

XML による分子系統樹データの表現とその検索手法

増山 智久[†] 岸 浪 り さ[†] 吉川 孝 伸[†]
松田 秀 雄[†] 橋 本 昭 洋[†]

分子系統樹とは、生物の持つ DNA 塩基配列等のデータから作成された、遺伝子の進化過程を木構造で表現したものである。分子系統樹はその有用性に関わらず、公表された分子系統樹を体系的に格納し、検索できるように整備されたデータベースが存在しなかった。そこで、本研究では、分子系統樹データを XML に基づいた書式でデータベースに格納し、検索できるシステムを開発している。XML を採用することで分子系統樹のデータそのものだけでなく出典などの分子系統樹に関する付加情報もデータとして扱えるようになる。また、分子進化解析の研究の進展に伴って発生する新たな関連や機能に関する情報の追加にも柔軟に対応することが可能となる。本論文では XML を用いて分子系統樹データを表現する方法を示し、そのデータベースの検索方法について説明する。

Representation and Retrieval Method for Molecular Phylogenetic Tree Data Based on XML

TOMOHISA MASUYAMA,[†] RISA KISHINAMI,[†] TAKANOBU YOSHIKAWA,[†]
HIDEO MATSUDA[†] and AKIHIRO HASHIMOTO[†]

A molecular phylogenetic tree is a tree-structured graph that represents the evolutionary process of genes, which is constructed from sequence data (such as DNA sequences) obtained from several organisms. Although molecular phylogenetic trees are fundamental data structures in evolutionary analysis, the database prepared so that to store the molecular phylogenetic tree announced systematically and it can look it up didn't exist. In this research we develop a system which stores molecular phylogenetic tree data by a form based on XML. By adopting XML, we can handle not only the data themselves of molecular phylogenetic trees but also the additional information related to the molecular phylogenetic trees such as their bibliographic information. It becomes possible to do with flexibly addition of the information about the new relations and functions which are obtained by the development of molecular evolution analysis. In this paper, we introduce the way to express the molecular phylogenetic tree data by using XML and explain about the way of referring to the database.

1. はじめに

近年のバイオテクノロジーの急速な発展により、多数の生物種から DNA 塩基配列やタンパク質アミノ酸配列等の遺伝子に関するデータが大量に得られつつあり、いくつかの微生物ではそれらの持つ DNA 塩基配列が完全に解読されている。

生物の持つ表現型の多様性は、多数の遺伝子の存在を反映していると考えられており、各生物の持つ全遺伝子集合を解析することにより、その生物の全体像を遺伝子レベルから理解しようという試みが盛んに行われている。そのような試みの一つとして、遺伝子の分子進化解析がある。この解析では、複数の生物種間で、

DNA 塩基配列またはそれが翻訳されたアミノ酸配列（以下、単に配列と表記する）が類似した遺伝子を抽出し、それらの配列上で塩基（またはアミノ酸）の異なる個数から遺伝子間の距離を定義し、その距離に基づいて遺伝子を階層的に分類する。この分類結果は、現在の遺伝子を葉節点とし、配列相互の類似度に応じて遺伝子を階層的にまとめた木構造となっている。

生物の進化の過程で、DNA 塩基配列上に突然変異により生じる塩基置換は時間とともに一定の割合で生じると考えられているので、前述の遺伝子間の距離は時間に比例しており、分類の結果得られる木構造は遺伝子の進化過程を反映していると考えられる。このように、遺伝子の配列相互の比較により得られる分類木を分子系統樹と呼ぶ。

分子系統樹には、対象としている遺伝子を持つ生物の種分化に伴う進化過程以外に、遺伝子重複により同じ生物の DNA 塩基配列上で遺伝子のコピーが生じ、

[†] 大阪大学大学院基礎工学研究科情報数理系専攻
Graduate School of Engineering Science, Osaka
University

それらが独立に進化する過程（遺伝子の機能分化過程）も含まれている。後者は、一般に遺伝子ごとに異なるため、分子系統樹の構造は遺伝子ごとに異なることになる。

分子生物学の研究者が分子進化解析を行うときは、配列データベースを検索したり、自身で直接、遺伝子の配列を読み取ることにより得られた配列を使って、次章で述べるような方法で分子系統樹を作成することになる。このとき、同じ遺伝子を使った分子系統樹が既に公表されているか、もし公表されているとすれば、それと自分が作った分子系統樹の構造が一致するかどうか調べたいという要求がある。また、遺伝子は一般に単独で働くことは少なく、別の遺伝子と相互作用することが多いが、相互作用をする遺伝子の組では、それぞれの分子系統樹の分岐関係が類似することが多いとされており、共進化の可能性が指摘されている¹⁾。このような解析でも、分子系統樹の構造比較が重要となる。

分子系統樹のデータは、いくつかのデータベースにまとめられ、WWW等によりインターネット上で公開されている^{2)~5)}。しかし、これらのデータベースでは、データベース中の分子系統樹を閲覧するインターフェースしか備えておらず、ユーザの作成した分子系統樹とデータベース中の分子系統樹との木構造を比較し、それらの間の相違点を調べることができない。また、生物の系統関係の解析を目的としているデータベースが多く、個々の遺伝子ごとに作成された分子系統樹を検索できるものは少ない。

そこで、本研究では、既に公表されている分子系統樹をXMLに基づいた書式で体系的に格納し、検索できるデータベースの開発を目指している。現在までに数多くの分子系統樹が公表されているが、これらの分子系統樹は、学術雑誌等の紙面上で図として公表されることがほとんどであるので、これらの図をスキャナ等で読み取って、パターン認識により木構造データに変換してデータベースに格納する。また、ユーザの入力した分子系統樹も同様の木構造データに変換し、木構造どうしの照合が行える機構を備えている。

以下では、2章で分子系統樹の原理と構造の比較についてまとめ、3章ではXMLでのデータベースの書式とデータベースの検索方法について述べ、最後に結論と今後の課題について述べる。

2. 分子系統樹

2.1 遺伝子配列と分子系統樹

分子系統樹を作成するための入力データとしては、生物や遺伝子の進化の過程を反映する様な配列、すなわちそれらの生物間で共通の祖先種における単一の遺伝子配列から受け継がれた遺伝子配列を使う必要がある。これは、実際には各生物間で共通の機能を持つ遺

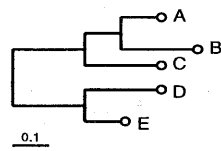


図1 分子系統樹の例

Fig. 1 Example of a molecular phylogenetic tree.

伝子の配列を選び出すことにより行われる。

共通の機能を持つ遺伝子配列であっても、生物の進化の過程で、塩基の挿入・欠損により配列長が多少異なっている。そこで分子系統樹を作る前に、多重アラインメント (multiple alignment) という処理を行い、複数の配列で同じ文字ができるだけ同じ位置にくるようにギャップ (挿入記号) を入れて補正する。これにより、分子系統樹の作成の際には、配列間での文字の置換だけを考えればよいことになる (文字の挿入・削除はそれぞれ、ギャップからの文字への置換、文字からギャップへの置換とみなせる)。

分子系統樹は、節点とそれらを結ぶ辺からなる木構造グラフで表される。節点には、接続の次数が1である葉節点と次数2以上の内部節点があり、葉節点は前述の多重アラインメントされた遺伝子配列を表し、内部節点はそれらの遺伝子の共通祖先となる遺伝子の配列を表す。また、節点間の辺は枝と呼ばれ、その長さはその枝の両端の節点が表す配列間の置換回数を表す。

図1に遺伝子配列の数が5のときの分子系統樹の表現例を示す。図1では左から右に遺伝子の分岐 (生物種の分化かまたは同一生物種中での遺伝子の重複) と遺伝子配列上での置換が生じたことを示し、葉節点と内部節点をそれぞれ○と垂直の線 (この線の長さには意味はない)、枝を水平の線で表している。葉節点に振られたA~Eのラベルはそれぞれの葉節点の配列に付けられている遺伝子名などに当たる。

枝の長さは両端の節点が表す配列間の置換回数を表す (一方の端の配列からその長さで表された回数の置換が起こるともう一方の端の配列に変化することを示す)。図1で0.1とラベルがついた線はスケールバーと呼ばれ、この線の長さが配列の各位置あたり平均0.1回の置換に対応する長さを示している。

2.2 分子系統樹の構造比較

分子生物学者は、遺伝子データから分子系統樹を作成するだけでなく、機能未知の遺伝子の機能を解析したり新しい知見を得るために、他の分子系統樹と構造を比較することがある。例えば、種系統樹と遺伝子による分子系統樹との違いから種分化に依らない遺伝子重複・遺伝子転移を探索する場合、複数の種類の遺伝子から作成した分子系統樹どうしを相互に比較することにより共進化関係にある遺伝子の組を探索する場合などである。

例えば、図2は、人間の発育因子 (growth factor)

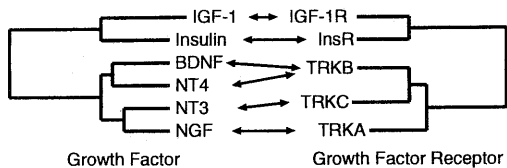


図2 機能的に関連のある遺伝子の分子系統樹間の比較例
Fig. 2 Example for comparing molecular phylogenetic trees of functionally-related genes.

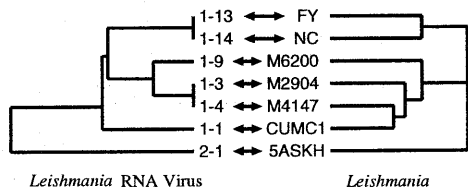


図3 ウィルスとその宿主の遺伝子の分子系統樹間の比較例
Fig. 3 Example for comparing molecular phylogenetic trees of genes obtained from viruses and their hosts.

の遺伝子配列から作成した分子系統樹(左側)と、発育因子からシグナルを受けとる受容体(receptor)の遺伝子配列から作成した分子系統樹(右側)を示す¹⁾。図2の中央の矢印は発育因子と受容体との間のシグナルの授受の対応関係を示している。図2からわかるように機能的に関連した遺伝子の分子系統樹は分岐関係が類似しており、両者の間の共進化関係が示唆される。しかし、発育因子と受容体はそれぞれ独立に突然変異の影響を受けるため、両者の系統樹の分岐関係や枝長にはいくつか食い違いが生じる。

他の例として、寄生体とその宿主との系統関係の類似性がある。図3はリーシュマニア(住血鞭毛虫)属に属するいくつかの種から採取した配列から作成した分子系統樹(右側)と、それらに寄生するウィルスから採取した配列から作成した分子系統樹(左側)であり、中央の矢印はウィルスとそれが寄生する宿主の対応関係を表している⁶⁾。図2の例と同様、これら2つの系統樹は分岐関係はある程度類似しているが、枝長に関しては非常に異なっている。これは、ウィルスとその宿主の間のライフサイクルやそれらの持つ遺伝子の機能の違いから生じる配列の置換速度の違いによるものと考えられる。

図2や図3の例からわかるように、共進化関係にある遺伝子の探索では、それらの分子系統樹が構造的に似ているか似ていないかを比較する必要がある。

比較方法には対数尤度⁷⁾の差に基づく比較、最大一致部分木⁸⁾による比較、split distance⁹⁾に基づいた比較などがある。ここではsplit distanceに基づいた比較方法を利用する。split distanceは分岐関係だけから分子系統樹の構造の違いを求めることができ、計算量が少ない。また、計算の結果、2つの系統樹が似て

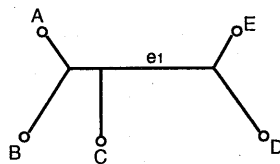


図4 図1の分子系統樹の木構造
Fig. 4 Tree structure of the molecular phylogenetic tree shown in Fig. 1.

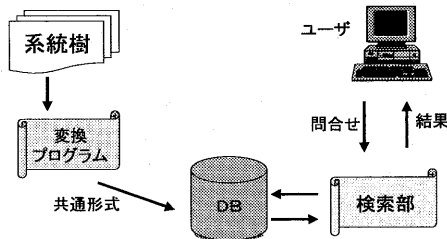


図5 分子系統樹データベースシステムの概要
Fig. 5 Outline of our molecular phylogenetic tree database system.

いるほどsplit distanceの値が小さくなるといった性質がある。

3. 分子系統樹データベース

3.1 データの書式

本研究では分子系統樹を体系的に格納し、検索できるように整備されたデータベースの実現を目指している。単純に分子系統樹を画像ファイルとして蓄積した場合、データ量が大きくなることと構造の比較が困難になるため、分子系統樹をNewick形式に基づくXML表現に変換して蓄積する。

分子系統樹の構造を表現するための形式として最も良く使われているものにNewick形式がある¹²⁾。Newick形式とは、木構造をカッコ“(”, “)”とカンマ“,”を使い構造を一意に決定できる表記法であり、枝の長さも情報として持たせることが可能である。例えば、図1をNewick形式に変換すると次のようになる。

((A:0.1, B:0.2):0.1, C:0.2, (E:0.1, D:0.2):0.4);

表記中のコロン“:”の後ろの数字は枝の長さを表している。一つの木構造から最初に注目する葉節点や内部節点の違いでいくつかのNewick形式の記述が得られるが、一つのNewick形式の記述からは木構造は一意に定まる。

```

<!ELEMENT phylotree (newick, description, reference)>
<!ELEMENT newick ((branch | leaf), (branch | leaf)+)
<!ELEMENT branch ((branch | leaf), (branch | leaf)+, length)>
<!ELEMENT leaf (geneid, label, length)>
<!ELEMENT geneid (#PCDATA)>
<!ELEMENT label (#PCDATA)>
<!ELEMENT length (#PCDATA)>
<!ELEMENT description(#PCDATA)>
<!ELEMENT reference (author, title, journal, volume, number, page, year, figNo)
<!ELEMENT author (#PCDATA)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT journal (#PCDATA)>
<!ELEMENT volume (#PCDATA)>
<!ELEMENT number (#PCDATA)>
<!ELEMENT page (#PCDATA)>
<!ELEMENT year (#PCDATA)>
<!ELEMENT figNo (#PCDATA)>
<!ATTRIBUTE phylotree treeid ID #REQUIRED>

```

(a) 分子系統樹を表現する DTD

```

<!ELEMENT gene (genesymbol?, species, description, type, related*, TreeRef*)
<!ELEMENT genesymbol (#PCDATA)>
<!ELEMENT species (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT type (#PCDATA)>
<!ELEMENT related (#PCDATA)>
<!ELEMENT TreeRef (#PCDATA)>
<!ATTRIBUTE gene geneid ID #REQUIRED>
<!ATTRIBUTE relatedtype CDATA #REQUIRED>

```

(b) 遺伝子を表現する DTD

図 6 分子系統樹データベースのデータ表現のための DTD
Fig. 6 DTD for entries of molecular phylogenetic tree database.

Newick 形式は、分子系統樹の木構造をコンパクトに表現できるため、必要な記憶容量が少なくすむという利点がある。しかし、表現できる情報が限られているため、例えば図 2 や図 3 のような分子系統樹どうしの関連や、分子系統樹の典拠等の文献情報、分子系統樹の葉節点に対応する遺伝子の機能情報を表現するのが困難である。

そこで、我々はデータベースに格納するデータの表現形式として、XML¹³⁾に基づく書式をとることにした。XML を選んだのは、(1) Newick 形式と同様テキスト形式での表現であるため、簡単なパーサで Newick 形式との間で相互に変換でき、既存のアプリケーション等の資産が利用できる、(2) 分子系統樹に付随する情報は基本的にその典拠である論文に書かれた文書情報であるため、構造化文書の表現形式である XML での記述に適している、(3) 分子系統樹どうしの関連の情報や葉節点の遺伝子の機能情報はデータの種類が非常に多くだけでなく、分子進化解析分野の研究の進展に伴い新たな関連や機能に関する情報が次々に追加されるので、あらかじめスキーマを決定するのが困難で、XML のように柔軟な表現形式が向いている、という理由による。

図 6 にデータベースの書式の DTD(Document Type Definition) を、図 7 にこの DTD を使って書

```

<phylotree treeid="T0001">
  <newick>
    <branch> <leaf> <geneid> G0001 </geneid>
      <label> IGF-1 </label>
      <length> 0.14 </length>
    </leaf>
    <leaf> <geneid> G0002 </geneid>
      <label> Insulin </label>
      <length> 0.09 </length>
    </leaf>
    <length> 0.39 </length>
  </branch>
  <branch> <leaf> <geneid> G0003 </geneid>
    <label> BDNF </label>
    <length> 0.09 </length>
  </leaf>
  <leaf> <geneid> G0004 </geneid>
    <label> NT4 </label>
    <length> 0.08 </length>
  </leaf>
  <length> 0.005 </length>
</branch>
<branch> <leaf> <geneid> G0005 </geneid>
  <label> NT3 </label>
  <length> 0.09 </length>
</leaf>
<leaf> <geneid> G0006 </geneid>
  <label> NGF </label>
  <length> 0.08 </length>
</leaf>
<length> 0.015 </length>
</branch>
</newick>
<description> Coevolution of the insulin-NGF
  gene family and their receptors
</description>
<reference>
  <author> K. J. Fryxell </author>
  <title> The coevolution of gene family trees
  </title>
  <journal> Trends in Genetics </journal>
  <volume> 12 </volume> <number> 9 </number>
  <page> 365 </page> <year> 1996 </year>
  <figNo> 1 </figNo>
</reference>
</phylotree>

```

(a) 図 2 の growth factor の分子系統樹の表現例

```

<!-- growth factor genes ----->
<gene geneid="G0001">
  <genesymbol> IGF-1 </genesymbol>
  <species> Homo sapiens </species>
  <description> insulin-related growth factor 1
  </description>
  <type> growth factor </type>
  <related type="receptor"> G0002 </related>
  <TreeRef> T0001 </TreeRef>
</gene>
...
<!-- receptor genes ----->
<gene geneid="G0002">
  <genesymbol> IGF1R </genesymbol>
  <species> Homo sapiens </species>
  <description> insulin growth factor 1 receptor
  </description>
  <type> growth factor receptor </type>
  <related type="growth factor"> G0001 </related>
  <TreeRef> T0002 </TreeRef>
</gene>
...

```

(b) 図 2 の系統樹の遺伝子表の表現例 (一部)

図 7 分子系統樹データベースのエントリの例
Fig. 7 Example of an entry of molecular phylogenetic tree database.

かれた分子系統樹データの例を示す。図 7 は図 2 の分子系統樹を表現したもので、分子系統樹のデータ (図 7(a)) からそれが持つ葉節点に対応する遺伝子の情報 (図 7(b)) へ関連は **geneid** という要素、逆の関連は **TreeRef** という要素に、それぞれ遺伝子の ID、分子系統樹の ID を記述することで表現されている。

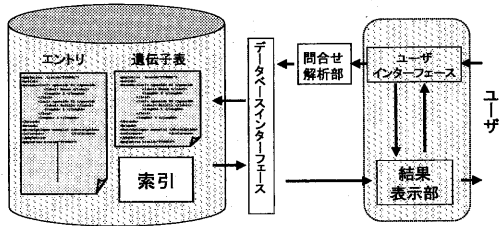


図8 分子系統樹データベースシステムの内部構成
Fig. 8 Internal configuration of our molecular phylogenetic tree database system.

また、図2の2つの分子系統樹の葉節点どうしの対応関係は遺伝子表で **related** という要素で表現されている。

3.1.1 システム構成

本データベースシステムは、系統樹や遺伝子のデータ、索引などを格納したデータベース部、ユーザインタフェース部、そしてそれらの2つをつなぐ部分からなる。この構成の概要を図8に示す。

データベース部分には、XMLで記述された系統樹のエントリを格納した1つのエントリファイル、ラベルと遺伝子名の対応関係などを記した遺伝子表、そして種々の索引ファイルからなる。データの蓄積方法については、特有のスキーマを用いるのではなく、前節で示したXMLで書かれたエントリをテキストファイルの形で蓄積する。このファイルに対して索引を用意しておき、検索を行う。索引は2種類あり、1つは内部でエントリの検索用に用いる索引、もう1つはキーワード検索用の索引である。キーワード検索用の索引は、遺伝子表とエントリファイルから作成する。

ユーザインタフェース部は、情報科学になれていない分子生物学者でも使い易いように特定の問合せ言語を用いずにWWWブラウザ等でテキスト入力、項目の選択を行うことでデータベースへの問合せを可能とする。

問合せに対する検索結果は、系統樹であることがほとんどである。だが、木構造をXML表現のまま出力したのではユーザには分かりにくい。そこでXML表現から木構造のイメージへの変換を行い、視覚的に分かりやすくユーザに提示する。

3.2 データベースの検索

分子系統樹を対象とする検索を考える。問い合わせ形式として以下のものがある。

(1) キーワード → **geneid** 集合

キーワード検索用の索引を引くことでそのキーワード

に関連のある遺伝子の **geneid** の集合を得ることができる。

(2) **geneid** → **treeid** 集合

エントリの検索用の索引を引くことで **geneid** に対応する遺伝子を含む系統樹の **treeid** の集合を得ることができる。

(3) **geneid** → 関連遺伝子の **gene** 要素

geneid に **type** を指定することで **geneid** に対応する遺伝子と **type** に指定した機能関係にある遺伝子の **gene** 要素を得ることができる

(4) **geneid** → 遺伝子表

geneid を指定することで **geneid** に対応する遺伝子の **gene** 要素を得ることができる

(5) **treeid** → 系統樹

treeid を指定することで **treeid** に対応する系統樹の **phylo** 要素を得ることができる

例としてキーワード「**insulin-related growth factor 1**」からキーワードに関連のある遺伝子を含む系統樹とその遺伝子と機能的に対応関係のある遺伝子を含む系統樹を求める手順は以下ようになる。以下のようにになる。

[Step 1] キーワード → **geneid** 集合よりキーワードに関連のある遺伝子の **geneid** 「G0001」を得る

[Step 2] **geneid** → **treeid** 集合より **geneid** 「G0001」から G0001 に対応する遺伝子が葉節点となっている系統樹の **treeid** 「T0001」を得る

[Step 3] **treeid** → 系統樹より **treeid** 「T0001」から対応する系統樹のエントリを得る。これでキーワードに関連のある遺伝子を含む系統樹を検索できる

[Step 4] **geneid** → 関連遺伝子の **geneid** より **geneid** 「G0001」から「G0001」に機能的に対応関係にある遺伝子の **geneid** 「G0002」を得る

[Step 5] **geneid** → **treeid** 集合より **geneid** 「G0001」から **treeid** 「T0002」を得る

[Step 6] **treeid** → 系統樹より **treeid** 「T0002」から対応する系統樹のエントリを得る。これでキーワードに関連のある遺伝子と対応関係にある遺伝子を含む系統樹を検索できる

また、検索条件として分子系統樹の構造データを与えることも可能である。分子系統樹の構造データが与えられた場合、出力としてその構造データに類似した分子系統樹の構造データを得る。その手順は葉節点に関する検索を行い、類似度を計算する分子系統樹の数を絞り **split distance** に基づいた比較計算をする。**split distance** の閾値を定めて計算結果が閾値よりも小さくなった分子系統樹を出力する。これにより指定した分子系統樹の構造データと類似した分子系統樹を得ることができる。

4. おわりに

XMLを用いた分子系統樹データの表現方法とデータを検索する方法について提案した。

提案した方法により分子系統樹データを体系的に格納し整備することが可能となる。また、検索条件となるキーワードを指定することでキーワードに関連のある遺伝子の系統樹だけでなくその遺伝子と機能的に関係のある遺伝子の系統樹を取り出すことができる。

今後は、格納している分子系統樹データの量を増やすとともに提案したDTDの検証及び改良、対話型の問合せユーザインターフェースの開発、XMLで記述された分子系統樹データをユーザに分かりやすく提示するインターフェースの開発、提案するデータベースシステムの実用化を目指す予定である。

参考文献

- 1) Fryxell, K. R.: The Coevolution of Gene Family Trees, *Trends in Genetics*, Vol. 12, No. 9, pp. 364-369 (1996).
- 2) UCMP Exhibit Halls: Phylogeny, *The Museum of Paleontology*, University of California, Berkeley, available at:
<http://www.ucmp.berkeley.edu/exhibit/phylogeny.html>.
- 3) Maidak, B. L., et al: A New Version of the RDP (Ribosomal Database Project), "Nucleic Acids Research, Vol.27, No.1, pp171-173, 1999, Internat address:
<http://www.cme.msu.edu/RDP/analyses.html>
- 4) Maddison, D.R. and Maddison, W. P.: The Tree of Life: A multi-authored, distributed Internet project containing information about phylogeny and biodiversity,(1998). Internet address: <http://phylogeny.arizona.edu/tree/phylogeny.html>
- 5) JUNGLE: Phylogenetic Tree Database, Internet address: <http://smiler.lab.nig.ac.jp/jungle/jungle.html>
- 6) Widmer, G. and Dooley, S.: Phylogenetic Analysis of *Leishmania* RNA Virus and *Leishmania* Suggests Ancient Virus-Parasite Association, *Nucleic Acids Research*, Vol. 23, No. 12, pp. 2300-2304 (1995).
- 7) Hasegawa, M. and Kishino, H.: Confidence Limit on the Maximum-Likelihood Estimate of the Hominoid Tree from Mitochondrial-DNA Sequences, *Evolution*, Vol. 43, pp. 672-677 (1989).
- 8) Finden, C. and Gordon, A.: Obtaining Common Pruned Trees, *J. Classification* Vol. 2, pp. 255-276 (1985).
- 9) Waterman, M. S.: *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman & Hall, London (1995).
- 10) Hilditch, C. J.: Linner Skeletons from Square Cupboards, *Machine Intelligence 4* (Meltzer, B. (ed.)), University Press, Edinburge, pp. 403-420 (1969).
- 11) Hough, P.: Method and Means for Recognizing Complex Patterns, U.S. Patent 3,069,645 (1962).
- 12) The Newick Tree Format, PHYLIP document, available at:
<http://evolution.genetics.washington.edu/phylip/newicktree.html>.
- 13) Bray, T., Paoli, J. and Sperberg-McQueen, C.M. ed.: Extensible Markup Language (XML) 1.0, *W3C Recommendation 10-Feb-98*, available at <http://www.w3.org/TR/REC-xml>.