1}(k+1)$ ならば $u^l(i,j)$, $u^t(i,j)$，そうでないならば，$u^t(i,j)$, $u^l(i,j)$ を追加することで，$\tilde{G}_y$ の後交差海藻図に更新できる．一方，$\tilde{G}_{y-1}$ の前交差海藻図は，$u^b(i,j)$ 海藻と $u^r(i,j)$ 海藻の先頭にそれぞれ $u^t(i,j)$, $u^l(i,j)$ を追加し，それらが交差する $u(i,j)$ 以外の開単位格子 $u(g,h)$ を $O(\log(m+n))$ 時間で二分探索した後に，もし見つかったならば，$u^r(g,h)$, $u^b(g,h)$ をそれぞれ先頭とする接尾辞を $O(\log(m+n))$ 時間で交換することで，$\tilde{G}_y$ の後交差海藻図に更新できる．また，$u(g,h)$ が見つかった場合は，$u(i,j) \longleftrightarrow u(g,h)$ であるから，$u(i,j)$ と $u(g,h)$ の間に辺を加えることで U を更新する．G_{y-1} の Q 側に単位格子が追加される場合も，同様の方針で β_{y-1}, U および海藻図を更新する．

上に述べた $O(mn)$ 領域アルゴリズムが，$\tilde{G}_z$ に含まれる $u(e,f) \longleftrightarrow u(g,h)$ である開単位格子 $u(e,f)$, $u(g,h)$ の対を逐次的にそれぞれ $O(\log(m+n))$ 時間で求め，この過程において，$y = 1, 2, \ldots, z$ の順に，G_{y-1} の海藻全単射 β_{y-1} を G_y の海藻全単射 β_y に定数時間で更新することは，容易に確認できる．

参考文献

[1] C.E.R. Alves, E.N. Cáceres, S.W. Song, An all-substirng common subsequence algorithm, Electr. Notes Discrete Math., 19 (2005) 133–139.

[2] Y. Sakai, An almost quadratic time algorithm for sparse spliced alignment, Theory Comput. Syst., 48 (2011) 189–210.

[3] A. Tiskin, Semi-local string comparison: Algorithmic techniques and applications, Math. Comput. Sci., 1 (2008) 571–603.