TR－H－049

進化システムを用いた遺伝子のコーディング領域予測システムの開発

田中真一（北陸先端科学技術大学院大学）
和田 健之介

1994. 2. 3

ATR人間情報通信研究所

〒619-02 京都府相楽郡精華町光台2-2 ☎07749-5-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

進化システムを用いた遺伝子のコーディング領域予測システムの開発

田中真一
北陸先端科学技術大学院大学 情報科学研究科
(s-tanaka@jaist.ac.jp)

和田健之介
(株) ATR 人間情報通信研究所 第6研究室
(kwada@hip.atr.co.jp)

１９９４年２月３日

1 はじめに

生体の主要な構成要素であるタンパク質は、DNA 上の遺伝情報によってコーディングされているが、実際にタンパク質が作り上げられるには、何段階もの過程が必要である。ヒトなどの高等生物では DNA 上の遺伝情報は、タンパク質へ翻訳される領域であるエクソンと、そうでない領域であるイントロンとに大別される。まず、2本鎖 DNA のうちの主鎖が mRNA に転写される。この時点での mRNA はエクソン、イントロンの両方からなり、mRNA 前駆体と呼ばれる。この mRNA 前駆体から、タンパク質に翻訳されない領域であるイントロンを切り出す作業をスプライシングと呼ぶ。スプライシングを経ることにより、エクソンだけからなる（成熟）mRNA が作られる。mRNA をリボソームが翻訳することによりアミノ酸の鎖が作られ、さらにこの鎖が立体的に折り畳まれてタンパク質が形作られる。

以上の過程のうち、スプライシングが行なわれる機構に我々は注目した。mRNA 前駆体が成熟 mRNA となるためには、スプライシングによってイントロンが除去される必要がある。スプライソゾーム（mRNA 前駆体の 5' スプライス部位, 3' スプライス部位、分岐部位を認識するリボ核タンパク質の集合体）がスプライシングを行なうためには、スプライス部位の正確な識別が行なわれなくてはならない。なぜなら、間違った部位でスプライシングが起こると、それ以降のコドンの読み枠がずれてしまうため、全く違ったタンパク質を形成してしまうからである。

スプライソゾームがスプライス部位の識別のために認識している配列を、何らかの方法で特定することが出来れば、DNA 上におけるコーディング領域の自動検出が可能となる。これは将来、DNA シーケンスのフルオートスキャンが実用化された際、欠かすことのできない技術になると考えられる。

そこで、エクソン・イントロン、イントロン・エクソン間のスプライス部位（前者を 5'

スプライス部位、後者を3' スプライス部位と呼ぶ) を検出するための、ニューラルネットを用いた進化システムを作成した。このシステムを遺伝子データベースDDBJ(DNA Data Bank of Japan) に対して適用した結果についての報告を行なう。

2 ニューラルネットと進化システム

ニューラルネットは学習によって識別能力の向上が可能であるが、ネットワークの構造や学習回数、学習アルゴリズム、各種パラメータ等の最適値をあらかじめ決定するためには非常に困難、かつ多くのコストが必要となる。これらの要素の最適値を自動的に探索するために、ニューラルネットを用いて進化システムを構築した。

個々のニューラルネットを進化システムにおける個体とし、集団を生成する。それぞれの個体に対して突然変異や選択といったGAサイクルを適用することにより、ニューラルネットの構造を進化させ、より能力の高い個体の出現を期待する。学習による結線値の修正といったミクロなチューニングに加え、進化システムによるさらにダイナミックなアーキテクチャの再構築を行なうわけである。

2.1 ニューラルネットの学習

ニューラルネットの学習アルゴリズムとしては、バックプロパゲーション、及びその高速化アルゴリズムを用いる。高速化アルゴリズムについては文献[1]) を参照されたい。

学習用教師サンプルとして、遺伝子データベースから切り出した部分塩基配列を用いる。ニューラルネットを5',3' どちらかのスプライス部位に反応させるために、片方のスプライス部位に注目してサンプルを作成する。今回の場合一つのサンプル単位として、長さ40塩基の部分配列を用いた。配列の中央にスプライス部位を含むものをPOSITIVEサンプル、そうでないものをNEGATIVEサンプルと呼ぶこととし、それぞれのサンプル集団に基づいて、ニューラルネットの評価、学習を行う。

POSITIVEサンプルを入力した場合の教師信号を1、NEGATIVEサンプルの場合を0とする。POSITIVE, NEGATIVEサンプルを交互にネットワークに入力し、ネットワークの出力した値と、サンプルの教師信号との誤差を学習信号として学習に用いる (図1)。POSITIVE,NEGATIVEのサンプル集団すべてについての評価、学習が終了すると1エポック完了する。これを各個体ごとにあらかじめ遺伝的に決められた学習回数だけ繰り返す。

また、学習に時間的コストがかかるとすると、学習回数が多い個体は成熟するまでより多くの時間が必要であると考えられる。逐次成熟した個体について評価、選択を行うことにより、学習回数による淘汰圧が集団中に働き、最適な学習回数の探索が行なわれる。よって、より高い識別能力を、より少ない学習回数で実現するネットワークの出現が期待できる。

2.2 ニューラルネットの構造

ニューラルネットは入力、中間、出力層の3層からなる。入力層にはATGC各塩基に特異的に反応するユニットがそれぞれ、サンプルの部分配列の長さだけ並ぶ。各入力ユニットは中間層のすべてのユニットと結合し、中間層のすべてのユニットは出力層の一つのユニットと結合する。

また、各ネットワークはシグモイド関数の勾配を決めるための温度パラメータを持ち、各ユニットの出力は次の関数fにより求まる。

$$f(u) = \frac{1}{1 + e^{-\frac{u}{T}}} \tag{1}$$

ここでTは温度をあらわし、温度パラメータも突然変異の対象となる。この突然変異により、ネットワーク出力の分離特性の向上が期待できる。

各入力ユニットにおいて、反応すべき塩基が存在すれば1、存在しなければ0を入力する。この入力パターンに応じて中間層、出力層の状態を変化させ、ネットワークの出力値を求める。

ネットワークによるスプライス部位の識別は次のように行なう。閾値を0.5として、ネットワークの出力値が閾値より大きい値ならば、入力されたサンプルは中央にスプライス部位を持つ部分配列であり、閾値以下ならば中央にスプライス部位は持たない部分配列であるとする。各ネットワークごとにそれぞれ、出力値と教師信号による誤差信号を用いて学習を行なう。

この問題においては初期値依存性が強いため、個体生成時に与えられたネットワーク構造では、学習だけによる正答率の向上には限界がある。このため、突然変異によってネットワークの初期構造を変化させる。まず、ユニット間の結線値に対して突然変異（図2）が施される。この突然変異により、ニューラルネットの局所解からの脱出が期待出来る。加えて挿入、欠失突然変異（図3,4）により中間層のユニットの数が増減し、最適なユニット数の探索が行われる。

各突然変異についての詳細は後述する。

2.3 GAサイクル

各個体に対して評価、選択、突然変異が施され、優れた適応度を持つ個体が次の世代により多くの子孫を残すことが出来る。このサイクルを繰り返すことにより、より環境に適応した個体、つまり、よりスプライス部位の識別能力の高いネットワークの出現が期待出来る。

2.3.1 ネットワークに対する評価、選択

各ネットワークはそれぞれ、POSITIVE,NEGATIVEサンプル集団に対してどれだけ正しくスプライス部位の有無を認識したかという評価値を持つ（図5）。評価値は、正しい認識をした回数をそれぞれのサンプル集団についてカウントし、正規化を行なったものである。

$$R = (\frac{P_c}{P_s} + \frac{N_c}{N_s}) \,/\, 2 \tag{2}$$

Rはネットワークの評価値、P_s, N_sはPOSITIVE, NEGATIVEサンプルそれぞれの総数、P_c, N_cは両サンプルに対してネットワークが正しくスプライス部位の識別を行なった回数である。

この評価値が、ネットワークの持つ純粋な識別能力にあたる。以下の式を用いて、i番目の個体（ネットワーク）の集団中での適応度を計算する。

$$R^i_{fit} = f^i / \sum_j f^j \tag{3}$$

$$f^i = P \times (R^i - R_{min}) + R_{min} \tag{4}$$

R^i_{fit} がi番目の個体の集団中での適応度、P が淘汰強度、R^{min} は集団中の評価値の最小値である。淘汰強度パラメータを導入することにより、個体ごとの評価値の違いを強調することが出来る。

適応度 R_{fit} の個体が選択されるかどうかは、以下の条件式によって決まる。

$$R_{fit} - V_{rand} \begin{cases} >= 0, & \text{selected} \\ < 0, & \text{not selected} \end{cases} \tag{5}$$

ここで、V_{rand} は0から1の一様乱数である。

ただし、優れた個体が必ずしも次世代に子孫を残せるわけではないという点に注意する必要がある。選択はあくまで適応度に応じて確率的に行なわれるために、優れた適応度を持つ個体が常に選択されるわけではない。これは集団遺伝学における遺伝的浮動に相当し、局所解に陥った個体の形質で集団中が満たされることを防ぐことが出来る場合もある。エリート保存の戦略を組み合わせることにより、現在の集団内で最も優れた形質を持つ個体を、常に次世代に残すことにする。こうして、優れた形質を残しつつ、なおかつ集団の多様性を保持することが可能となる（図6）。

2.3.2 突然変異

突然変異には以下の4通りがある。

- 結線値に対する突然変異
- 挿入突然変異
- 欠失突然変異
- 各パラメータに対する突然変異

結線値に対する突然変異

中間層の各ユニットにおける全ての結線値に対して、等確率で結線値の増減が起こる（図2）。この突然変異はネットワークが局所解に陥った場合に有効である。特に、中間層のユニット数が少ない場合には初期値依存性がかなり強く現れるので、このような場合に局所解からの脱出に有効に働く。

挿入突然変異

中間層の各ユニットに対して、等確率でユニットの挿入が起こる。挿入が起きたユニットについては、そのユニットのコピーがもう一つ作られ、結線値が分割される（図3）。ただし、このとき結線値を完全に二等分してしまうと、元のユニットと新たに挿入されたユニットにおける出力値、学習による結線値の変更値が同一になってしまい、以降の二つの

ユニットの挙動が全く同じになってしまう。そのため結線値を分割する比率をパラメータによって決定する。

このため、分割前と同一の機能を保持しつつ、学習を行なうごとにそれぞれのユニットが異なった働きをするようになる。

欠失突然変異

中間層の各ユニットに対して、より使われていないユニットほど欠失の対象となりやすい。欠失が起きたユニットは、ネットワーク中からそのまま除かれる。この欠失によって、ネットワークの機能を大きく損なうことなく、より少ない素子数でネットワークを実現できる期待もある（図4）。

各パラメータに対する突然変異

各個体は、学習アルゴリズムごとに持つ特定の定数の集合や、学習アルゴリズムを選択するフラグを持つ。また、学習回数、シグモイド関数における温度パラメータを持つ。それぞれのパラメータに対して確率的に突然変異が発生する。

これらの突然変異により、元のネットワークより優れた識別能力を持つネットワークの生成が期待できる。特に、挿入、欠失の突然変異によって中間層のユニット数を増減し、ネットワークの構造をダイナミックに変化させている。3層ニューラルネットワークにおいては、中間層のユニット数はネットワークの汎化能力に大きな影響を与えるため、必要かつ最小限であることが望まれる。しかし前もって最適なネットワークの構造を決定することは困難であるため、進化システムと組み合わせて優れたネットワークの生成を行っている。

3　遺伝子データベースに対する適用

以上のようなシステムを用いて、POSITIVE, NEGATIVE それぞれ 1000 サンプルに対して適用してみたところ、識別率 100% のネットワークが得られた（数量化理論との比較については文献 [2]) を参照）。

この識別率 100% のネットワークを用いて、遺伝子データベースの塩基配列から直接スプライス部位の検出ができるかどうかについて実験を行った（図7）。

具体的には、遺伝子データベース DDBJ(DNA Data Bank of Japan) の塩基配列情報、ロケーション情報を用いて、ネットワークの能力の評価を行なう。エクソン、イントロン等のロケーション情報は、塩基配列情報フィールドの先頭から数えて何塩基から何塩基目までがエクソンであるといったように、各 LOCUS ごとに FEATURES テーブルに記述されている。

各 LOCUS における塩基配列情報フィールドから、先頭から順に一塩基ずつずらして切り出した部分配列をネットワークに入力する。イメージとしては、ネットワークの入力層が DNA 上をスイープすることによって、部分塩基配列を入力しているものと考えられる。入力した部分配列の中央部においてのスプライス部位の有無と、ネットワークの出力とを比較して正誤判定を行なう。

3.1 途中結果及びシステムの拡張

POSITIVE, NEGATIVE それぞれ 1000 サンプルの場合において識別率 100% を達成したネットワークを用いても、全遺伝子データベース（ddbjpri.seq, 24.5Mbp）に対する実行では、最終的な識別率は 90% 程度であった。識別率が低下するのは、本来認識すべきスプライス部位では反応せずに、スプライス部位でない場所で誤って反応してしまう回数が圧倒的に多いためである。つまり、実際の生体になぞらえるなら、誤った場所でスプライシングを行なってしまい、意味のないタンパク質を生成してしまうことに相当する。

スプライシング部位はそうでない部位に比べて圧倒的に少ないため、過ったスプライシングが多くの部位で生じてしまう。例えば、データベースに登録されている遺伝子の塩基配列の長さが 5000 塩基であるとする。ネットワークの正当率が 90% だとすると、誤った判断を下した部位は 10%、つまり約 500 個もの部分配列に対して誤ったスプライシングを行なっているわけである。なおかつ、この程度の長さの塩基配列においては、通常スプライス部位はたかだか数箇所しか存在しない。ネットワークの教師用データ数に非常な偏りがあるため、ネットワークによる汎化が困難となっている。

しかもこれは、研究者が特定の機能遺伝子をシーケンスしたデータに対しての結果である。将来膨大な量の遺伝子データがオートシーケンスされるようになれば、コーディング領域の占める比率はさらに減少するであろう。つまりコーディング領域を自動抽出するためには、さらに優れた識別能力を持つネットワークが必要である。

このことからも、生体におけるスプライス機構がいかに精密で、かつ巧妙な手段を用いているのか感嘆せずはいられない。

よって、ネットワークの予測精度をさらに向上させるために、システムに次の拡張を行なった。あらかじめ設定した識別率を持ったネットワークが出現するか、あるいは一定回数の GA サイクルを実行するまで、前述のシステムと同じく教師サンプルセットを用いて学習を行なってゆく。その後に、集団中で最も優れた識別率を持つネットワークを用いて、データベース ddbjpri.seq に対してスプライシング部位の予測を行わせる。このときに、入力される部分配列のうち、誤ってスプライス部位だと反応した配列を、ネットワークの教師サンプルセットへ追加してゆく。このモードをスイープモードと呼ぶことにする。教師サンプルセットに一定数のサンプルが追加された後、この教師サンプルセットを用いて再び GA サイクルを実行する。以上の手続きを繰り返す（図 8）。

このように、GA サイクルとデータベースによる評価とを組合せ、繰り返し実行することによって、識別能力の向上が向上する。

3.2 実行結果

このシステムにより生成されたネットワークを遺伝子データベース DDBJ における ddbjpri.seq に適用した。当初の教師サンプルセットのみで学習を行った場合に比べて、識別率の向上を確認した。実際に 5' スプライス部位において識別率 95% のネットワークを得、特に、スプライス部位でない領域での識別率が大幅に向上した。

学習の初期においては、エリートネットワークを用いているにかかわらず、スイープモードでの識別誤りが頻発し、スイープモードを終了してしまう。しかし、世代を経て学習サンプルも増加するにしたがって、スイープモードでの識別誤りの頻度が急速に減少してゆくの

が観察される。

5' スプライス部位において、ネットワークが識別を誤った部分配列を、POSITIVE, NEGATIVE サンプルについていくつか示す（表 1, 2, 3, 4）。各表の第一要素はその部分配列に対するネットワークの出力におけるエラー値、第二、第三要素は部分配列が存在する LOCUS 名とその位置、第四要素が誤った識別を行なってしまった部分配列である。

表 1, 2はスプライス部位であるにもかかわらず、ネットワークがスプライス部位でないと判断した領域を示し、表3,4はスプライス部位でないにもかかわらず、そうであると判断した領域を示している。特に表 4においては、いわゆる Chambon 則の 5' 側のパターンがほぼ満たされているため、一見スプライス部位と見間違えかねない。逆に表 2、特にローカス名 HUMALIDN02 などはとてもスプライス部位とは考えられない配列となっている。

ここに挙げたように、エラーサンプルの中にはそれぞれ明確な違いを持たない配列だけでなく、誤って当然ともいえる配列が数多く存在する。

4 結論

ニューラルネットに進化システムを適用したシステムを開発し、このシステムがニューラルネット単体の場合よりも優れたスプライス部位の識別能力を持つことを確認した。また、GA サイクルとスイープモードを交互に実行することにより、動的かつ効果的に教師サンプルを獲得することが可能となり、さらに識別能力が向上した。現在のシステムにおいて、DDBJ における ddbjpri.seq 上で、識別率 95% 以上を示すネットワークを安定して生成することが可能である。

前章で触れたように、エラーサンプル配列中には POSITIVE, NEGATIVE 互いに識別不可能な配列が多く存在することから、現在のシステムにおけるニューラルネットの汎化能力は、ほぼ限界に達していると推測される。実際、表 4で挙げたような部分配列は、もはや現在の方法では識別不可能であろう。Chambon 則を満たすような配列はスプライス部位以外にも大量に存在し、かつスプライス部位の両側数十塩基の比較では識別出来ない領域が膨大な遺伝子データ中に数多く存在するからである。

このような領域をも正確に識別し、精度をさらに向上するには、塩基配列のレベルとは異なる情報を利用したシステムが必要であろう。例えば、コドンの使用頻度やモザイク構造の情報、さらに塩基配列上の文脈情報などを考慮する必要があると考えられる。このような遺伝子自体の構造情報を考慮することによって、より精度の高いシステムを作成できるものと考える。

謝辞

本研究を進めるにあたり、ATR 人間情報通信研究所の東蔵洋一社長、下原勝憲室長、ならびに北陸先端科学技術大学院大学の木村正行教授に研究の機会を与えていただいたことに深く感謝致します。

表 1: ポジティブエラーの領域 (エラーの小さいもの)

ERROR	LOCUS	POSITION	EXON : INTRON
0.00486	GCRHBBA7	[265]	cgcgaccaggctctagagag : gtgggggcaggccaggcgat
0.00895	HUMALBGC	[15116]	tgccctgtgcagaagactat : gtgagtctttaaaaaaatat
0.00960	HSL7A	[4164]	acgacgtggatcccatcgag : gtgcgtttgcctgttgactg
0.00960	HSSURF3	[2321]	acgacgtggatcccatcgag : gtgcgtttgcctgttgactg

表 2: ポジティブエラーの領域 (エラーの大きいもの)

ERROR	LOCUS	POSITION	EXON : INTRON
0.50000	HUMALIDN02	[1687]	cggctggactacatctccct : ccacaggaaggtgcgccctg
0.50000	HUMAMD04	[725]	cgtatgaattctgactgttg : gttgtttaatgcaattttgt
0.50000	HUMAMD04	[966]	ctgcaaaggatgtcactcgt : gtaagcatttttagtaataa
0.50000	HUMAMPD1X	[6642]	attctctggatgttcatgct : gtaggttgaaacggcaatgt

表 3: ネガティブエラーの領域 (エラーの小さいもの)

ERROR	LOCUS	POSITION	EXON : INTRON
0.00089	HSNMYC	[2617]	caggggtgggcttagagagc : ttccaattaagctattggca
0.00037	HSLCATG	[6540]	tactcgggaggctgaggcag : gagaattgtttgaacctggg
0.00037	HSMPOG	[5959]	tcaggggcagggaactcccg : gacagagaacgctgtggtgc
0.00022	HSPRB4S	[1714]	ctttttctgcttacaaatgg : gtcatttctccagtgtcttc

表 4: ネガテイブエラーの領域 (エラーの大きいもの)

ERROR	LOCUS	POSITION	EXON : INTRON
0.49985	AGGGLINE	[2963]	cgggtgaggcccgggggccg : gtgggtggctagggatgaag
0.49985	ATRINS	[1189]	acctcggagggcacggcagg : gtagggtcctccctccacgt
0.49985	BABAPOE	[272]	ccgacccgctagaaggtggg : gtggggagagcatgtggact
0.49985	BABAPOE	[3867]	gccgcgtgcgggccgccact : gtgggctccctggccagcca

参考文献

[1] Tetsuya Maeshiro, Ken-nosuke Wada: Evolutionary System for the Computer Screening of the Coding Region of Human Genome, ATR technical report TR-H-??? (1993)

[2] 和田健之介他: 遺伝子コーディング領域のコンピュータ・スクリーニング, 数理科学 (10/1993)

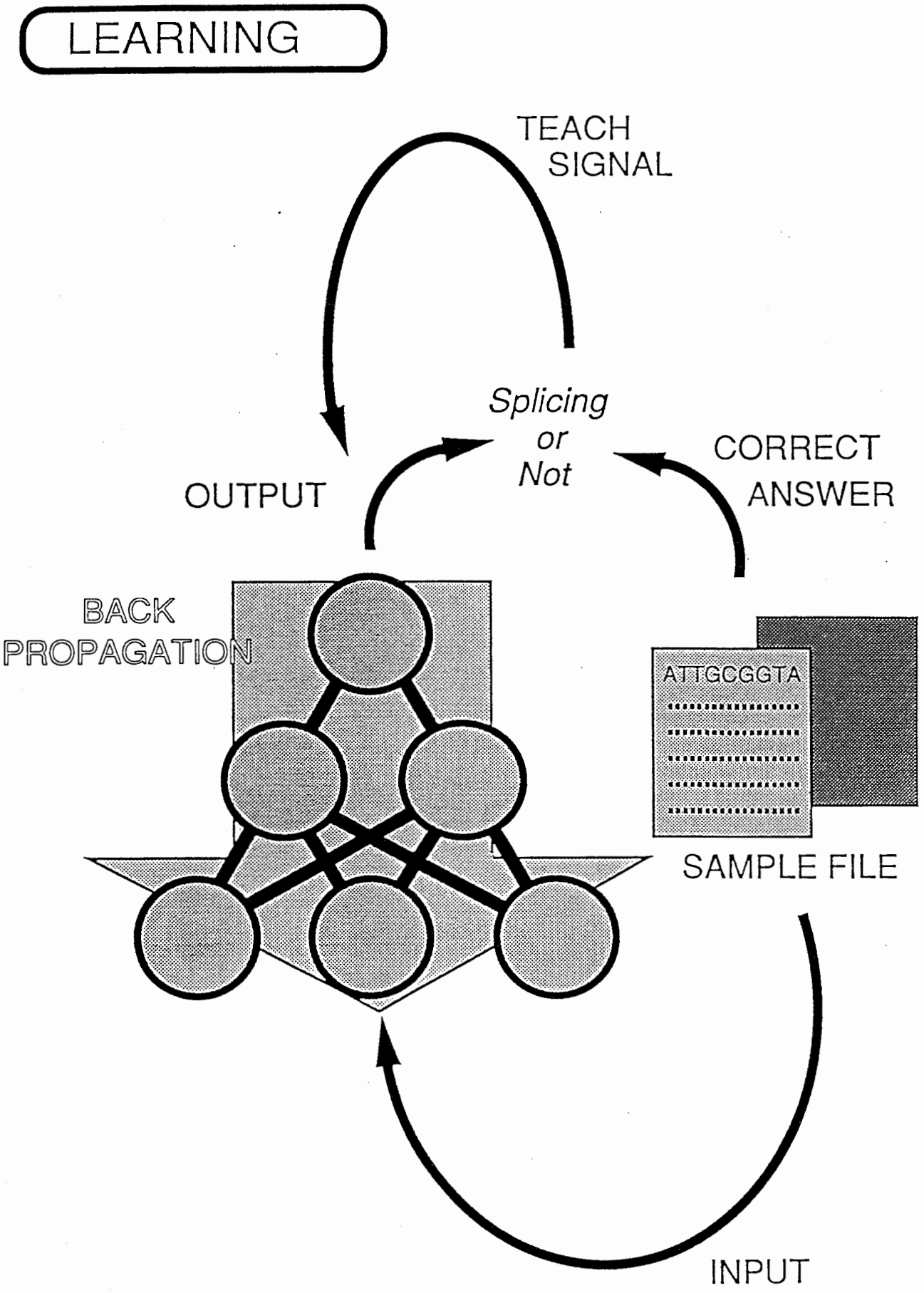
LEARNING
TEACH
SIGNAL
Splicing
or
Not
CORRECT
ANSWER
OUTPUT
BACK
PROPAGATION
ATTGCGGTA
SAMPLE FILE
INPUT

図 1

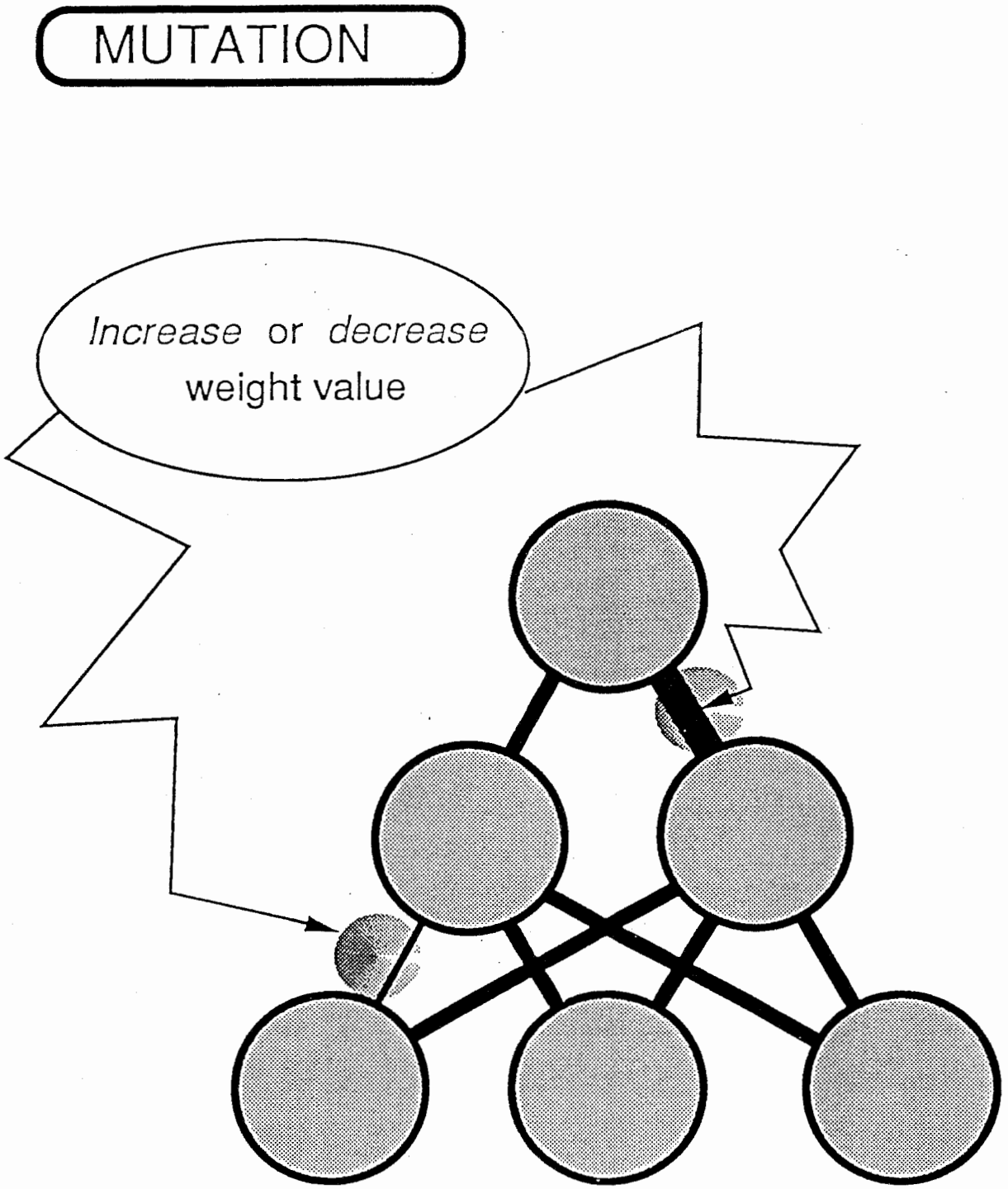

図 2

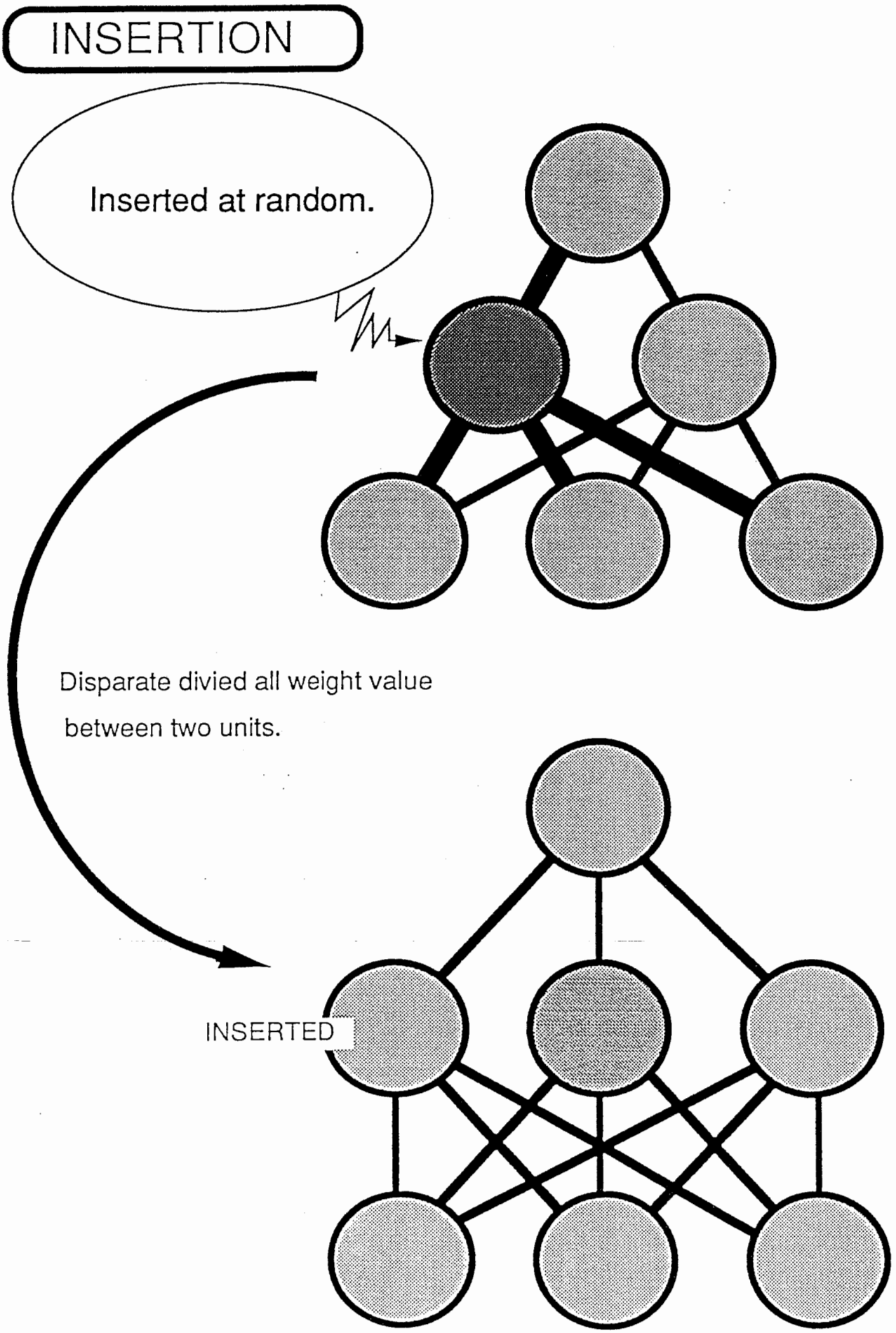
INSERTION
Inserted at random.
Disparate divied all weight value
between two units.
INSERTED

図 3

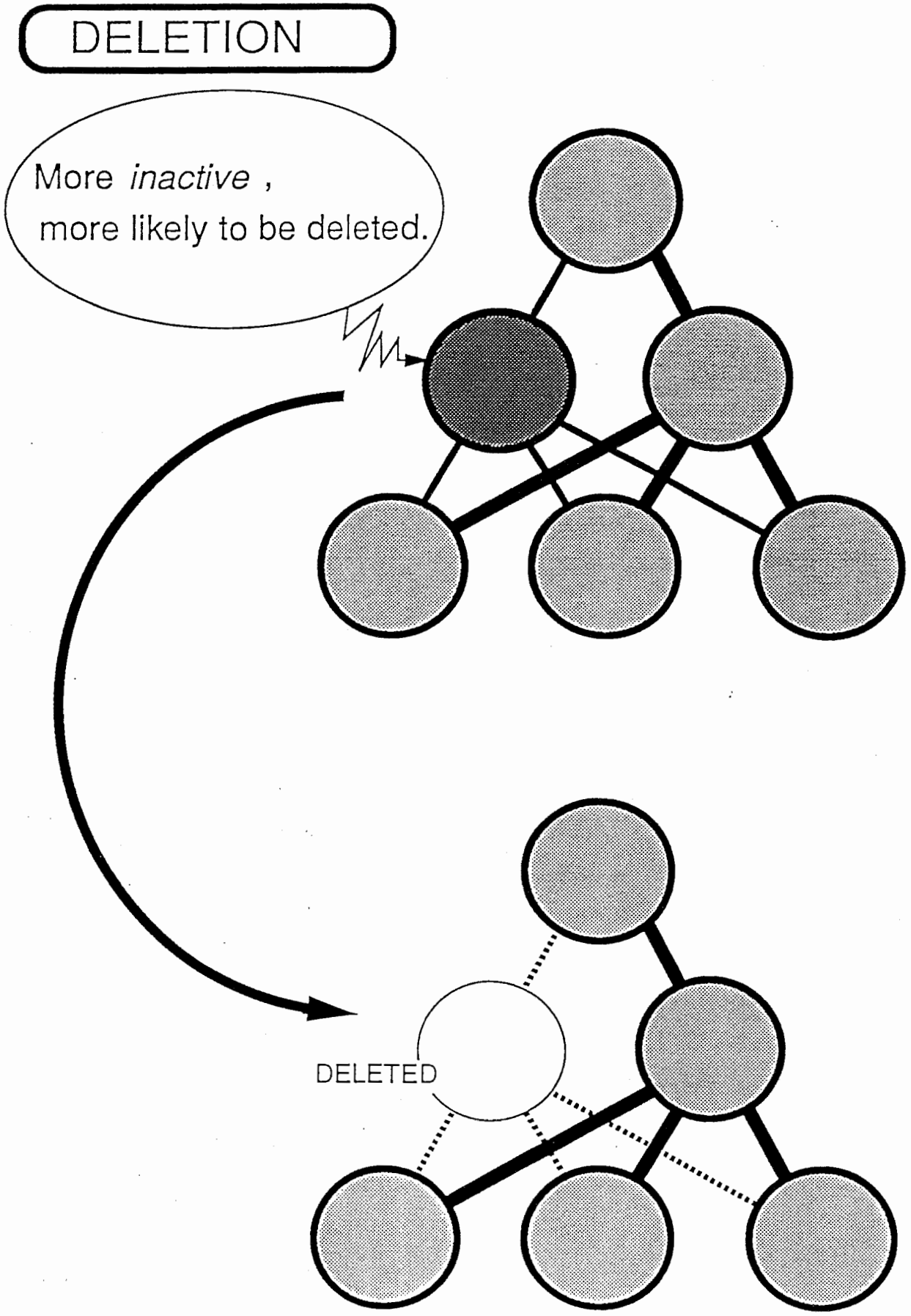

図 4

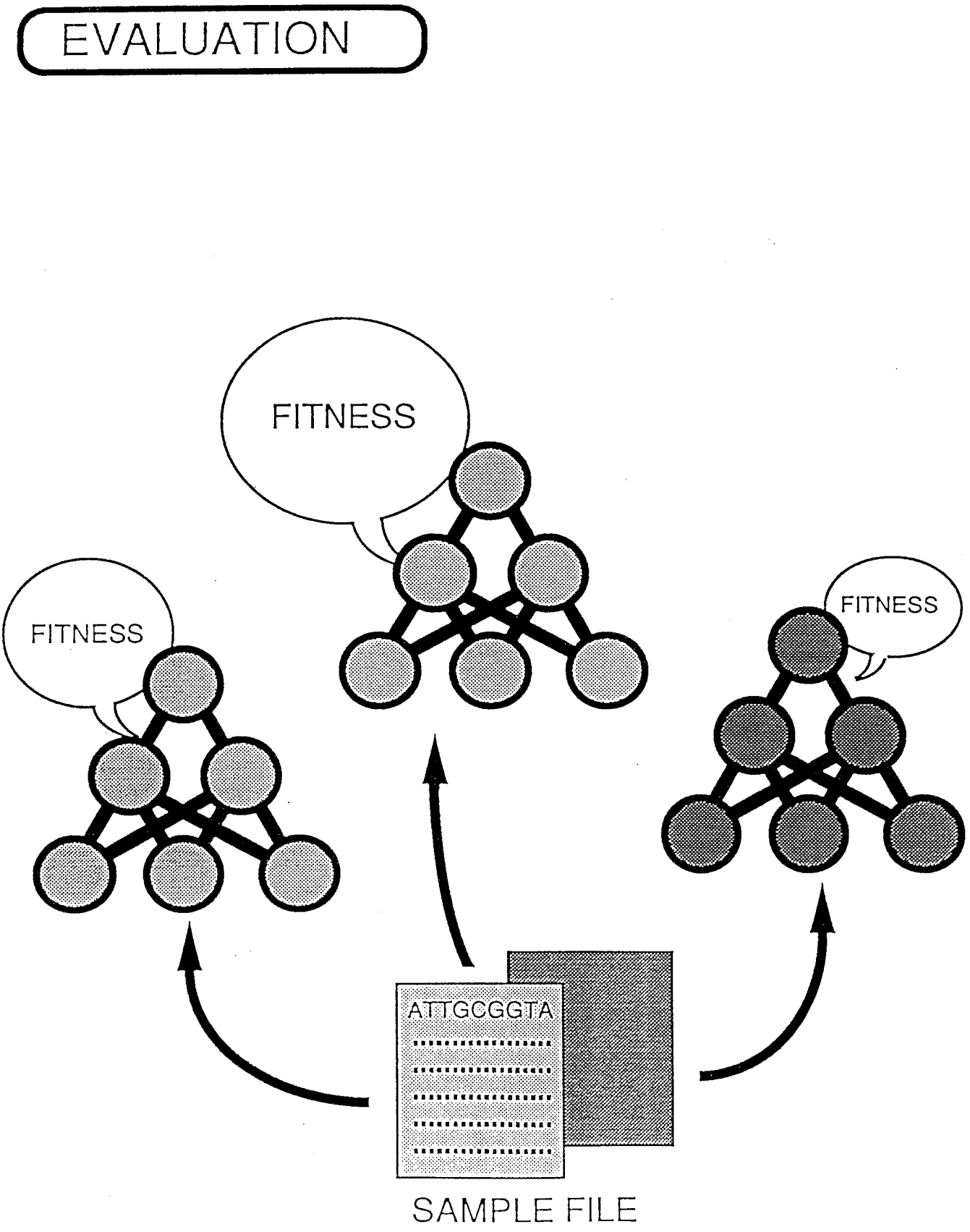
EVALUATION
FITNESS
FITNESS
FITNESS
ATTGCGGTA
SAMPLE FILE

図 5

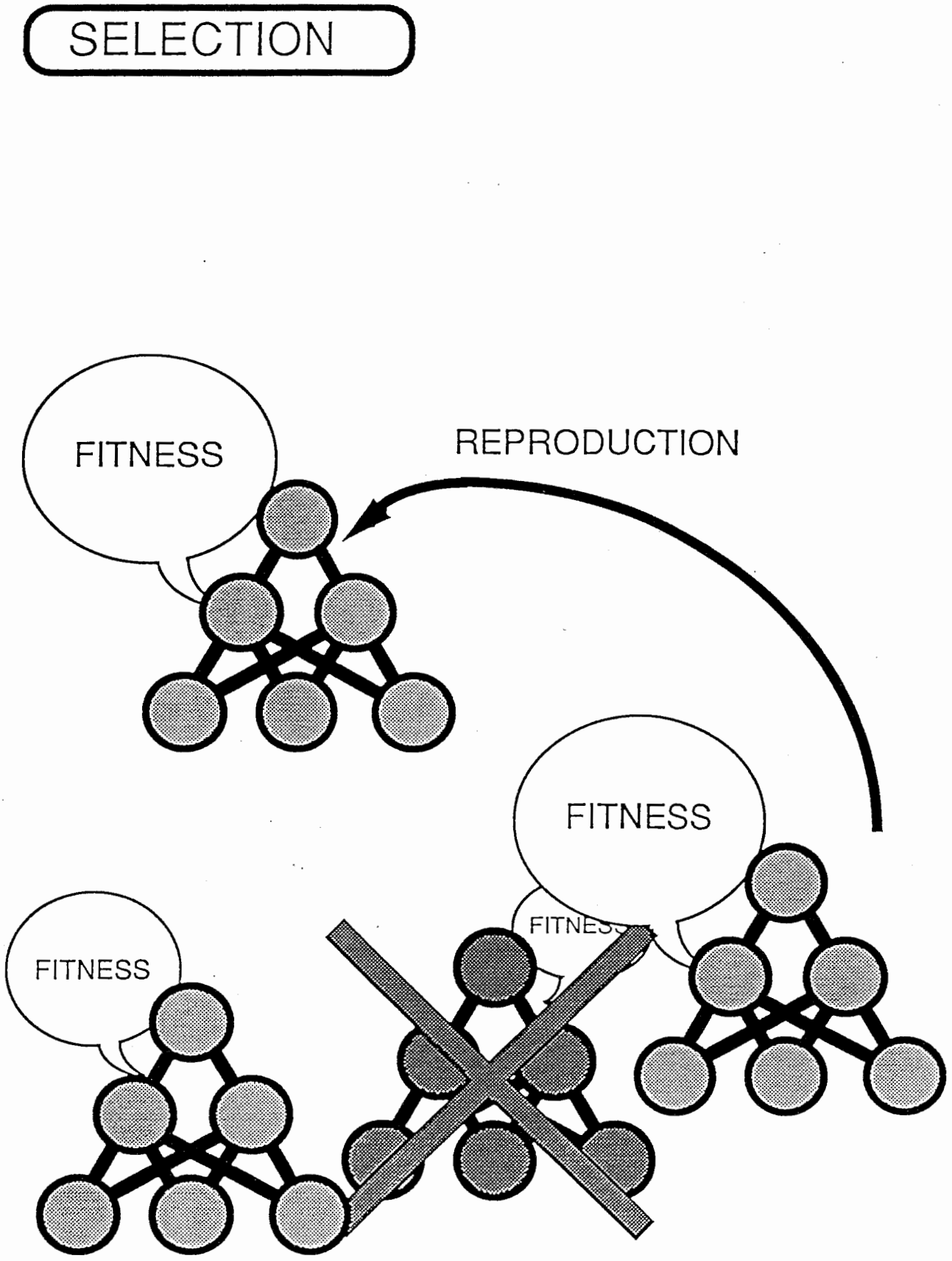

図 6

Sweep DNA sequence

Auto recognizer for DNA
coding region

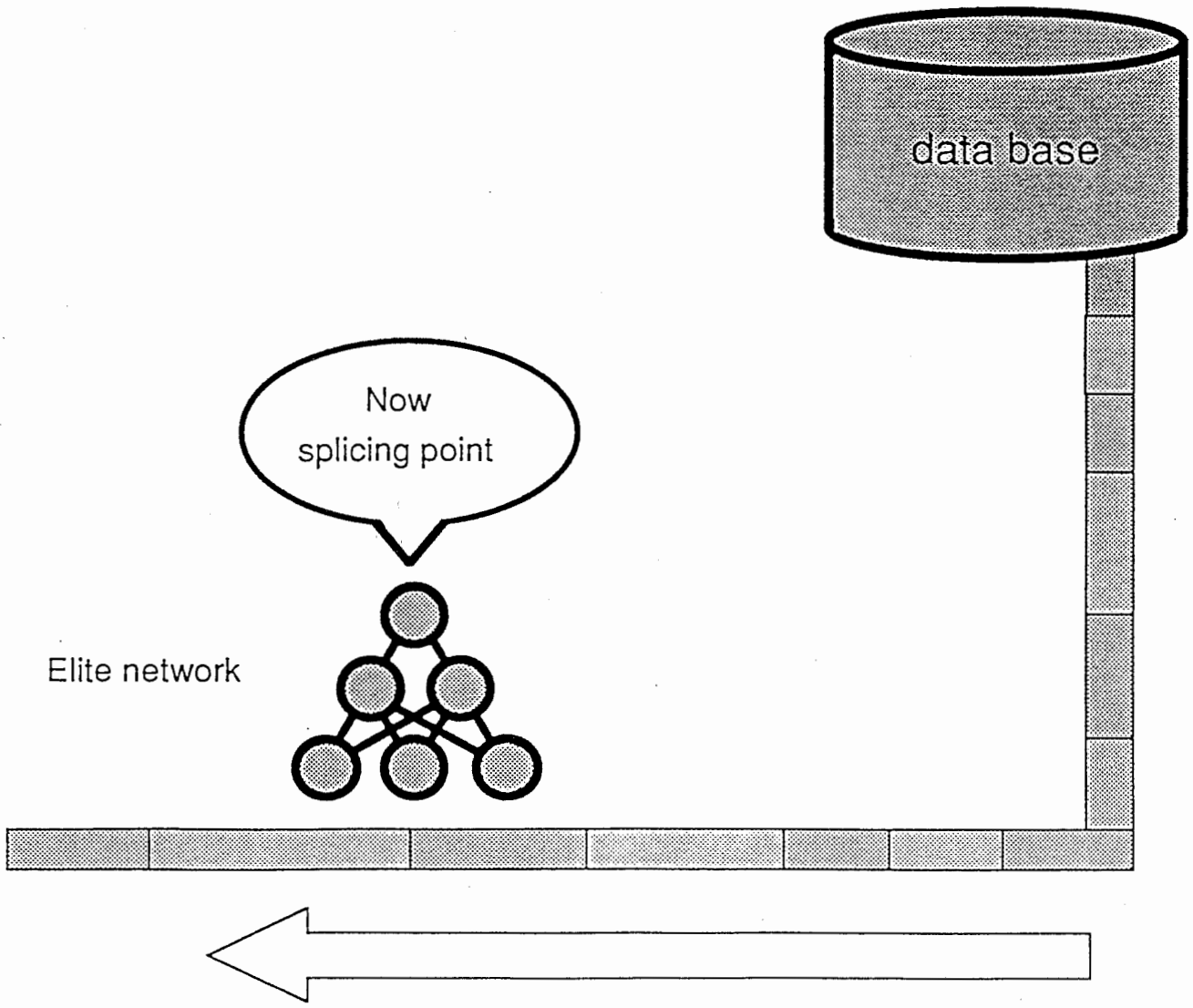

図 7

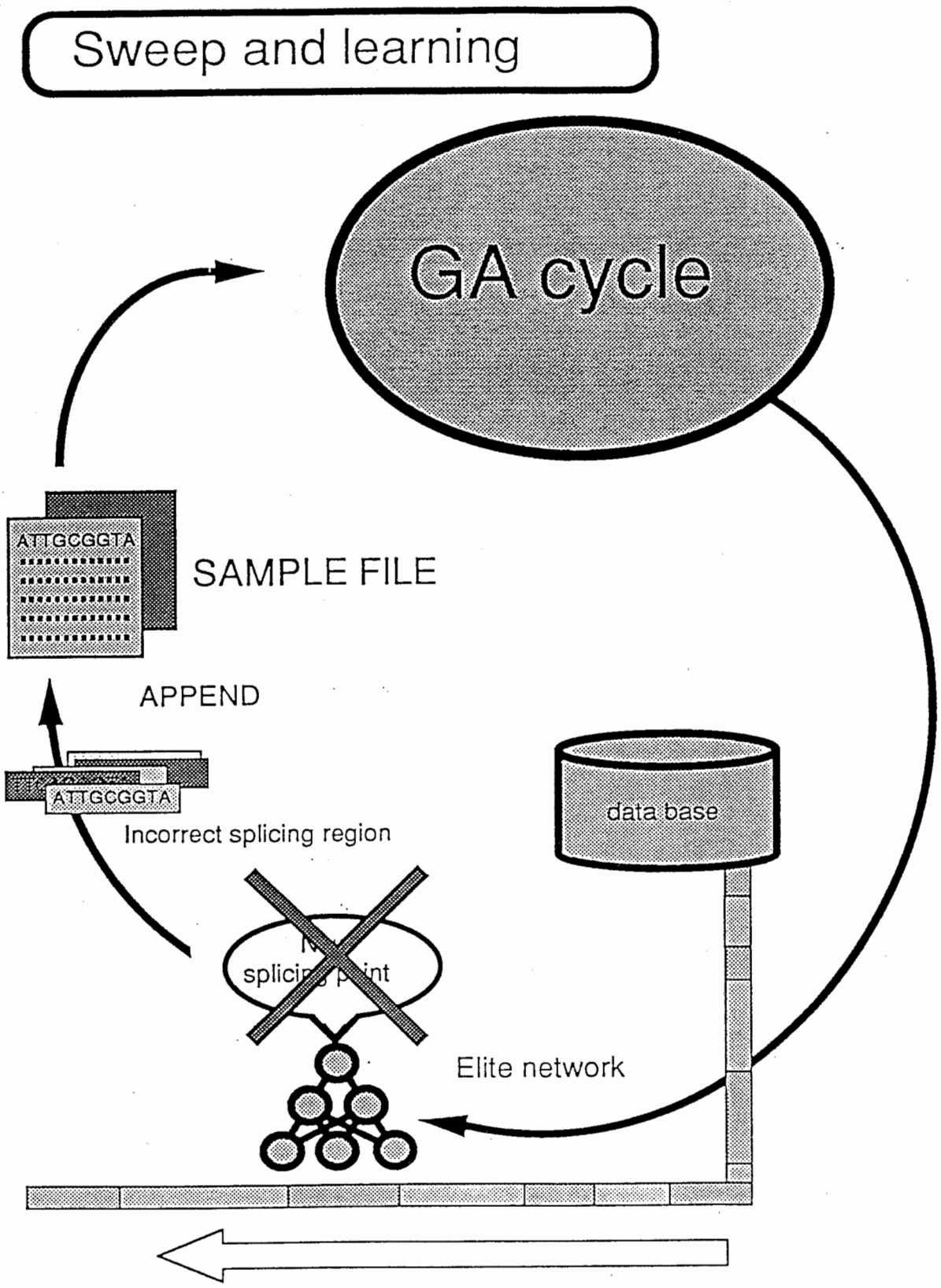
Sweep and learning
GA cycle
ATTGCGGTA
SAMPLE FILE
APPEND
ATTGCGGTA
Incorrect splicing region
data base
Elite network

図 8