着手記号列の出現頻度に基づく囲碁棋譜からの定型手順獲得

中 村 貞 吾†

囲碁は探索空間が広くまた静的局面評価が難しいため，定石，手筋，石の形といった様々なパターン知識の使用が不可欠となる．これまで，このようなパターンは，定石書などから人手で収集したり，ごく狭い固定された窓の範囲内で形パターンを獲得するといった手法がとられていた．しかし，これらの手法では長さの異なる様々な手順パターンを効果的に収集することは難しいため，棋譜から定型的であると認められる手順を網羅的に獲得する手法の確立が望まれる．本論文では，棋譜を着手ごとの着点が符号化されてできた文字列（棋譜テキスト）であるととらえ，部分文字列の n-gram に基づいて手順の定型性を評価することによって定型手順の獲得を行う手法を示す．本論文で提案する手法は，固定窓のような局面範囲の限定を必要とせず，長さの異なる様々な定型手順を棋譜から直接抽出できるという特長を持つ．そして，日本棋院「棋譜データ集 96」に収録されている全棋譜（約 34,000 局，総手数約 700 万）を対象として行った定型手順抽出実験の結果より，本手法の有効性を検証する．

Acquisition of Move Sequence Patterns from Encoded Strings of Go Moves

Teigo Nakamura†

Move sequence patterns such as Joseki or Tesuji are essential to reduce a lot of search efforts and speed up the search in the game of Go, because Go has an enormous search space and it is difficult to evaluate the board positions precisely. The various sequence patterns with the different length appear in the various stages of the real game. To acquire these patterns, we regard a game record as a string of characters encoding each move into a character and evaluate the degree of patterns for all substrings. We propose a new method for encoding each move and evaluating the degree of patterns based on n-gram statistics. This method can acquire a lot of move sequence patterns of various length and extent from game records. We show the result of acquisition from "Kifu Database 96" which contains about 34,000 games and seven million moves and verify that our method is effective to acquire the sequence patterns.

1. はじめに

囲碁は，チェスや将棋などと比べるとはるかに探索空間が広く，また，静的な局面評価が難しい．その最大の要因は，囲碁では，個々の石に先験的な役割が定まっていないため，プレイヤは石の配置や周囲の状況によって石の集まりを識別して認識する必要があるからである．したがって，限られた時間の中で効率良く局面を探索，評価するためには，定石，手筋，石の形といった様々なパターン知識の使用が不可欠なものとなる．これまで，このようなパターンは，定石書などから人手で収集したり，ごく狭い固定された窓の範囲内で静的な形パターンを獲得したりするといった手法がとられていた[5),11)]．しかし当然のことながら，固定窓を使う手法では，あらかじめ与えた範囲を越えるようなパターンを獲得することはできないし，また，定石などのような長さの異なる手順パターンを効果的に獲得することも難しい．小島らは生態学的アナロジーに基づいたより柔軟なパターン知識の獲得法を提案している[9),10)]が，そこでは主に詰碁などに出現する比較的短手数の手筋の獲得が対象となっている．一般に，定石とは序盤に部分的に出現する一定の石の形およびそこに至る手順を指すが，中盤の定石という類のものもあることから分かるように，序盤に限らず，碁の法則から導かれる一定の理にかなった着手の応酬というものが存在する．そこで，手順の長短，局面範囲にかかわらず，一局の棋譜を通じて定型的であると認められる手順をすべて獲得することが望まれる．

一局のゲームは，個々の着手の着手位置を記録した棋譜によって記述され，棋譜からは，互いの石の配置などの静的局面情報や着手系列によって局面がどのよ

† 九州工業大学
Kyushu Institute of Technology

うに変化していったかなど，ゲームの進行に関する情報をすべて再現することができる．また，囲碁は別名「手談」ともいわれるように，その対局は着手を通じたコミュニケーションであると見なすことができ，個々の着手にはこれ以外にもその局面における役割やプレイヤの意図などの様々な情報が内在されていると考えられる．そこで本論文では，棋譜を個々の着手を符号化してできたテキスト（棋譜テキスト）であると見なし，この棋譜テキストに対して自然言語処理の分野で行われている n-gram 統計に基づいた定型表現の抽出法を用いて，このような長さの異なる定型手順の獲得を行う手法を示す．

以下では，まず 2 章で棋譜からの定型手順獲得の概略を述べた後，2.1 節で着手をテキスト化するための符号化について説明する．次に，2.2 節で文字列の定型性を評価するための手法をいくつか紹介し，本論文で提案する定型性評価法である部分列頻度プロファイル法について述べる．そして最後に 3 章でこれらの手法を適用して実際に棋譜データベースから定型手順獲得の実験を行った結果を示し，ここで提案した手法の有効性を検証する．

2. 定型手順獲得法

ゲームの法則から導かれる理にかなった一定の着手系列である定型手順は，過去の棋譜中の様々な場所に頻繁に出現している．囲碁におけるこのような定型手順は，自然言語テキストにおける定型表現と類似した特徴を持っている．

- 自然言語テキストにおける定型表現
  - 頻繁に使用される
  - ひとまとまりの表現
  - 単語よりも大きい単位
- 定型手順（定石）
  - 頻出
  - 単位性のある連続着手
  - ある程度の長手順

日本語のように単語の間に空白を置かずにべた書きされるテキストデータの場合，単語やそれが連なった定型的な表現を抽出することを目的として，辞書と文法による形態素解析を行わずに，文字列の出現頻度情報のみを用いて定型表現を抽出する様々な手法の研究が行われており[1)~4),6)~8)]，これらの手法は，棋譜テキストからの定型表現獲得にただちに応用することが可能である．

2.1 着手の符号化

n-gram 統計ではパターンが一致するかどうかは文字列の一致性によって判断される．したがって，文字列の出現頻度に基づく手順パターンの定型性評価が正しく行われるために，着手の符号化において以下のことが要請される．

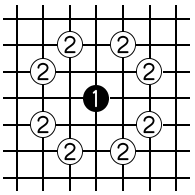


図 1 同一の位置関係にある着手の例
Fig. 1 Moves with the identical positional relationship.

（1） 個々の着手に対して，時間的，空間的に局所的な情報のみを符号化する．

（2） 盤上での回転，鏡像，移動の関係にある手順が同一の符号列になる．

（3） 形の異なる手順は，同一の符号列にならない．

本論文では，連続した一連の着手のみを定型手順としての獲得対象としている．要請 (1) は周辺配石などの余分な情報を排除し符号数を増大させないことを意図した要請である．また，囲碁では駒の進む方向というものがないため，着手間の関係は盤上の相対的な位置関係によって規定されることから，要請 (2)，(3) が導かれる．

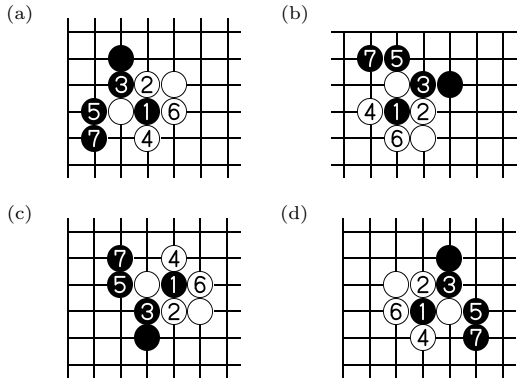
図 1 は「ケイマ」と呼ばれる位置関係にある着手の例である．これら 8 つの着手にはいずれも同じ符号が与えられることが要請され，そのためには以下の符号化法が考えられる．

単純符号化法（E_S） 現在の着手 m の着点を (x, y)，直前の相手着手の着点を (x_p, y_p) とするとき，m の着手符号 $c_{x',y'}$ を以下のようにして定める．

$$\begin{cases} x' = \min(|x - x_p|, |y - y_p|), \\ y' = \max(|x - x_p|, |y - y_p|). \end{cases}$$

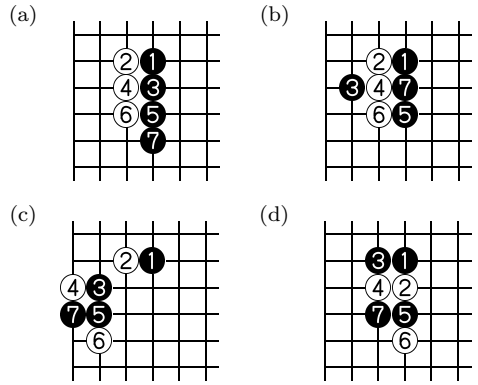
たとえば，現在の着点が $(7, 5)$ で，直前の相手方の着点が $(5, 4)$ である場合には，符号 $c_{1,2}$ を与える．これにより，**図 2** に示すように，回転，鏡像，移動の関係にある着手列を同一の符号列に変換することができる．しかし一方で，この符号化法は一連の着手のなす形状が異なるパターンをも同一の符号列に変換してしまう（**図 3**）ため要請 (3) を満足しない．

そこで本論文では，要請 (1)–(3) を満足する符号化法として単純符号化法を改良した以下に示す符号化を行う．この符号化法は，棋譜中の各着手に対して直前の相手方の着手との相対的な位置の差分をもとにした符号化を行うという点に関しては上記の単純符号化



手順 (a)〜(d) はいずれも同一の符号列
“$c_{0,1}\ c_{0,1}\ c_{1,2}\ c_{1,2}\ c_{0,3}\ c_{1,3}$” になる．

図 2 回転，鏡像，移動に関する一致性
Fig. 2 Identity in terms of rotation, mirroring and moving.

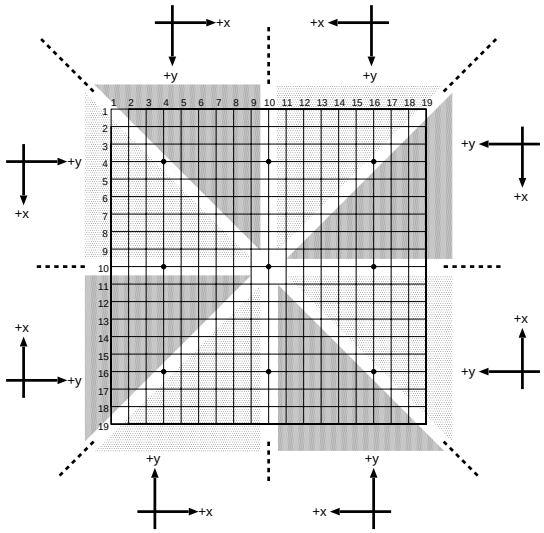


形の異なる手順 (a)〜(d) はいずれも同一の
符号列 “$c_{0,1}\ c_{1,1}\ c_{0,1}\ c_{1,1}\ c_{0,1}\ c_{1,1}$” になる．

図 3 単純符号化法の問題点
Fig. 3 Problem of simple coding method.

法と同様である．しかし，単純符号化法において座標軸の向きと順序が現在の着手との相対的な関係に基づいて決定されていたのに対して，新しい符号化法では，直前の着手の絶対座標に基づいて座標軸の向きと順序の決定を行うという点が異なっている．これは，手順を構成する各着手がなす形状に応じて着手符号化における座標軸の選択を固定する効果があり，それによって要請 (3) が満足される．

相対符号化法（E_R） 直前の相手方の着点を座標原点とする．座標軸（x 軸，y 軸）の選択と正負の方向は，直前の着点が**図 4** で示した 8 つの領域のどれに属するかに応じて決定する．直前の着点が領域の境界上に位置する場合，境界面で接する 2 つの領域の座標軸間で，x 軸，y 軸の選択および



直前の着点が属する領域に付随する座標軸を採用する．

図 4 直前の着点と座標軸との対応
Fig. 4 Correspondence between the posision of previous move and the axis.

軸の向きのうちで一致しているものについてはそれを採用し，一致していないため一意に決定できないものについては，直前の符号化で使用した座標軸のものを継承することとする．そして得られた座標軸を用いて現在の着点の相対位置 (x, y) を求め，符号 $c_{x,y}$ を与える．なお，初手は近隣の絶隅からの差分とし，他の着手とは異なる符号系列を与えることにする．

たとえば，現在の着点が $(3, 9)$，直前の着点が $(4, 7)$ の場合には，符号 $c_{2,-1}$ を与える．

図 4 にある座標軸は以下のようにして決定している．まず，盤面を対称な 8 つの領域に分割する．そして各領域について，その領域の重心が天元を通る最寄りの線分に向かう方向を x 軸の正方向に，最寄りの対角線に向かう方向を y 軸の正方向とした．

この符号化法を用いて，実際に棋譜データベース中の棋譜の符号化を行った．棋譜データベースとしては，日本棋院「棋譜データ集 96」CD-ROM に収録されているプロ棋士の対局約 34,000 局，総手数約 700 万手分のデータを用いた．**表 1** は，E_S，E_R の符号化におけるアルファベット数と，符号化された棋譜中に出現する様々な長さの部分文字列の異なり数を集計したものである．E_R では E_S に比べてアルファベット数が約 3 倍に増えているが，頻度 2 以上の異なり部分文字列の数は逆に 17%程度減少している．これは，E_S においては同一符号列となっていた異形手順が E_R に

表 1 符号化法の比較（1）アルファベット数
Table 1 Comparison of each coding method (1): number of alphabets.

符号化法	アルファベット数		部分文字列の異なり数	
	出現数	総数	頻度 2 以上	総数
単純符号化（E_S）	202	244	2.59×10^7	2.89×10^8
相対符号化（E_R）	738	738	2.15×10^7	2.99×10^8

表 2 符号化法の比較（2）同一符号列の異形数
Table 2 Comparison of each coding method (2): number of variants with the identical code string.

符号列長	単純符号化（E_S）における異形数		相対符号化（E_R）における異形数	
	最大	平均	最大	平均
4	61	3.66	6	1.30
5	98	3.62	5	1.13
6	78	3.07	4	1.04
7	44	2.44	4	1.01
8	27	1.89	3	1.00
9	34	1.50	2	1.00
10	24	1.26	2	1.00
15	4	1.06	2	1.00
20	4	1.04	2	1.00
25	3	1.04	2	1.00
30	2	1.03	1	1.00

おいて分割された結果，出現頻度が 1 となる部分列が多数生成されたためであると考えられる．

次に，同一符号に符号化された手順列中に形の異なる手順がどの程度含まれているかを符号長ごとに集計した結果を**表 2** に示す．これによると，E_S では同一符号列であっても実際には形の異なるものが多数あった（たとえば，長さ 5 の符号列において，最大で 98 個，平均でも 3.62 個）のに対して，E_R では同様に長さ 5 の符号列に対して，最大でも 5 個，平均では 1.13 個と異形手順数は著しく減少している．このことから，E_S 符号化によって誤って同一符号になった異形手順の多くは，E_R 符号化を用いることによって正しく分割されるものと考えられる．これによって，真に同一の手順のみが n-gram において正しくカウントされ，結果的に，定型手順獲得の精度は向上する．

2.2 出現頻度に基づく定型性評価

近年，自然言語処理，特に日本語処理の分野では大量のテキストデータから定型的な表現パターンを自動的に獲得する試みがさかんに行われている．その中でも特に辞書と文法による形態素解析を行わずに，文字列の n-gram 統計に基づいて定型表現を抽出する手法は，棋譜テキストからの定型手順獲得における定型性評価にただちに利用することができる．n-gram 統計とは n 個の文字が隣接した文字列がテキスト中にどのような頻度で出現するかを調査したものを指すが，テキストからの定型表現獲得においては，種々の n に対する n-gram を用いて表現単位の定型性を判断する様々な試みが行われており[1)~4),6)~8)]，これらの手法のうち文献 3)，4) で述べられている手法を定型性評価に利用することを試みる．

また，本論文では定型性評価のための新たな手法として部分列頻度プロファイル法を提案し，上記の手法との比較を行う．

2.2.1 正規化頻度法[3)]

文字列 x の対象テキスト中の出現頻度を $f(x)$ で表すものとする．このとき，文字列 x，y に対して $f(x) = f(y)$ であったとしても，$|x| < |y|$（$|x|$ は文字列 x の長さ）であれば y の方が重要であると考えることができる．これは，文字列の長さが長くなるにつれて文字列の可能な種類が増加し，長い文字列の出現頻度分布が総体的に小さくなるためである．中渡瀬は，これを補正するために出現頻度に文字列の長さ n に応じた係数 $\alpha(n)$ を乗じた正規化を行う方法を提案している[3)]．そこでは，対象テキスト中に実際に出現した n-gram の異なり数を $\beta(n)$ として，正規化係数 α と正規化頻度 nf は次のように計算される．

$$\alpha(n) = \sum_{i=1}^{n} \beta(n),$$
$$nf(x) = (f(x) - 1) \cdot \alpha(|x|).$$

そして，この正規化頻度の大きい方から順に文字列を獲得する．

2.2.2 隣接文字エントロピー法[4)]

文字列 x と文字 c に対して，x の直後に c が生起する確率 $P(c|x)$ は次のように求められる．

$$P(c|x) = \frac{f(xc)}{f(x)}.$$

x に後接する文字集合を $C(x)$ とすると，x の後接文字のエントロピー $H_R(x)$ は次式で計算される．

$$H_R(x) = -\sum_{c \in C(x)} P(c|x) \cdot \log P(c|x).$$

$H_R(x)$ は，後接文字の種類が多く出現の度合いが均等であるほど大きくなり，すべての後接文字が等確率で出現するときに最大となる．逆に，後接文字の種類が少なく出現の度合いが偏っているほど $H_R(x)$ は小さくなり，$|C(x)| = 1$ のときに 0 となる．x の前接文字に対するエントロピー $H_L(x)$ も同様にして計算し，$H_R(x)$ と $H_L(x)$ の小さい方の値をエントロピーの有効値 $H(x)$ とする．そして，$H(x)$ の高い順に定型表現として抽出する．

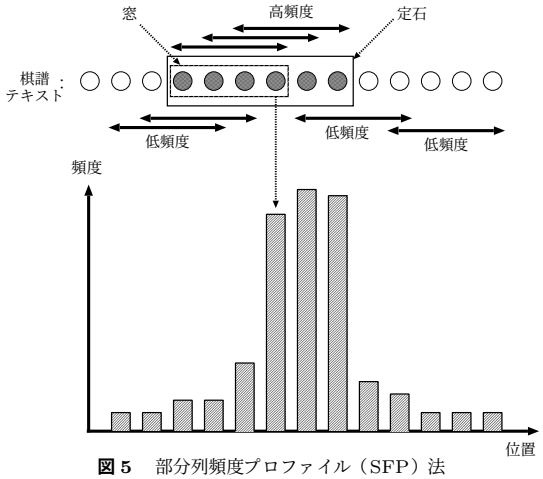


図 5 部分列頻度プロファイル（SFP）法
Fig. 5 Substring Frequency Profiling Method (SFP Method).

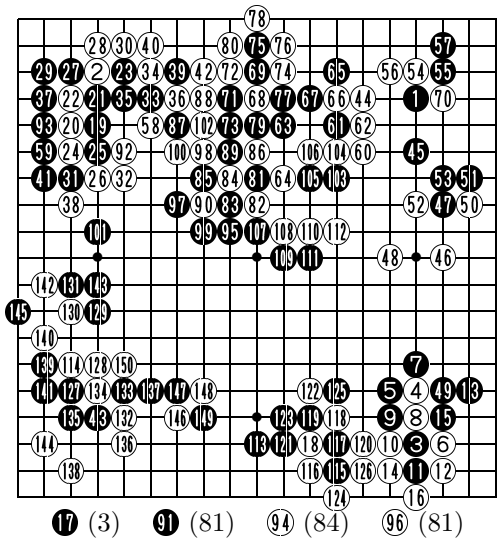


図 6 サンプル棋譜
Fig. 6 Sample game.

2.2.3 部分列頻度プロファイル法（SFP 法）

正規化頻度法は，n-gram を直接用いる方法に比べて断片的な文字列の抽出を避けるような改良がなされているが，基本的にセグメンテーションを行わない手法であるため，それでもなお互いに重なりを持つ文字列や断片的な文字列を抽出してしまうことがある．一方，隣接文字エントロピー法は，単位性の認定を重視して定型手順を獲得する手法であるが，一般に出現頻度の高い短い手順を優先して抽出する傾向にある．そこで本論文では，n-gram 統計を用いて文字列から定型表現を直接切り出すための新しい手法として，部分列頻度プロファイル法（Substring Frequency Profiling Method; SFP 法）を提案する．

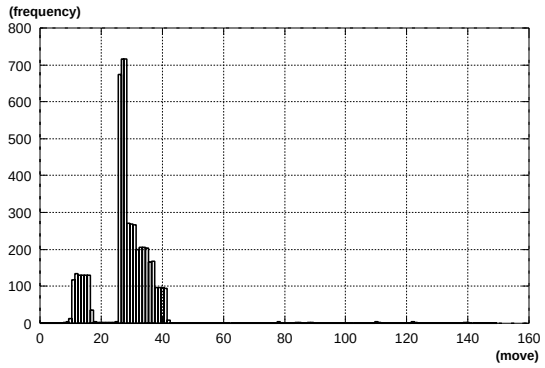


図 7 窓幅 8 の SFP
Fig. 7 SFP with width 8.

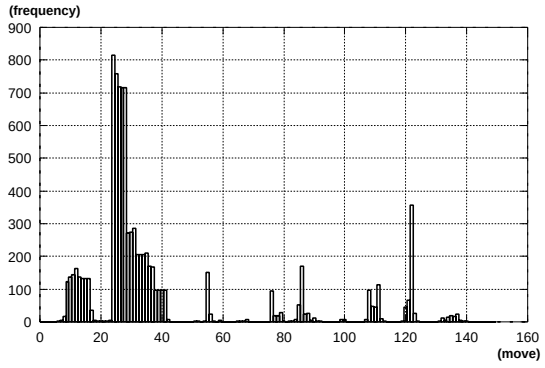


図 8 窓幅 6 の SFP
Fig. 8 SFP with width 6.

SFP とは，注目する部分列の長さ（窓幅）w を固定し，ある棋譜データについて $i-w+1$ 手目から i 手目までの長さ w の着手列がデータベース中に出現した頻度を各 i に対して記録したものである．

複数の n に対する n-gram を用いた場合，ある文字列 x が対象テキスト中に出現する頻度 $f(x)$ と x の部分文字列 y の出現頻度 $f(y)$ の間には $f(x) \leq f(y)$ が成立する．すなわち，ある文字列の部分列は元の文字列よりも出現頻度が高いため，注目している部分列が頻出する基本定石手順の内部にあるときはその出現頻度は高い値をとる．一方，定石は単位性のある連続着手であるため，定石が一段落した境界をまたぐ部分やその外部では出現頻度は低い値となる．したがって，SFP の値は一般に**図 5** に示すような形状をなす．そこで，SFP の値の変化量に適当な閾値を設定して “山” を切り出すことによって高頻度かつ単位性のある手順パターンを獲得することができる．

図 6 の棋譜（日本棋院「棋譜データ集 96」より）に対して，$w=8$ および $w=6$ で作成した SFP を**図 7**，**図 8** に示す．図 6 では，序盤に右下と左上に基本定石

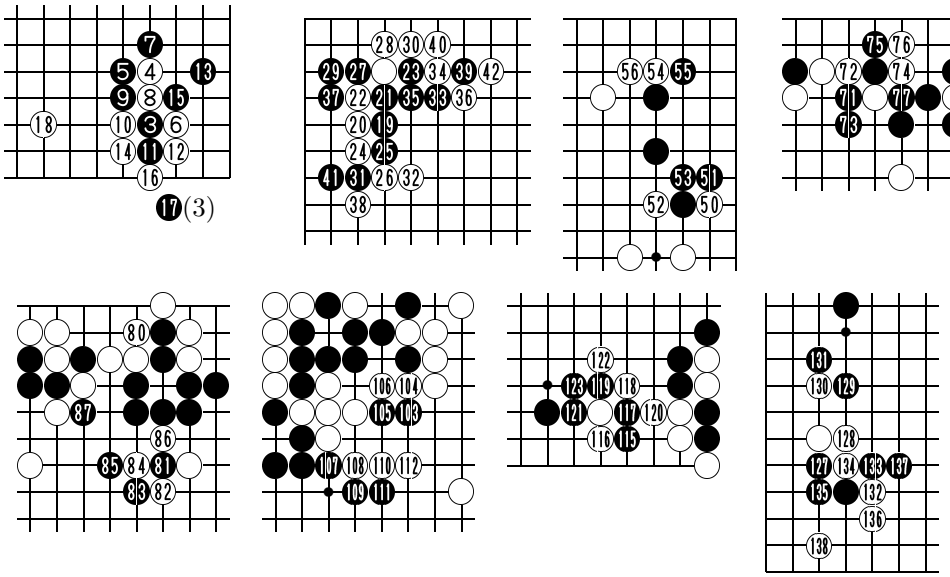


図 9 窓幅 6 の SFP より獲得された定型手順
Fig. 9 Sequence patterns acquired from SFP with width 6.

が出現しており，図 7 の窓幅 8 の SFP では，基本定石部分が明瞭に浮き出てきている．図 8 では，この 2 つの基本定石以外に 6 つの “山” が認められる．これに対応する部分を抽出したものを**図 9** に示す．

2.2.4 多重部分列頻度プロファイル法（MSFP 法）

SFP を用いて定型手順を獲得する際に変化量の閾値を用いる方法の場合，ある定形手順が山として切り出された回数を用いて優先順位付けを行うことになる．このとき，獲得される定型手順数を増加させるためには山を判定する閾値を緩和させてやる必要があるが，山の判定と獲得された手順の優先順位付けは独立しているので，質の良い定型手順を多数獲得するための二値的な閾値の設定は難しい．また，固定した窓幅における SFP では，2 つの独立した定型手順が近接しているとき，次に示すような誤判定の可能性がある．

図 10 の下図は，上図に示す手順に対して作成した窓幅 4 と 5 の SFP である．窓幅 4 の SFP では，位置 5–10 までの範囲が山として認められ，結果的に白 1 〜黒 10 が定型手順であると判断されるが，窓幅 5 の SFP では，山は位置 9–10 であると判断され，結果的に黒 4〜黒 10 が定型手順であると認定される．

この問題を解決するために，SFP 法を以下のように拡張する．まず，部分列に対して山となる度合いを計る指標として次のようなスコアを導入する．対象テキストを $t = c_1c_2c_3 \ldots c_n$ とし，$t[i,j]$ を t の位置 i から j の部分列，すなわち $t[i,j] = c_{i+1} \ldots c_j$ とする．窓幅 w のときの文字列 $t[i,j]$ のスコア $S_w(t,i,j)$ は

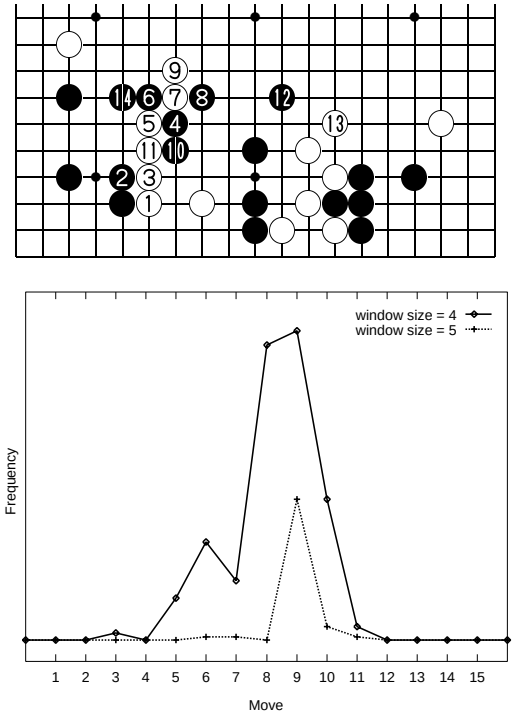


図 10 窓幅による SFP の相違
Fig. 10 Difference of SFPs according to the width of window.

次式で定義される．

$$S_w(t,i,j) = \left(1 - \frac{f(t[i-1,i-1+w])}{f(t[i,i+w])}\right) \times \left(1 - \frac{f(t[j+1-w,j+1])}{f(t[j-w,j])}\right) \times f(t[i,j]).$$

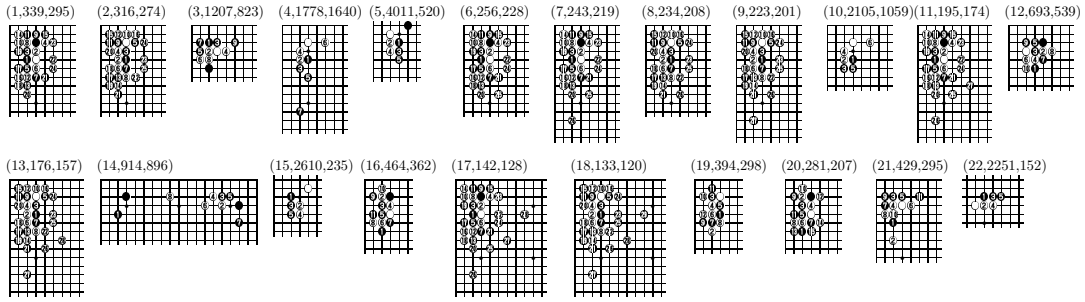


図 11 正規化頻度法による獲得手順
Fig. 11 Acquired sequences by Normalized Frequency Method.

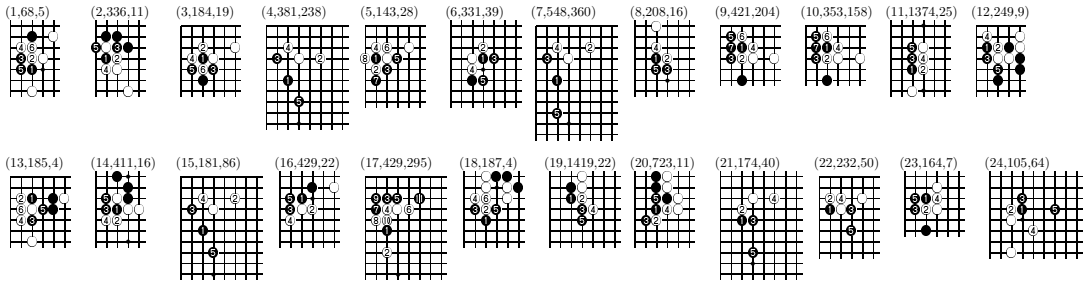


図 12 隣接文字エントロピー法による獲得手順
Fig. 12 Acquired sequences by Entropy Method.

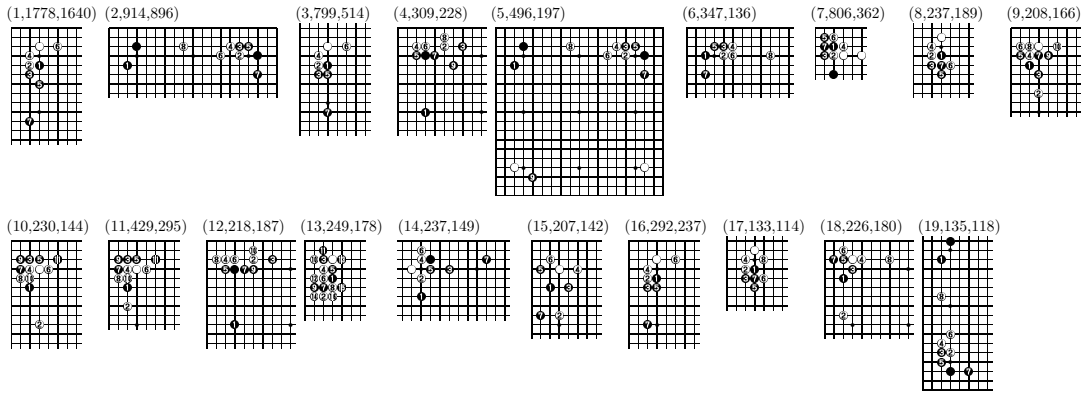


図 13 部分列頻度プロファイル（SFP）法（$w = 6$）による獲得手順
Fig. 13 Acquired sequences by Substring Frequency Profiling Method with $w = 6$.

右辺の第 1 項は山の立上りに対する評価，第 2 項は立下りに対する評価，第 3 項は部分列全体の出現頻度である．S_w の値は，SFP の山が高く，かつ，部分列の境界における頻度の変化量が大きいほど増大するという特徴を持つので，この値を定型性評価の際の優先順位付けに用いることができる．

次に，窓幅 w の違いによる定型性判断の食い違いを吸収するために，考慮の対象とする窓幅集合を $W = \{w_1, w_2, \ldots, w_m\}$ として，各 $w_i \in W$ による SFP のスコアの平均値を最終的なスコア S_W として採用する．

$$S_W(t,i,j) = \frac{1}{|W|} \sum_{w \in W} S_w(t,i,j).$$

この拡張された SFP 法を多重部分列頻度プロファイル法（MSFP 法）と呼ぶことにする．

3. 定型手順獲得実験

日本棋院「棋譜データ集 96」CD-ROM に収録のプ

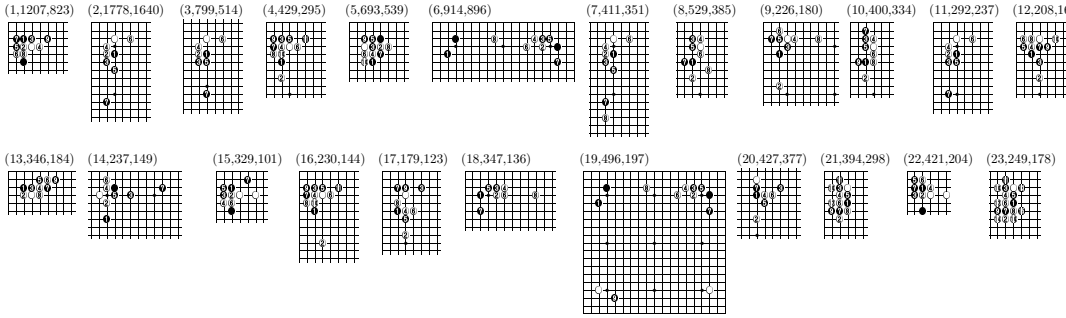


図 14 多重部分列頻度プロファイル（MSFP）法（$w = \{4, 5\}$）による獲得手順
Fig. 14 Acquired sequences by Multiple Substring Frequency Profiling Method with $w = \{4, 5\}$.

表 3 定型手順獲得結果
Table 3 Result of acquiring move sequence patterns.

順位	正規化頻度法			隣接文字エントロピー法			SFP 法（w=6）			MSFP 法（W={4,5}）	
	E_S	E_R	E_R 順位補正	E_S	E_R	E_R 順位補正	E_S	E_R	E_R 順位補正	E_R	E_R 順位補正
1000	19.1	8.4	15.2	1.2	0.0	1.9	29.6	28.6	30.9	28.5	34.5
2000	25.8	15.9	25.1	3.6	0.0	13.3	35.9	34.7	37.4	33.8	40.3
3000	30.2	21.6	30.9	6.5	0.2	17.9	38.6	38.1	40.3	37.0	44.7
4000	33.6	23.8	32.6	11.6	0.2	20.5	39.6	40.3	42.8	42.0	47.6
5000	35.7	26.7	36.0	13.3	0.7	23.2	42.2	42.5	44.2	45.2	49.3
6000	37.9	29.1	37.4	15.0	1.0	26.1	42.5	43.7	45.9	46.9	51.2
7000	39.4	31.0	38.6	16.4	1.4	29.2	43.2	44.7	46.1	49.0	51.7
8000	41.1	33.2	40.3	17.6	1.4	30.9	44.2	45.9	46.1	49.8	52.7
9000	42.0	34.9	40.6	19.1	2.2	34.1	45.1	46.4	46.1	50.2	53.9
10000	43.2	35.8	41.3	19.6	2.9	34.8	45.4	46.4	46.4	50.7	55.3
15000	45.9	38.9	47.1	23.2	5.5	40.3	48.1	46.6	46.9	56.0	58.0
20000	49.5	43.3	50.0	26.3	8.7	43.7	49.5	47.3	47.8	57.0	58.9
25000	52.4	47.1	53.1	29.5	10.3	46.6	—	48.1	—	58.5	59.7

表中の数字は，基本定石の再現率（%）

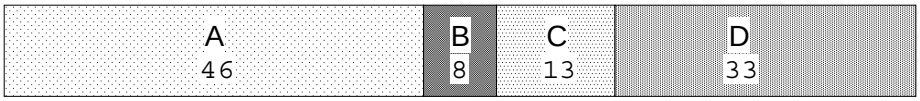
ロ棋士の対局約 34,000 局，総手数約 700 万手分のデータを符号化した棋譜テキストに対して，2 章で述べた 4 つの手法を適用して定型手順の獲得を行った．各手法によって上位に獲得された手順長 5 以上の定型手順を**図 11**〜**図 14** に示す．これらの手法で直接獲得されるものは注目している手順のみであり，そこには周辺の配石は含まれていない．そこで，その手順が適用される際の周辺の配石は棋譜から別途収集する必要がある．今回は，手順中の石をちょうど含む矩形から外側に各々2だけ拡大した矩形の範囲にある配石を収集した☆．なお，図中の（ ）内の数値は，左から順に，順位，着手符号列の頻度，周辺の石を含めた盤面パターンの頻度を表す．

各手法で獲得した個々のパターンを眺めてみると，まず，正規化頻度法では「大ナダレ内マガリ定石」の部分的な手順が多数抽出されていることが目につく．正規化頻度法は長手順でかつ頻出するものに高い優先順位を与えるが，それがまとまった単位として機能しているかどうかの判定をしていないため，互いに重なり合う文字列の組合せが多数抽出される傾向にある．それに対して，その他の 3 つの方法では，単位性の判定が行われているため，正規化頻度法の抽出結果のような違和感はない．

次に，これらの手法によりいわゆる定石手順がどの程度獲得されたかを定量的に比較するために，既存の定石書である「基本定石事典[12]」に代表的な基本定石として掲載されている定石☆☆が獲得できたかどうか，獲得できた場合は第何位で獲得されたかを調査した．その集計結果を**表 3** に示す．

☆ 手順の適用の可否を厳密に判断するための配石を獲得するには，清ら[5]のようにマンハッタン距離を用いるべきかもしれないが，ここでは抽出した手順の表示が主たる目的なので簡単のためこのようにした．

☆☆ 「基本定石事典」には代表的な基本定石に「◆◆」印が付されている．ここでは，「◆◆」が付された定石と，これ以外にも基本定石と思われるものを若干数加えた計 414 個を使用した．



A : 完結した定型手順
B : 定型手順（周囲の状況によっては続けて着手することもある）
C : いわゆる定石の直前または直後の数手を含む定型手順
D : 定石途中（複数の候補手がある分岐点）で終了している

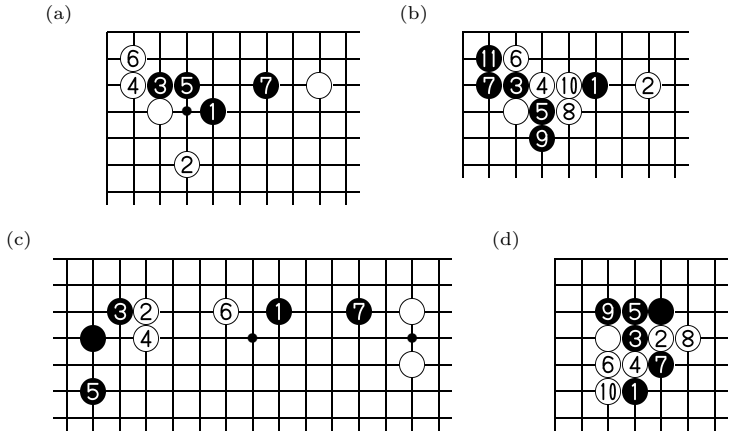


図 15 MSFP 法による獲得手順の分類
Fig. 15 Classification of acquired sequences by MSFP Method.

表中の値は再現率（recall ratio）でこれは次式で計算される．

$$\text{再現率} = \frac{\text{抽出された基本定石数}}{\text{定石事典中の基本定石数}} \times 100(\%).$$

まず，エントロピー法において，E_S と E_R の結果を比較すると E_R の方がかなり再現率が低い結果となった．これは，E_S の符号化では，多数の異形手順が同一符号列となっているが，E_R では異なる形の手順は異なる符号列に符号化されるため，総体的に順位が下がったものと考えられる．そこで，E_S と E_R を正当に比較するために，E_R において獲得された定型手順のうち，E_S 符号化で一致するものについては下位の手順の順位を上位の手順の順位と同一視する補正を行った☆．その結果を「E_R 順位補正」欄に示す．これによれば，E_R 符号化は E_S 符号化よりも基本定石の再現率に関して優れていることが確認できる．また，SFP 法においても，E_R の方に若干の優位性が認められる．さらに，SFP 法と MSFP 法の比較においては，順位が上位のものについては若干 SFP 法の再現率が高いが，順位が下がるにつれて MSFP 法の再現率の方が優位に立っているのが分かる．また，正規化頻度法は再現率に関しては SFP 法と比肩するが，前述のとおり単位性判定に難があるため実用的ではない．

次に，MSFP 法で獲得された定型手順の有効性を評価するために，上位 100 個の手順について単位性および意味のある手順であるかという観点からの分類を行った．その結果を**図 15** のグラフに示す．A の「完結した定型手順」に分類された 46 個の手順の中には，所謂基本定石以外にも (a) に示すような手順も含まれている．また，B，C，D に分類したものは，それぞれ (b)，(c)，(d) のような手順である．(b) は一般的な定石では引き続いて着手されるのが普通であるが，この時点でひとまず落ち着いたと考えることもできるので，着手を継続するかどうかは外部の状況に依存する．したがって，この中で D を除いた 67 個の手順については「単位性のある定型手順」と見なすことができる．なかでも，C に分類された手順は，ある部分手順が適用されるタイミングを教示しており，着手候補生成には有効な知識であると考えられる．D に分類された手順に関しては，局面が一段落しておらず継続して着手する必要があるが，もし継続着手に定型性があるならばその着手は別途獲得されていると考えられるため，これはある定型手順と他の定型手順を合成することができるかどうかの問題に帰着される．棋譜中に 2 つの定型手順の連続着手の実例が存在する場合は合

☆ この順位補正はあくまでも E_S 符号化との比較のためのものであり，E_R 符号化における真の能力は補正なしの値で評価される．

成できることは自明であるが，そうでない場合に合成の可否を判定する手法については今後の課題である．また，この 100 個の手順中,「押し，ノビ，押し，ノビ」のようなあまり価値のない定型手順は 2 例のみで，それ以外はいずれも定石あるいは定形といえるものであった．このことは，基本定石事典[12]に基本定石として掲載されていない定形手順が多数あってそれらが自動的に獲得できたことを意味しており，本手法の有用性を示すものである．

以上の結果より，本論文で提案した MSFP 法が棋譜からの定型手順の獲得に有効であると考えられる．

4. おわりに

棋譜から定石などの定型的な手順を獲得するために，棋譜を着手が符号化されてできたテキストであるととらえ，n-gram 統計に基づいて定型性を評価する手法を示した．まず，着手列をテキスト化するにあたって着手列が構成する石の配置形状の一致性を文字列の一致性で判断することを可能とする着手符号化法として，相対符号化法を提案した．次に，棋譜テキスト上で n-gram に基づいて定型性を評価する手法として，新たに部分列頻度プロファイル法（SFP 法）と多重部分列頻度プロファイル法（MSFP 法）を提案し，これらの定型性評価法が，自然言語テキストからの定型表現獲得で用いられる手法よりも定型手順獲得において優れていることを大量の棋譜データからの定型手順獲得実験を通じて検証した．

参考文献

1) 池原 悟，白井 諭，河岡 司：大規模日本語コーパスからの連鎖型および離散型共起表現の自動抽出法，電子情報通信学会 NLC 研究会報告，Vol.NLC 95-3, pp.17–24 (1995).
2) 長尾 真，森 信介：大規模日本語テキストの n グラム統計の作り方と語句の自動抽出，情報処理学会 NL 研究会報告，Vol.NL 96-1, pp.1–8 (1993).
3) 中渡瀬秀一：統計的手法によるテキストからのキーワード抽出法，電子情報通信学会データ工学研究会報告，Vol.DE 95-2, pp.9–16 (1995).
4) 下畑さより，杉尾俊之，永田淳次：隣接文字の分散値を用いた定型表現の自動抽出，情報処理学会 NL 研究会報告，Vol.NL 110-11, pp.71–78 (1995).
5) 清 愼一，川嶋俊明：「局所パターン」知識主導型の囲碁プログラムの試み，ゲームプログラミングワークショップ，Vol.94, pp.97–104 (1994).
6) 新納浩幸，井佐原均：コーパスからの関係表現の自動抽出，情報処理学会論文誌，Vol.35, No.11, pp.2258–2263 (1994).
7) 新納浩幸，井佐原均：コーパスからの付属語的表現の自動抽出，人工知能学会誌，Vol.10, No.3, pp.429–435 (1995).
8) 新納浩幸，井佐原均：擬似 N グラムを用いた助詞的定型表現の自動抽出，情報処理学会論文誌，Vol.36, No.1, pp.32–40 (1995).
9) Kojima, T. and Yoshikawa, A.: A Two-Step Model of Pattern Acquisition: Application to Tsume-Go, *Computers and Games*, van den Herik, H.J. and Iida, H. (Eds.), Lecture Notes in Computer Science, No.1558, pp.146–166, Springer (1998).
10) 小島琢矢，植田一博，永野三郎：生態学アナロジーを用いた囲碁パターン知識の獲得，ゲームプログラミングワークショップ，Vol.96, pp.133–140 (1996).
11) 斎藤康己，吉川 厚：囲碁プログラムを強くするにはどうしたらよいか？，情報処理学会 AI 研究会報告，Vol.AI 91-7, pp.55–64 (1993).
12) 石田芳夫：基本定石事典（上，下巻），日本棋院 (1975).

(平成 14 年 2 月 20 日受付)
(平成 14 年 9 月 5 日採録)

中村 貞吾（正会員）

昭和 34 年生．昭和 59 年九州大学大学院工学研究科電子工学専攻修士課程修了．昭和 62 年九州大学大学院工学研究科電子工学専攻博士後期課程満期退学．同年九州大学工学部助手．自然言語処理の研究に従事．平成 4 年より九州工業大学情報工学部講師．自然言語処理，ゲームプログラミングに関する研究に従事．工学博士．人工知能学会，電子情報通信学会，Computer Go Forum 各会員．