ゲーム情報学 1−15
(1999. 6. 25)

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, 梶山 貴司

九州工業大学 情報工学部 知能情報工学科
E-mail: {teigo,kajiyama}@dumbo.ai.kyutech.ac.jp

概要

囲碁は探索空間が広く静的局面評価も難しいため，定石，手筋といった様々なパターン知識の使用が不可欠となる．従来，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取られていた．これに対して筆者らは，棋譜を着手毎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，文字列 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なう手法を示した．本論文では，さらに手順獲得の精度を高めるために着手符号化法と定型性評価法を改良し，棋譜データベースからの定型手順獲得実験を通じてその効果を検証する．

Automatic Acquisition of Move Sequence Patterns from Encoded Strings of Go Moves

Teigo NAKAMURA, Takashi KAJIYAMA

Department of Artificial Intelligence, Kyushu Institute of Technology
E-mail: {teigo,kajiyama}@dumbo.ai.kyutech.ac.jp

Abstract

Move sequence patterns such as Joseki and Tesuji can be used to reduce a lot of search space and time. There are many types of Joseki in the game of Go, and their lengths and shapes are various. To acquire these sequence patterns automatically, we proposed a new method based on n-gram statistics in our former article, where we encoded each move into a character and considered a game record as a text string. In this paper, we improve the method of encoding moves and the method to evaluate degree of formulas to acquire move sequence pattern more precisely. Experimental results are shown to compare the former method with improved one.

1 はじめに

囲碁では，チェスや将棋とちがって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が，人間のプレイヤはこれを言葉を用いてネーミングすることで認識し，さらに，着手においても言語的な思考を行なっているということが認知科学的研究により明らかになってきている．また，囲碁は，チェスや将棋と比べるとはるかに探索空間が広く，静的な局面評価も難しいため，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．したがって，このようなパターン知識と人間の言語的思考との間の関係を明らかにすることは，人間の思考形態を模倣する認知科学的なアプローチに基づくコンピュータ囲碁システムの構築において重要な役割をはたすことになる．

一方で，囲碁は別名「手談」とも言われるように，対局は着手を通じたコミュニケーションであると考えられる．通常，一局のゲームは棋譜を用いて記述され，棋譜中には，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うに変化していったかなど，ゲームの進行に関する情報が全て含まれている．これは，個々の着手を符号化してできたテキストであるとみなすことができるので，この「棋譜テキスト」に対して適切な言語モデルを作成することは，上記の認知科学的アプローチに対する一つの方法論となると考える．我々は，この言語モデルが，盤面認識や着手候補選出などゲームプレイシステムの中核として利用できるだけでなく，ゲーム記述言語と自然言語との相互変換を通じて，棋譜からの解説文生成や自然言語による棋譜データベース検索など，知的ゲームに関する自然言語システムへと幅広く応用することができ，ひいては，言語以外の分野における人間の言語的思考過程の解明へと繋がることを期待している．

このような背景の下で，棋譜テキスト中で使用される表現単位となりうる定型的な着手列パターンを適切に設定するために，筆者らは，文献[8]において，符号化された着手記号列の出現頻度に基づく定型手順知識の獲得を試みた．そして，文字列 ***n*-gram** に基づく定型性の評価によって，これまでの固定窓を使う手法では獲得が難しかった長さの異なる様々な定型手順パターンを自動的に獲得できることを示した．しかし，その一方で，使用した着手符号化法の問題のため，定型的とは認め難い手順も多く獲得してしまっていた．

そこで本論文では，より効果的に精度良く定型手順を獲得するために，着手符号化法と定型性の評価法について改良を行なう．そして，その効果を検証するために行なった実験結果について述べる．

2 定型手順獲得法

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わかるように，序盤に限らず碁の法則から導かれる理にかなった一定の着手の応酬というものが存在する．このような定型手順は，過去の棋譜中のさまざまな場所に頻繁に出現しており，一局の棋譜を通じて定型的であると認められる手順を全て知識として獲得することが望まれる．

個々の着手を符号化した棋譜テキストというものを考えた場合，定型手順は，通常の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なわずに，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なわれており[1][2][3] [4][5]，これらの手法は，棋譜テキストからの定型表現獲得に直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なわれるために，着手の符号化において以下のことが要請される．

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．

文献 [8] では，盤面上で回転，鏡像，移動の関係にあるパターンを同一視するために，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おこなっていた．

旧符号化法 (R_o)　現在の着手の着点が (x, y)，直前の相手の着点が (x_p, y_p) であるとき，着手符号 $c_{x',y'}$ は以下のようにして定める．

$$x' = \min(|x - x_p|, |y - y_p|)$$
$$y' = \max(|x - x_p|, |y - y_p|)$$

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．

例えば，現在の着点が (7, 6) で，直前の相手方の着点が (4, 5) である場合には，符号 $c_{1,3}$ を与える．

この符号化法では，図 1に示すように，着手毎に，異なる 8 点を同一の符号に符号化する．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，この符号化法 R_o では，一連の着手のなす形状 (着手の型) が異なるパターンをも同一の符号列に変換してしまう場合がある (図 3)．

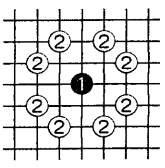
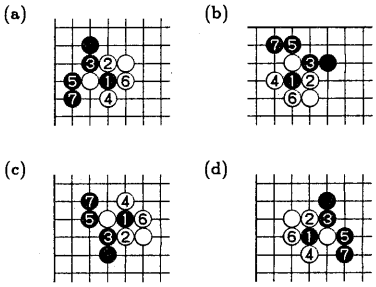
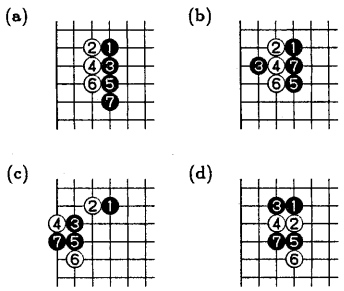


図 1: 同一符号になる着点の例



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



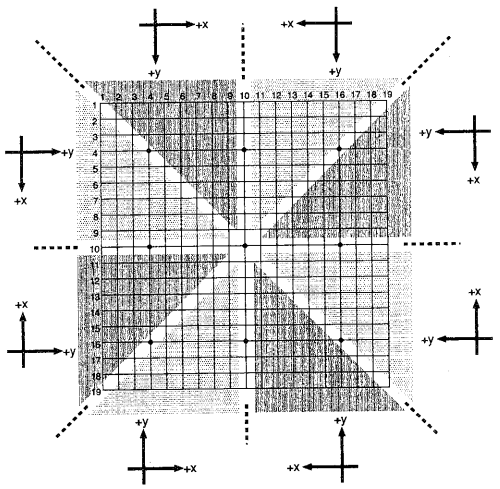
形の異なる手順 (a) ～ (d) はいずれも同一の
符号列 "$c_{0,1}\ c_{1,1}\ c_{0,1}\ c_{1,1}\ c_{0,1}\ c_{1,1}$" になる．

図 3: 従来の符号化法の問題点

囲碁における手筋や定石等のパターンの抽出に適用する場合，前者の特性は少ない棋譜データベースからでも高い出現頻度が得られる利点であるが，後者のそれは定型的でない手順まで抽出してしまう問題点となる．そこで本論文では，従来の符号化法に対し「回転，鏡像，平行の関係にある手順の同一符号列化」と「異形手順は異なる符号列にする」ことの両立をめざした符号化の改良を行なう．新符号化法も，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おこなうという点に関しては従来の符号化法と同様である．しかし，従来の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なうという点が異なっている．

新符号化法 (R_n)　直前の相手方の着点を座標原点とする．直前の着点の位置に応じて x 軸，y 軸の向きと正負の方向を一意に決定する．x 軸，y 軸の向きと正負の方向は，直前の着点が図 4 で示した各領域のどれに属するかに応じて決定する．そし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．

例えば，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．



直前の着点が属する領域の傍らにある座標軸に定める

図 4: 直前の着点と座標軸との対応

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．

これら 2 つ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なった．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，R_o, R_n の符号化におけるアルファベット数と，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．

R_n では R_o 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，R_o においては同一符号列に符号化されていた異形手順が R_n において分割された結果，出現頻度が 1 であるような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毎に集計した．その結果を表 2 に示す．

これによると，R_o 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，例えば，長さ 5 の符号列において，最大で 98 個，平均でも 3.62 個あったのに対して，R_n 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，R_o 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，R_n 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が向上する．

符号化法	アルファベット数		部分文字列の異り数	
	出現数	総数	頻度 2 以上	総数
旧符号化法 (R_o)	202	244	2.59×10^7	2.89×10^8
新符号化法 (R_n)	738	738	2.15×10^7	2.99×10^8

表 1: 符号化法の比較 (1)　アルファベット数

符号長	旧符号化法 (R_o) の異形数		新符号化法 (R_n) の異形数	
	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
11	21	1.15	2	1.00
12	10	1.10	2	1.00
13	7	1.08	2	1.00
14	5	1.06	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

表 2: 符号化法の比較 (2)　同一符号列の異形数

2.2 出現頻度に基づく定型性評価

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．そして，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 を比較することによって表現単位の定型性の判断が行なわれる．以下では，文献 [8] で示した定型性評価の手法のうち，隣接文字エントロピー法，部分列頻度プロファイル法 (SFP 法) について簡単に説明し，最後に，SFP 法を改良した多重部分列頻度プロファイル法 (MSFP 法) の提案を行なう．

2.2.1 隣接文字エントロピー法 [6]

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．

2.2.2 部分列頻度プロファイル法

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かい手順を優先して抽出する傾向にある．そこで筆者らは，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 (Substring Frequency Profile; SFP) 法を考案した．

SFP とは，注目する部分列の長さ (窓幅) n を固定し，ある棋譜データについて $i-n+1$ 手目から i 手目までの n 長さ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り，これは，各着手の近傍における定型性の指標とみなすことができる．

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．

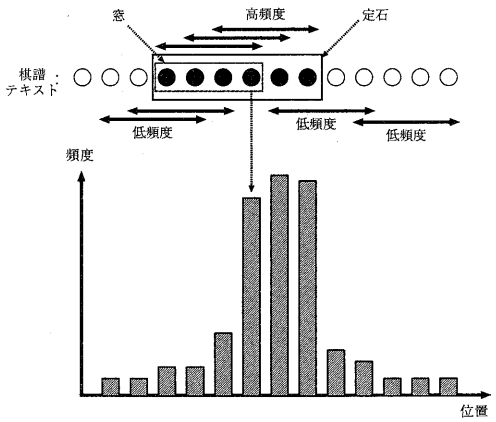


図 5: 部分列頻度プロファイル (SFP) 法

したがって，適当な閾値を設定して SFP における“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．

2.2.3 多重部分列頻度プロファイル法 (MSFP 法)

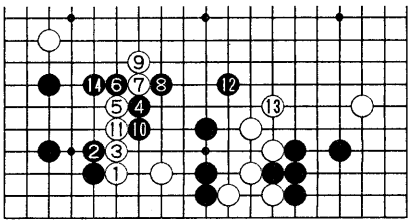
SFP 法では，個々の棋譜から作成した SFP において，部分列頻度の形状を山と認めるか否かの二値的な判断を行ない，そうして得られた手順を集計して出現頻度順に優先順位付けを行なっていた．この手法によって獲得される定型手順数を増加させるためには，山を判定する閾値を緩和させてやる必要があるが，山の判定と獲得された手順の優先順位付けは独立しているので，質の良い定型手順を多数獲得するための二値的な閾値の設定は難しい．そこで，部分列に対して山となる度合を計る指標として次のよ

うなスコアを導入する．いま，窓幅を k とすると，文字列 $c_1c_2c_3\ldots c_n$ の部分列 $c_2c_3\ldots c_{n-1}$ のスコア S_k は次式で計算される．

$$S_k = (1 - \frac{f(c_1\ldots c_k)}{f(c_2\ldots c_{k+1})})(1 - \frac{f(c_{n-k}\ldots c_{n-1})}{f(c_{n-k+1}\ldots c_n)})$$

棋譜中のすべての部分列に対してスコアリングを行ない，それをデータベース中のすべての棋譜に対して集計して定型手順の優先順位付けを行なう．

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．



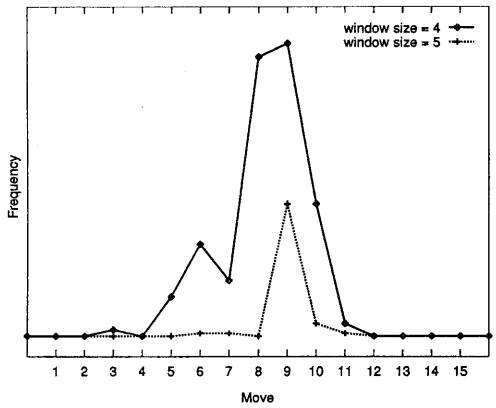


図 6: 窓幅による SFP の相違

図 6 は，棋譜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．一方，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．窓幅 w の違いによるこのような定型性判断の食い違いを吸収するために，MSFP 法では，$W = (w_1, w_2, \ldots, w_m)$ として，各 $w_i \in W$ による SFP のスコアの総和を新しくスコア S_W として用いる．

$$S_W = \sum_{w \in W} S_w$$

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，2.2 で述べた 3 つの手法を適用して定型手順の獲得を行なった．そして，既存の定石書である「基本定石事典 [6]」に掲載されている代表的な基本定石が何位で獲得されたかを集計した結果を表 3 に示す．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，R_o と R_n の結果を比較すると R_n の方がかなり再現率が低い結果となった．これは，R_o 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，R_n 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，総体的に順位が下ったものと考えられる．そこで，R_o と R_n を正当に比較するために，R_n において獲得された定型手順のうち，R_o 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なった．その結果を「R_n 順位補正」欄に示す．これによれば，R_n 符号化は R_o 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，R_n 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，順位が下るにつれて MSFP 法の再現率の方が優位に立っているのがわかる．また，SFP 法では二値的な判断を行なっていたため獲得できる手順数が限られていたが，MSFP 法では必要なところまで任意に獲得数を拡大することが可能となっている．

また，$W = (4, 5)$ の MSFP 法を用いて獲得した定型手順のうち優先順位が上位のものを付録に示す．

順位	隣接文字エントロピー法			SFP 法 (w=6)			MSFP 法 (W=(4,5))	
	R_o	R_n	R_n 順位補正	R_o	R_n	R_n 順位補正	R_n	R_n 順位補正
1000	1.2	0.0	1.9	29.6	28.6	30.9	28.5	34.5
2000	3.6	0.0	13.3	35.9	34.7	37.4	33.8	40.3
3000	6.5	0.2	17.9	38.6	38.1	40.3	37.0	44.7
4000	11.6	0.2	20.5	39.6	40.3	42.8	42.0	47.6
5000	13.3	0.7	23.2	42.2	42.5	44.2	45.2	49.3
6000	15.0	1.0	26.1	42.5	43.7	45.9	46.9	51.2
7000	16.4	1.4	29.2	43.2	44.7	46.1	49.0	51.7
8000	17.6	1.4	30.9	44.2	45.9	46.1	49.8	52.7
9000	19.1	2.2	34.1	45.1	46.4	46.1	50.2	53.9
10000	19.6	2.9	34.8	45.4	46.4	46.4	50.7	55.3
15000	23.2	5.5	40.3	48.1	46.6	46.9	56.0	58.0
20000	26.3	8.7	43.7	49.5	47.3	47.8	57.0	58.9
25000	29.5	10.3	46.6	—	48.1	—	58.5	59.7
30000	32.1	11.8	49.5	—	—	—	58.9	60.9
35000	33.3	13.2	53.4	—	—	—	59.4	60.9
40000	35.7	14.9	53.6	—	—	—	60.1	61.8
45000	37.9	16.3	54.6	—	—	—	60.9	62.8
50000	39.4	18.0	55.8	—	—	—	61.4	62.8

表 3: 定型手順獲得結果

4 おわりに

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文字列の出現頻度情報に基づいて定型手順を獲得する手法において，着手符号化法の改良を行ない，同一符号列になる異形手順数を削減することによって，定型手順獲得の精度向上を行なうことができた．また，部分列頻度プロファイル法 (SFP 法) を拡張した多重部分列頻度プロファイル法 (MSFP 法) を用いることで，SFP 法における閾値設定の問題や固定窓幅による獲得誤りの問題を改善することができた．

参考文献

[1] 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”, 情報処理学会自然言語処理研究会報告 NL 96-1, pp.1–8, 1993.

[2] 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.

[3] 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”, 電子情報通信学会言語理解とコミュニケーション研究会報告 NLC 95-3, pp.17–24, 1995

[4] 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 DE 95-2, pp.9–16, 1995.

[5] 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”, 情報処理学会自然言語処理研究会報告 NL 110-11, pp.71–78, 1995.

[6] 石田芳夫：“基本定石事典”, 上, 下巻, 日本棋院, 1975.

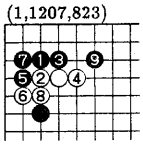
[7] 小島琢矢, 植田一博, 永野三郎：“生態学アナロジーを用いた囲碁パターン知識の獲得”, ゲームプログラミングワークショップ’96, pp.133–140, 1996.

[8] 中村貞吾：“*n*-gram 統計を用いた棋譜データベースからの定型手順の獲得”, ゲームプログラミングワークショップ’97, pp.96–105, 1997.

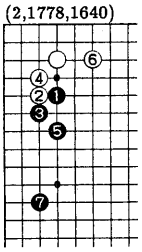
付録　W=(4, 5) の MSFP 法で獲得された定型手順

() 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．

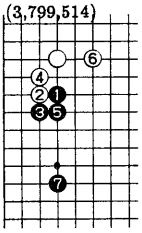
(1,1207,823)



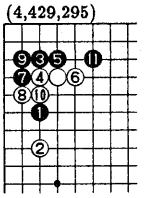
(2,1778,1640)



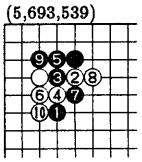
(3,799,514)



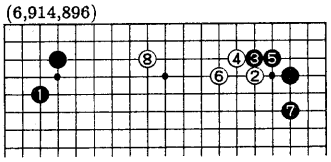
(4,429,295)



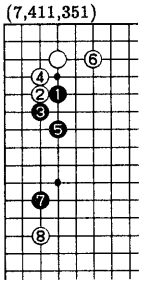
(5,693,539)



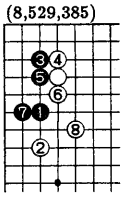
(6,914,896)



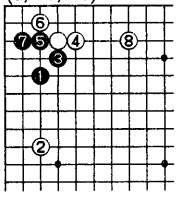
(7,411,351)



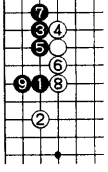
(8,529,385)



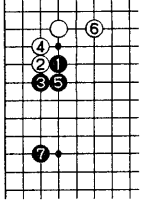
(9,226,180)



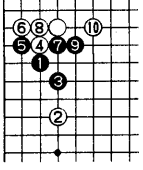
(10,400,334)



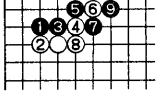
(11,292,237)



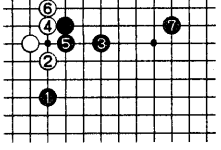
(12,208,166)



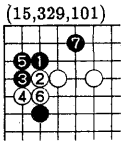
(13,346,184)



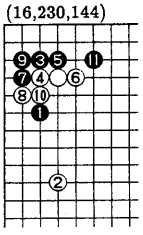
(14,237,149)



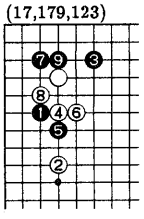
(15,329,101)



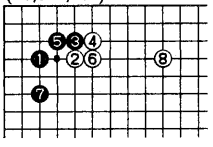
(16,230,144)



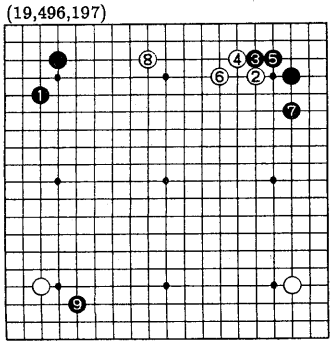
(17,179,123)



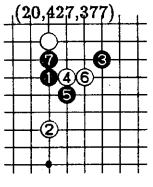
(18,347,136)



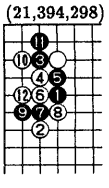
(19,496,197)



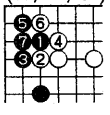
(20,427,377)



(21,394,298)



(22,421,204)



(23,249,178)