Machine Learning, 12, 69–87 (1993)

Discovery by Minimal Length Encoding: A Case Study in Molecular Evolution

ALEKSANDAR MILOSAVLJEVIĆ* MILOSAV@ANL.GOV
JERZY JURKA JURKA@JMULLINS@STANFORD.EDU
Linus Pauling Institute of Science and Medicine, 440 Page Mill Rd., Palo Alto, CA 94306

Abstract. We apply the Minimal Length Encoding Principle to formalize inference about the evolution of macromolecular sequences. The Principle is shown to imply a combination of Weighted Parsimony and Compatibility methods that have long been used by biologists because of their good practical performance. The background assumptions are expressed as an encoding scheme for the observed data and as heuristic rules for selection of diagnostic positions in the sequences. The Principle was applied to discover new subfamilies of Alu sequences, the most numerous family of repetitive DNA sequences in the human genome.

Keywords. Alu sequences, repetitive elements, molecular evolution, machine discovery, data compression.

1. Introduction

In medieval times scholastic philosophers argued that Nature always follows the simplest rules, and that scientists should consequently try to discover them. William of Occam, a fourteenth-century philosopher, opposed this metaphysical thesis about the simplicity of Nature as unnecessary. Instead of assuming that Nature is governed by simple laws, he proposed that preference for simpler hypotheses should be part of the scientific method, regardless of whether or not Nature indeed follows simple rules (Losee, 1980). This methodological principle is often referred to as "Occam's Razor Principle," the "Principle of Parsimony," or, most recently, the "Minimal Length Encoding Principle."

The modern algorithmic formulation of the Minimal Length Encoding Principle has almost concurrently been proposed by Solomonoff (1964), Kolmogorov (1968), and Chaitin (1966). Since the review of these most general formulations is beyond the scope of the present article, we refer the interested reader to the recent reviews by Vitanyi and Li (1989) and by Cover and Thomas (1991). Sober (1988) gives an extensive discussion of the Parsimony Principle in the context of evolutionary reconstructions. In the following we only illustrate the Principle by a historical example.

Consider the geocentric model of the solar system of Ptolemy versus Kepler's improved version of the Copernican heliocentric model (Kuhn, 1957). One way to explain why the heliocentric model is preferred is to apply the Minimal Length Encoding Principle. Toward that goal, assume that the model that reproduces the motions of the planets and the sun most exactly and that uses fewest numbers of parameters is preferred. For each planet, the geocentric model requires the description of a deferent (the main orbit) around the earth, plus the descriptions of a set of epicycles (minor orbits) to account for the nonuniformities in the apparent motion of planets around the earth. In contrast, the heliocentric model

*Current address: Genome Structure Group, Biological and Medical Research Division, Argonne National Laboratory, Argonne, IL 60439

of planetary motions based on the heliocentric model better fits the observed astronomical measurements. Thus, the heliocentric model reproduces the planetary motions more exactly and in fewer parameters, and thus it is preferred by the Minimal Length Encoding Principle to the geocentric model.

The current accumulation of information about the genetic DNA sequences of different organisms exceeds the volume of astronomical measurements that preceded modern astronomy and physics. Also, the living world appears to be much more "information rich" in the sense that it may not be possible to capture the phenomena of life in a few laws as simple as, say, the laws of mechanics. In order to discover patterns in the overwhelming amounts of genetic data, biologists are turning more and more toward formal methods of inference.

The long-term motivation behind this work is to show that the process of discovery in biology that is based on macromolecular genetic sequences can be viewed as Minimal Length Encoding of observations. In this sense, our work falls into the same category with other attempts at the application of Minimal Length Encoding in molecular evolution (Cheeseman & Kanefsky, 1990; Allison & Yee, 1990) and molecular biology (Babcock, Olson, & Pedrault, 1990; Jiang & Li, 1991; Konagaya & Yamanishi, 1991; Reichert, Cohen, & Wong, 1973; Jimenez-Montano, 1984).

In the present article we study the discovery of new subfamilies of so-called Alu sequences, the most numerous family of short interspersed repetitive DNA fragments that account for about 5%–10% of the human genome (Hwu, Roberts, Davidson, & Britten, 1986). The very presence of distinct subfamilies of Alu sequences, even though currently accepted (Willard, Nguyen, & Schmid, 1987; Britten, Baron, Stout, & Davidson, 1988; Quentin, 1988; Jurka & Smith, 1988; Jurka & Misolavljević, 1991), has until recently been disputed (Bains, 1986). Alu sequences are about 300 letters long (a typical Alu sequence is given in figure 2) and are present in several hundred thousand copies in the genome. Several hundred Alu sequences can be found in the current editions of DNA sequence databases. The function of Alu sequences is still unknown, and the understanding of their evolution may be the first step towards discovering it.

The current hypothesis is that most Alu sequences are *pseudo-genes* that have originated from one or more *master genes* (Jurka & Smith, 1988). An Alu pseudogene is a replica of an Alu master gene that is created through *transcription* of the DNA sequence of the master gene into an RNA sequence, subsequent *reverse transcripton* of the RNA sequence into a DNA sequence, and final *insertion* of the DNA sequence back into the genome. This three-stage process of transcription, reverse transcription, and insertion is also referred to as *retroposition.* According to the current hypothesis, the Alu pseudogenes that have originated from the same master gene during a short period of evolutionary time or from a subset of very similar master genes should be distinguishable from the rest of Alu pseudogenes by a set of common features.

A detailed discussion of the evolution of Alu sequences can be found in the companion paper by Jurka and Milosavljević (1991). In the present article, the emphasis is on the method of inference that was used to reconstruct the evolution of Alu sequences, a point not emphasized in the companion paper because of its biological audience. The biological discussions will be limited to what is necessary to understand the process of inference.

The key to the process of evolutionary reconstruction is the notion of *derived homology* (Ridley, 1986). For the purpose of a simplified example, consider the following eight hypothetical sequences of length 5 in the four-letter alphabet of DNA and ask the question about their possible common ancestor:

1 GAGCC
2 AAGCT
3 GGACC
4 TGGCT
5 ACCCT
6 GCGCG
7 GTGCC
8 ATGGT

The sequences are most familiar in the third and fourth positions (letters G and C, respectively). We may hypothesize that this similarity is due to their common ancestry. The similarity that is due to the inheritance of features from a hypothetical common ancestor is called *homology*.

By taking the majority letter in each positon (in case of a tie, the alphabetically first letter is taken). we may compute the *consensus sequence* GAGCT representing the hypothesized common ancestor. A total of 17 mutations can explain the differencs from the ancestor.

For the purpose of a simplified example, let us assume for a moment that the encoding length of the sample equals the number of letters to specify the ancestor plus the number of letters to specify the mutations. (This is an introductory oversimplification because we ignore for a moment the encoding of the positions that contain mutations and some other parts of an exact encoding). Under this assumption, the total encoding length would be $5 + 17 = 22$.

But upon closer inspection, one may observe that the sequences 1, 3, 6, and 7 almost always have letters G and C in the first and the last positions, respectively, while the sequences 2, 4, 5, and 8 almost always have letters A and T. To take advantage of this observation, we may postulate that the sequences from the first subfamily had an ancestor GAGCC, while the sequences from the second subfamily had an ancestor AAGCT.

The total number of differences of the sequences from their respective ancestors is now only 11. By adding the five letters to describe one of the ancestors and two letters to describe the differences of one ancestor from the other, we obtain a total of 18 letters to encode the whole sample. Note that this is four letters less than the 22 letters that were required initially.

A subset of sequences forms a *monophyletic group* relative to the complete set of sequences if all the members of the subset share a common ancestor that is not an ancestor of any other sequence from the complete set. For example, let us assume an evolutionary tree where the first ancestor (GAGCC) is at the root, and where its direct offsprings are the sequences from the first subfamily plus the second ancestor (AAGCT), and where the sequences from the second subfamily are direct offsprings of the second ancestor. In this case, the first subfamily would not be a monophyletic group while the second subfamily would.

The assumed tree implies that the first and last positions in the sequences from the monophyletic group contain letters that are changed in the lineage from the first to the second ancestor. Such positions are called *diagnostic* because they can be used to identify members of the monophyletic group. In general, the features that are changed in the lineage leading to an ancestor of a monophyletic group are called *derived homologies*, and they are the key to the process of evolutionary reconstruction.

An important problem is to decide whether a given family of sequences contains within it a monophyletic subfamily. To decide this, according to the Minimal Length Encoding Principle, we should compare the encoding length of the observed sequences with and without postulating a monophyletic subfamily. Toward that goal, in the next section we first propose a realistic encoding scheme.

2. The encoding scheme

In order to present the encoding scheme, we first introduce some notation. Let a *sample* S consist of m sequences $(S_1, \ldots, S_i, \ldots, S_m)$. Each sequence S_i consists of n letters. The letters come from a finite alphabet Γ of size $|\Gamma| = g$. In the case of DNA sequences, the alphabet is $\{A, G, C, T\}$, representing the four nucleotides that form DNA. By $\mathcal{S}(m, n, \Gamma)$ we denote the set of all samples consisting of m sequences of length n over alphabet Γ.

Due to evolutionary insertions and deletions of letters in some of the real DNA sequences, if we simply write the sequences underneath each other, the corresponding parts of different sequences will not appear in the same columns. To correct this, the corresponding parts are aligned by inserting the special *gap*-character " - ". While the process of alignment is part of an evolutionary reconstruction and should be incorporated into the inference program (Hein, 1990; Cheeseman & Kanefsky, 1990), in the present article we assume that the sequences were aligned in advance. We should also note that gap-characters were not treated as yet another letter in the alphabet. The gap-characters were treated as "unknown" letters, and they are implicitly replaced by any of the other four letters so that the overall encoding length is minimized.

To understand the discussion that follows, the knowledge of basic information theory is required (say, Cover & Thomas, 1991; Hamming, 1986).

An evolutionary model M for sample S has three parts, defined as follows:

Part (1): The encoding of the number m_1, which tells how many sequences are outside the monophyletic subfamily. If a single ancestor is postulated, then $m_1 = m$. This number is encoded in lg m bits. (Here and in what follows, we assume that information can be encoded in fractions of bits, so we write lg m instead of $\lceil \lg m \rceil$. In some cases, arithmetic coding (Bell, Cleary, & Witten, 1990) justifies this assumption in the limit, while in other cases we assume that the overall roundoff error is small.)

Part (2): The encoding of the number d, which tells how many diagnostic positions there are and the encoding of the location of individual diagnostic positions. This requires $(1 + d) * \lg n$ bits in case of two ancestors and 0 bits in case of a single ancestor.

Part (3): The encoding of the distribution of frequencies of letters in each position. A single frequency can be encoded in lg m bits. All but one frequency need to be encoded (the last frequency can be inferred based on all the other frequencies), and thus $(g - 1)$

lg m bits are required per distribution. Since the d diagnostic positions require separate encodings for each subfamily, a total of $(n + d)(g - 1)$ lg m bits is required.

The encoding length of the model can now be obtained by summing up the lengths of the three parts above.

$$I(M) = \begin{cases} \lg m + & n(g-1)\lg m & \textit{(single ancestor)} \\ \lg m + (1+d)\lg n + & (n+d)(g-1)\lg m & \textit{(two ancestors)} \end{cases}$$

In the example from the previous section, $m = 8$ and $n = 5$, so for a single-ancestor model, $I(M) = \lg 8 + 5(4 - 1) \lg 8 = 46$ bits, while for a two-ancestor model with $d = 2$ diagnostic positions, $I(M) = \lg 8 + (1 + 2) \lg 5 + (5 + 2)(4 - 1) \lg 8 \approx 73$ bits.

By $\mathfrak{M}$ we denote a class of evolutionary *models* that either may postulate a single ancestor for all the sequences from the sample or may postulate two ancestors, one of them being the ancestor of a monophyletic subfamily. The model does not specify which of the two subfamilies is monophyletic.

Note that each diagnostic position contributes to $I(M)$ because its location and two distinct distributions (instead of one for a nondiagnostic position) need to be encoded. Thus, it pays off to postulate diagnostic positions only if the increase in encoding length due to the more complex model M is less than the simultaneous decrease in the encoding length of the sample S relative to that model, which we discuss next.

The second part of the complete encoding is the encoding of the sequences from the sample S given a model M. To describe this encoding, we first introduce some additional notation. By $m_2 = m - m_1$ we denote the number of sequences in the putative monophyletic subfamily, and by $f_1' = m_1/m$ and $f_2' = m_2/m$ we denote the corresponding frequencies. Let $f_l''(x)$ denote the frequency of occurrence of the letter $x \in \Gamma$ in the $l - th$ position among all the sequences in the sample. Also, let $f_{j,l}''(x)$ denote the frequency of occurrence of the letter $x \in \Gamma$ in the $l - th$ position among the members of the $j - th$ subfamily ($j = 1, 2$). The encoding of the sample then consists of the following two parts.

1. The encoding of the subfamily membership of the individual sequences. This requires $m H'$ bits, where H' is the entropy function of the subfamily membership; in other words, $H' = -f_1' \lg f_1' - f_2' \lg f_2'$.
2. The encoding of the letters in all the positions. This requires $m\Sigma_{l \text{ not diagnostic}} H_l'' + \Sigma_{j=1,2} m_j \Sigma_{l \text{ diagnostic}} H_{j,l}''$ bits, where $H_l'' = \Sigma_{x\in\Gamma} - f_l''(x) \lg f_l''(x)$ and $H_{j,l}'' = \Sigma_{x\in\Gamma} - f_{j,l}''(x) \lg f_{j,l}''(x)$.

The encoding length of a sample relative to a model can now be obtained by summing up the lengths of the two parts above.

$$I(S|M) = m H' + m \sum_{l \text{ not diagnostic}} H_l'' + \sum_{j=1,2} m_j \sum_{l \text{ diagnostic}} H_{j,l}'' \qquad (2)$$

Continuing our example, $H' = 0$ for a single-ancestor model and $H' = 1$ for the two-ancestor model. Also, $H_1'' \approx 1.4$, $H_{1,1}'' = 0$, $H_{2,1}'' \approx 0.81$, $H_2'' = 2$, $H_3'' \approx 1.06$, $H_4'' \approx 0.54$, $H_5'' \approx 1.4$, $H_{1,5}'' \approx 0.81$, and $H_{2,5}'' = 0$. After some addition and multiplication we obtain that, for a single-ancestor model, $I(S|M) \approx 51$ bits, while for a two-ancestor model $I(S|M) \approx 43$ bits.

The total encoding length of a sample S, as encoded by model M, is the sum

$$I(S, M) = I(M) + I(S|M), \tag{3}$$

where $I(M)$ and $I(S|M)$ are given by (1) and (2), respectively.

Given sample S from $\mathcal{S}(m, n, \Gamma)$ we consider the problem of discovering a model M from $\mathfrak{M}$ that minimizes $I(S, M)$ given by (3). Continuing our example, the approximate encoding length for the single-ancestor model is 46 + 51 = 97 bits, while for the two-ancestor model, it is 73 + 43 = 116 bits. Thus, the single-ancestor model is preferred.

Recall that in the introductory examle the two-ancestor model was preferred. This was due to an oversimplification in the measurement of the encoding length (e.g., the bits needed to encode the locations of diagnostic differences were ignored).

3. Computing the optimal model

The discussion above leads to the following problem: for a given sample S, find the model M_{opt} that minimizes the encoding length $I(S, M)$ given by (3). Even under certain simplifying assumptions, this problem remains NP-hard (Milosavljević, 1990).

Since the globally optimal model may be hard to find, we propose a local search algorithm (figure 1). Local search trials are repeated t times. An individual trial starts from a random partition of m sequences into two subfamilies. The putative diagnostic positions are then selected (as described below), and the encoding length is computed. The algorithm then cycles through the sequences searching for a sequence such that, if it is moved from one

```
procedure MASC (sample S, number of trials t)
  I_opt ← ∞;
  repeat t times
    pick a random initial split of S into two subfamilies;
    choose the diagnostic positions;
    compute I(S, M) by (3);
    while there is a move that decreases I(S, M)
      perform the move;
      choose the diagnostic positions;
      compute I(S, M) by (3);
      if ( I(S, M) < I_opt )
        M_opt ← M;
        I_opt ← I(S, M_opt);
  return M_opt;
```

Figure 1. A high-level description of the procedure MASC (Multiple Aligned Sequence Classification). The number of local search trials is denoted by t, the current optimal model by M_{opt}, and the current minimal encoding length by I_{opt}.

subfamily into the other, the encoding length decreases. If no such sequence exists, the model corresponds to a local minimum. The algorithm returns the model that corresponds to the best found local minimum.

In order to present the heuristic rule for the selection of diagnostic positions for Alu sequences, we first introduce some additional notation. Let *majority*(j, l) denote the majority letter in the $l - th$ position among the sequences from the $j - th$ subfamily (ties resolved by taking the alphabetically lowest letter). In the example from the introduction, for the two-ancestor model, *majority*(1, 1) = G, *majority*(2, 1) = A, *majority*(1, 2) = A, *majority*(2, 2) = A, and so on.

A position l is selected to be diagnostic under the following conditions:

```
diagnostic(R) if
  (majority(1,R)) ≠ majority(2,R) and
  ((majority(1,R) ≠ ' - ') and (majority(2,R) ≠ ' - ')) and
  (not CpG(R))
```

The first condition eliminates the positions that are poor candidates to be diagnostic: if the majority base is the same across the subfamilies, the introduction of the separate encoding schemes for individual subfamilies is not likely to pay off. The second condition eliminates the positions that do not contain sufficient information. The third condition is aimed at eliminating the positions called "CpG" (Watson, 1987) that are known to mutate rapidly: a "C", if followed by "G", will tend to be replaced by "T", while the "G" will tend to be replaced by "A". The "CpG" positions may give rise to dependencies across positions that are not due to the evolution of Alu sequences, and they must be eliminated. The following definition of a "CpG" position was applied.

```
CpG(R) if for both j=1 and j=2 the following is true
  ((majority(j,R) = T) and (majority(j,R+1) = G)) or
  ((majority(j,R) = C ) and (majority(j,R+1) = G or A)) or
  ((majority(j,R) = A) and (majority(j,R-1) = C)) or
  ((majority(j,R) = G ) and (majority(j,R-1) = C or T))
```

4. The prediction test

The predictive power of biological theories has been formulated by Ernst Mayr (1961): "If I have identified a fruit fly as an individual of *Drosophila melanogaster* on the basis of bristle pattern and the proportions of face and eye, I can "predict" numerous structural and behavioral characteristics which I will find if I study other aspects of this individual. If I find new species with the diagnostic key characters of the genus *Drosophila*, I can at once "predict" a whole set of biological properties."

In the following, we directly apply this idea to test the predictive power of the newly discovered subfamilies of macromolecular sequences. We assume that two subfamilies, one monophyletic and the other consisting of the rest of the sequences, have been proposed based on an initial set of sequences and that their predictive power is tested on a new set.

Assume that the sequences from the test set are presented to us with one of their diagnostic positions blanked out. Let our goal be to guess the hidden letters. Our strategy would be to predict based on the knowledge of the subfamilies and the frequencies of letters in the diagnostic position across the subfamilies. A strategy that ignores subfamilies would predict based only on the frequencies of letters in that position. In both strategies the frequencies of letters are obtained based on the initial set.

Our strategy is to first identify the subfamily membership of the sequences from the test set based on all the letters except the hidden one and then to guess that the hidden letter is the majority letter for the guessed subfamily. If the subfamilies were ignored, the best one could do is to always guess based on the majority for the whole family. To compare the two guessing strategies, assume the following letter counts in the hidden position of the new sequences:

	T	C	A	G	
subfamily 1	14	0	1	0	
subfamily 2	4	2	36	0	
total	18	2	37	0	57

Based on two subfamilies and assuming that the majority letters in the initial set and the test set are the same, we would have correctly guessed $c = 14 + 36 = 50$ out of $m = 57$ letters. If the subfamilies were ignored the probability of success would be $s = 37/57$. In that case, each of our $m = 57$ guesses would be a Bernoulli trial with probability of success $s = 37/57$. But, by binomial distribution, the probability of having succeeded 50 or more times would be only $p = 0.0001$. Such a low probability strongly supports the presence of the postulated subfamilies.

Since the value of p is typically very small, we instead use $w = -\log p$, where the logarithm is base 10; for example $w = -\log 0.0001 = 4$. This quantity we use not only in the prediction test, but also in a post-hoc way to measure how diagnostic a position is for a particular split of a family into subfamilies.

5. Discovering Alu subfamilies

The Alu sequences were extracted from the GenBank DNA sequence database by applying the same programs that were recently used to extract L1 sequences (Jurka, 1989). Three sets of Alu sequences were extracted. The first set consisted of 125 Alu sequences from the Release 46.0 of the GenBank database (this is the same set that was used to infer the two major Alu subfamilies (Jurka & Smith, 1988). The second set consisted of 259 Alu sequences from the Release 55.0 of GenBank. The third set consisted of 369 sequences that were present in the Release 63.0 but absent from the Release 55.0.

Following their extraction, the sequences within each set were aligned against the Alu consensus sequence (Jurka & Smith, 1988) (figure 2) by the alignment algorithm of Smith and Waterman (1981). Finally, the sequences within each set were mutually aligned by an algorithm that combines their pairwise alignments against the consensus into a multiple alignment. In the process of multiple alignment, the sites of homologous insertions (fragments that are inserted in only a small set of possibly related sequences) were not

```
GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACCTGAGG
^1       ^10       ^20       ^30       ^40       ^50       ^60       ^70
TCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCG
         ^80       ^90       ^100      ^110      ^120      ^130      ^140
GGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGG
         ^150      ^160      ^170      ^180      ^190      ^200      ^210
AGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTC
         ^220      ^230      ^240      ^250      ^260      ^270      ^280
TCAAAAAAAA
         ^290
```

Figure 2. The Alu consensus sequence.

aligned well because of their absence in the consensus, so they were skipped. (It should be noted that an alignment represents an implicit reconstruction of insertion and deletion events during evolution and that no alignment is absolutely certain.)

There were three stages of experimentation, each using the respective set of Alu sequences. In the first stage, based on the first set of Alu sequences, the inference method presented here and a more standard method were tested on the problem of reproducing the discovery of known Alu subfamilies (Jurka & Smith, 1988). In the second stage, the second set of Alu sequences was used to rediscover the old subfamilies and to discover some new ones. In the third stage, the third set of Alu sequences was used to test the predictive power of the Alu subfamilies that were discovered in the second stage.

In the first stage, a number of experiments were tried where the encoding length was computed as in (4) (see next section). As explained in the next section, this measure is obtained by introducing the assumptions that are implicit in Weighted Parsimony and Compatibility methods. But this measure was not as good for rediscovering the Alu subfamilies (Jurka & Smith, 1988) as the more general formulation given by (3), so we omit the details of these experiments. In addition to the implicit and unjustified assumptions, the presence of "CpG" positions may also have contributed to the failure to rediscover known Alu subfamilies by this more standard method. Thus, in the rest of the experiments the encoding length was computed by (3).

In the second stage, the initial sample was partitioned into a set of most specific subfamilies by performing a series of binary splits, as illustrated in figure 3. To perform a single split, local search (figure 1) was repeated $t = 100$ times, each time starting from a new random split of a set of Alu sequences into two subsets. The splitting procedure was then repeated on the newly obtained subsets.

The problem with the approach outlined above was to decide when to stop the splitting procedure. If the Minimal Length Encoding criterion was strictly applied, the splitting procedure would stop whenever the single-ancestor encoding length is less than the lowest found two-ancestor encoding length. If this strict criterion was applied, only three subfamilies—Alu-J, the union of Alu-Sb and Alu-Sc, and the union of Alu-Sp, Alu-Sq, Alu-Sr, and Alu-Sx—would be found. Thus, the criterion was relaxed, and the splitting procedure was continued even in cases where the best found two-ancestor model gave longer encoding length than the single-ancestor model, provided multiple diagnostic positions could be identified. The diagnostic positions were then used in the third stage to justify the splits by a prediction test. The computations of the particular splits are summarized in table 1.

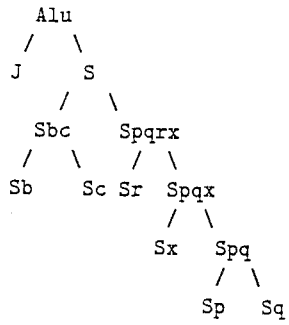

Figure 3. The process of discovery of subfamilies of Alu sequences. Each branching point in this binary tree denotes a discovery of a split of a set of sequences into two subsets of sequences, each consisting of one or more subfamilies.

Table 1. The search for splits of Alu sequences into subfamilies.

Split into subfamilies	Number of seq.	CPU seconds per search	Frequency of finding best minima	Number of diag. pos.	Bit-length for best local minima	
					Two-anc.	Single-anc.
J/S	259	168	100%	13	48958	49420
Sbc/Spqrx	208	230	82%	4	36552	36600
Sb/Sc	54	37	65%	3–4	11323–11325	11273
Sr/Spqx	154	119	11%	2	28651–28666	28624
Sx/Spq	63	39	29%	2	14030–14064	13971
Sp/Sq	38	22	33%	2	9594	9544

The program was implemented in C++ and was run on a Sun SPARCStation 330. The leftmost column indicates the splits into subfamilies; e.g., Sx/Spq denotes the split of the set containing subfamilies Sx, Sp, and Sq into two subsets, one containing subfamily Sx, and the other containing subfamilies Sp and Sq. For the split Sb/Sc there were two very close local minima, one of them having an additional diagnostic position, this is why the number of diagnostic positions is denoted by the range 3–4. Some of the single-ancestor models have a shorter encoding length than the best local minimum for two subfamilies—the presence of subfamilies in such cases was proven by prediction.

In the third stage, for each split of the sequences from the second set, the positions were assigned weights w as described in the previous section. The position with the highest weight was then identified and blanked out in the sequences from the third set. The subfamily membership of the sequences from the third set was the identified by running MASC on the mixture of sequences from the second and third sets, with the subfamily membership of sequences from the second set fixed and with the single position blanked out. For each sequence from the third set, the letter in the hidden position was then guessed based on its subfamily membership. The probability of success in guessing the hidden letters was then computed (table 2), as described in the previous section. Based on the results of the prediction test, we conclude that all the splits are strongly supported, except the split Sp/Sq.

Table 2. The splits of Alu sequences into subfamilies and the letter counts in the hidden positions for the third (test) set.

Split and position	T	C	A	G	-	T	C	A	G	-	*w*
J/S 94	1	12	5	*49	14	12	*242	5	14	15	6.1
Sbc/Spqrx 78	4	2	*45	0	3	*198	5	8	3	21	10.4
Sb/Sc 219	1	*31	0	3	0	0	4	3	*19	0	4.4
Sr/Spqx 154	1	0	9	*109	0	2	2	*74	28	3	10.5
Sx/Spq 272	0	1	1	*38	12	1	0	*21	9	2	2.9
Sp/Sq 244	*16	3	0	0	1	8	*0	0	0	0	0.2

The letter counts for both subfamilies are given for each split. For example, the first row indicates the letter counts in the position 94 for the set of sequences consisting of subfamilies J and S. The sequences were first classified as J or S based on all the positions except position 94. If a sequence was classified as J, it was guessed that the letter in position 94 is G, while if it was classified as S it was guessed that the letter is C. The guessed letters are indicated by asterisks. The total number of correct guesses for subfamily J was 49, while for subfamily S it was 242. The last column indicates that the probability of making 49 + 242 or more correct guesses if the subfamilies were ignored is at most $10^{-6.1}$.

But the independent evidence based on a diagnostic insertion after positon 264 in the sequences from the subfamily Sp suggests that this subfamily indeed exists (the diagnostic insertions and deletions cannot be discovered by the current approach because the sequences are aligned in advance).

Table 3 gives a brief summary of the Alu subfamilies that have been proposed so far. Note that the "Major" subfamily of Willard et al. (1987) corresponds to two subfamilies, II and III, of Britten et al. (1988), to three subfamilies, Alu-Sc, -Sd, and -Se, of Jurka and Smith (1988), and to three subfamilies, C, E, F, of Quentin (1988); all the smaller subfamilies that occur in identical columns are approximately equivalent.

The main discovery performed by MASC is that subfamily Alu-Sd is not homogeneous and that it contains at least three subfamilies, Alu-Sp, Alu-Sq, and Alu-Sr. In table 3, note that the subfamily Alu-Sq, discovered by MASC Version 0.6 (Jurka & Milosavljević, 1991), corresponds to two subfamilies, Alu-Sr and Alu-Sq, that were discovered by MASC Version

Table 3. The discoveries of Alu subfamilies.

Reference	Alu subfamilies						
Willard et al. (1987)	Diverged	Major					Conserved
Britten et al. (1988)	I	II				III	IV
Jurka & Smith (1988)	J	Sa				Sc	Sb
		Se	Sd				
Quentin (1988)	D	F	E			C	A
MASC 0.6 (Jurka & Milosavljević 1991)	J	Sx	Sq		Sp	Sc	Sb
MASC 0.7 (this article)	J	Sx	Sr	Sq	Sp	Sc	Sb

0.7. Since both MASC 0.6 and MASC 0.7 were run on the same set of Alu sequences, the discovery of the new subfamily *Sr* was only due to the improvements in the encoding scheme for two-ancestor models that were suggested by Peter Cheeseman in his review of the present article.

The discovery of the subfamily *Sp* has an interesting evolutionary consequence. A number of sequences from this subfamily turned out to be younger (measured by the degree of decay due to random mutations) than the members of the subfamily *Sc*. Since the lineage consisting of subfamilies *Sb* and *Sc* is generally younger than the lineage consisting of subfamilies *Sp*, *Sq*, and *Sr*, this indicates that there were at least two Alu genes that were being retroposed during overlapping periods of time. This important point is discussed in more detail in the companion biological paper (Jurka & Smith, 1988).

6. Relation to other methods

In this section we discuss the relation of the proposed method for the discovery of evolutionary relationships to other methods of evolutionary reconstruction and to the methods of categorization in general.

6.1. Weighted Parsimony and Compatibility methods

We next present a short overview of Weighted Parsimony and Compatibility position-weighting methods and show how a combination of these methods can be derived from the Minimal Length Encoding Principle. While Weighted Parsimony and Compatibility methods are typically used to infer a complete evolutionary branching pattern, for simplicity we here present the methods only as they pertain to the inference of a single monophyletic subfamily.

We will first have to introduce some additional notation. Let $\Delta_{j,l} = m_j - majority\ (j, l)$ denote the number of differences from *majority* (j, l) within the $j - th$ subfamily in the $l - th$ position. In addition, let $\Delta_l = \Delta_{1,l} + \Delta_{2,l}$, denote the number of differences in the $l - th$ position across the subfamilies. The simplest Parsimony criterion minimizes the number of differences across the positions

$$\Delta = \sum_{l=1}^{n} \Delta_l.$$

The Compatibility criterion (Felsenstein, 1981) is based on the assumption that if the probability of change varies across positions, then the positions that have low probability of change and are most compatible with the putative monophyletic subfamily should have the highest weight. If the probability of change in the l–th position is estimated by the frequency $f_l = \Delta_l/m_l$, then we can define the indicator of compatibility of the l–th position by

$$compatible_l = \begin{cases} 1 & \text{if } f_l \leq threshold \\ 0 & \text{if } f_l > threshold \end{cases}$$

where the value of the threshold is chosen a priori. The Parsimony criterion, where the positions are weighted by their compatibility, is then given by

$$\sum_{l=1}^{n} compatible_l \, \Delta_l.$$

Continuing our example from the introduction, if we choose *threshold* = 0.25, all the positions except position 2 are compatible with the proposed two-ancestor model. Thus, the value of the criterion is 5, and is obtained by summing the numbers of mutations in all the positions except the position 2.

In place of the threshold function for the compatibility indicator $compatible_l$, Farris (1969) experimented with continuous weighting functions w that are monotonically decreasing on the interval (0, 1]. The weighted parsimony criterion

$$\sum_{l=1}^{n} w(f_l)\Delta_l$$

was used with linear, convex, concave bounded and concave unbounded weighting functions w. The best agreement with the evolutionary trees that were independently proposed by biologists was obtained for concave unbounded weighting functions of the form $w(x) = ((x)^{-c} - 1)$, where c is positive.

In contrast to Weighted Parsimony, which minimizes the number of changes of values of reliable positions, Compatibility methods (LeQuesne, 1969) minimize the total number of incompatible positions as given by the following:

$$\Lambda = \sum_{l=1}^{n} (1 - compatible_l).$$

Continuing our example, we may verify that in the proposed two-ancestor model, only the position 2 is incompatible. Thus, the value of the compatibility criterion is $\Lambda = 1$. Also, one can easily verify that no other two-ancestor model can do better.

For the purpose of comparison, let us now turn to the Minimal Length Encoding method. Let us apply the Minimal Length Encoding principle under the following two simplifying assumptions: 1) all the positions are diagnostic; 2) all the letters that are not majority within a subfamily are encoded using the same number of bits. Under these two assumptions, the formula (2) can be rewritten as follows (for a detailed derivation, see Milosavljević, 1990):

$$I(S|M) = m\,H' + \sum_{j=1}^{k}\sum_{l=1}^{n} \Delta_{j,l} \log \alpha_{j,l} + \sum_{j=1}^{k} m_j \sum_{l=1}^{n} \log \beta_{j,l} \tag{4}$$

where $\alpha_{j,l} = (1/f_{j,l} - 1)(g - 1)$ and $\beta_{j,l} = 1/(1 - f_{j,l})$ and where $f_{j,l} = \Delta_{j,l}/m_j$ denotes the frequency of differences from the majority letter in the $l - th$ position among the sequences from the $j - th$ subfamily.

The Weighted Parsimony method minimizes only the second term in (4); the "weights" $\log \alpha_{j,l}$ are concave and unbounded in $f_{m,l}$ as suggested to be the best by the experimental results of Farris that were discussed above. Similarly, the third term corresponds to a weighted version of the Compatibility method. It is interesting to mention at this point that the minimization of the sum of the last two terms of (4) has also been independently proposed in phytosociology (Orloci, 1968).

6.2 Bayesian and maximum likelihood methods

Another way to define $I(M)$ and $I(S|M)$ would be to postulate two probabilistic processes; the first process generates a probabilistic model M, which in turn probabilistically generates a sample S. Let $P(M)$ denote the a priori probability of a model M, and let $P(S|M)$ denote the probability of S given M. The details of this approach are explained elsewhere (Milosavljević, Haussler, & Jurka, 1989, Milosavljević, 1990); in the following we only sketch the main relationships between encoding lengths and probabilities. Ignoring additive constants, the relationship between encoding lengths and probabilities for an optimal encoding scheme are as follows (e.g., Cover & Thomas, 1991):

$$I(M) = -\log P(M) \tag{5}$$

and

$$I(S|M) = -\log P(S|M). \tag{6}$$

From (5) it follows that the lower the complexity $I(M)$ of a model M, the higher its a prior probability $P(M)$; in other words, simpler models are more likely a priori. From (6) it follows that the smaller the complexity $I(S|M)$ the larger the likelihood $P(S|M)$ of the model M; in other words, the model that fits the data better is more likely.

By using (5) and (6), we obtain

$$I(S, M) = I(M) + I(S|M) = -\log(P(M)P(S|M)) = -\log P(S, M), \tag{7}$$

where $P(S, M)$ is the joint probability of S and M. Hence, the model that minimizes the total complexity $I(S, M)$ is the same one that maximizes the joint probability $P(S, M)$. Since $P(S, M) = P(M|S)P(S)$, where $P(S)$ is fixed, the same model also maximizes $P(M|S)$, the a posteriori probability of M given S. Hence, this method of defining $I(M)$ and $I(S|M)$ reduces the Minimum Length Encoding method to the standard Bayesian method (Cheeseman et al., 1988). If we define $I(M) = 0$, then we obtain the Maximum Likelihood

method (Duda & Hart, 1973). A detailed probabilistic interpretation of Minimal Length Encoding in the context of evolutionary reconstructions can be found elsewhere (Milosavljević, 1990).

AutoClass (Cheeseman et al., 1988), a program that implements Bayesian technique for unsupervised classification, has been applied to the first set of 125 Alu sequences that were discussed above. This generic program proposed nine classes of Alu sequences that did not overlap very well with the subfamilies accepted by biologists. An obvious problem was that the correlated occurrences of letters in the *CpG* positions have misled AutoClass to suggest subfamilies where there were none and to overlook the existing ones. The problem is that AutoClass assumes that mutations are independent across positions, which is not true of *CpG* positions in Alu sequences.

6.3. Other minimal length encoding methods

The Minimum Message Length approach to classification, pioneered by Wallace and Boulton (1968), provides a principled way of choosing the complexity of the inferred model by minimizing the combined encoding length of the model and the data given the model. While the most recent extensions of this work (Wallace, 1990) are very similar to the approach presented here, there are a few important differences. One important difference is that, in addition to a theoretically sound optimization criterion (Wallace, 1990) we also present the algorithm for the minimization of encoding length with a verified performance. The combinatorial problem of finding an optimal split of a family of sequences into two subfamilies at a time, as performed by MASC, not only has a valid biological justification but is also likely to be easier (Milosavljević, 1990) than the problem of finding an optimal set of arbitrary many subfamilies at once, as suggested by Wallace and Boulton (1968). Another difference is that MASC dynamically selects diagnostic positions during the search for an optimal model. Our experiments (the first kind of experiments from the previous section) indicate that the selection of diagnostic positions performed by MASC is necessary for reproducing the discovery of Alu subfamilies, particularly because of the elimination of mutations that are due to *CpG* noise.

6.4. Conceptual clustering

Although the present article deals with the methods for inference of evolutionary relationships, there are some striking similarities to conceptual clustering (Michalski & Stepp, 1983; Gennari, Langley, & Fisher, 1989).

As pointed out by Michalski and Stepp (1983), not all attributes are considered equally relevant in the process of categorization. Indeed, when comparing it to Weighted Parsimony, we have shown that the Minimal Length Encoding approach implicitly weighs certain positions more and the others less. An important difference between MASC and CLUSTER/2, the program implemented by Michalski and Stepp (1983), is that the latter defines clusters by logical formulae, an approach certainly not applicable in case of macromolecular sequences, where the boundaries between different subfamilies of sequences are not sharply defined.

Fisher's COBWEB (1987) constructs clusters that maximize *category utility*, a combined criterion that measures both the *predictiveness* of category membership based on attribute values and the *predictability* of attribute values based on category membership. We may recall that the prediction test that we have employed to prove the existence of the subfamilies of Alu sequences uses a combination of the predictiveness property (when subfamily membership of a sequence is guessed) and the predictability property (when the hidden letter is guessed). The subfamilies of Alu sequences turned out to have predictive value even though the category utility criterion was not explicitly applied in the process of inference. Indeed, it is widely recognized in the domain of data compression that the data that can be well compressed can also be well predicted, and vice versa (Bell, Cleary, & Witter, 1990). Thus, the minimization of encoding length may be viewed as an implicit maximization of predictive power of the inferred models.

The approach to concept formation implemented in COBWEB (Fisher, 1987) is incremental in the sense that the extensive reprocessing of the already presented data is not allowed. In the case of Alu sequences, this would certainly create a problem because subfamilies of Alu sequences can be inferred only after a total of more than 20–30 Alu sequences are observed. Thus, the subfamilies that are inferred based on a small sample may have to be completely revised at the point when the sample becomes sufficiently large. An additional problem is that COBWEB does not provide for elimination of correlated attributes, a necessary feature in case of the presence of *CpG* positions.

7. Conclusion

The rules embodied in biological Weighted Parsimony (Farris, 1969; Felsenstein, 1981) and Compatibility (LeQuesne, 1969) methods for evolutionary reconstructions that have long been used by biologists were shown to follow from the more general Minimal Length Encoding Principle. The Principle was applied to discover new subfamilies of Alu sequences (table 3, Jurka & Milosavljević, 1991) that for the first time provide evidence for coexistence of multiple retropositionally active Alu genes.

The process of discovery consists of two main phases. In the first phase, the subfamilies of Alu sequences are computed by performing binary splits of the sample (figure 3). In the second phase, the proposed subfamilies are tested for their predictive power on new data.

An important feature of the proposed heuristic algorithm is the dynamic selection of diagnostic positions. The positions that are known to contain correlated mutations due to the "CpG" noise are eliminated by a heuristic rule. The dynamic selection of positions is easily accommodated within the Minimal Length Encoding framework. The importance of the elimination of correlations has been emphasized by Felsenstein (1982):

> A major assumption of all parsimony methods is that the characters evolve independently. There has been no attempt to provide methods of detecting character correlation and removing its effects from the analysis.... How to develop methods to remove the effects of character correlation is perhaps the most important unsolved problem facing phylogenetic inference. The absence of a solution is the greatest weakness of existing methods.

A drawback of the method is that it tends to underestimate the number of subfamilies if the Minimal Length Encoding Principle is applied strictly. The predictive test does eliminate this problem, but it also requires new data for testing. It would be desirable to improve the method and eliminate the need for new data. Two improvements may be tried. The first possible improvement is to introduce the uncertainty in sequence assignment to subfamilies, as suggested by Wallace (1990) and Cheeseman et al. (1988). This may help because the exact assignment of a sequence to a particular subfamily is an excess of information if the sequence is not clearly a member of either of the subfamilies. The second possible improvement is to use fewer bits to encode the frequency distributions of letters in the particular positions. This may help because the exact specification of a distribution may not pay off in terms of a reduction in the overall encoding length.

An important extension would be to remove the assumption that the sequences are aligned in advance. While the insertions and deletions may provide invaluable evolutionary evidence (e.g., in the case of diagnostic insertions that prove the split into Alu subfamilies Sp and Sq), some previous attempts (Cheeseman & Kanefsky, 1990) indicate that the resulting search problem may be very hard. One possible approach would be to alternate the alignment and splitting steps in an iterative algorithm, thus replacing a hard combinatorial problem by two easier ones.

8. Acknowledgments

David Haussler supervised and contributed to this work in its earlier stages. This work was supported by the ONR grant N 00014-86-K-0454 while the first author was at the Baskin Center for Computer and Information Sciences at the University of California at Santa Cruz. The final stages of this research have been supported in part by the DOE grant DE-FG03-91ER61153. Peter Cheeseman's suggestions led to the improvements of the MASC program that resulted in the discovery of the subfamily Alu-Sr. We thank Wray Buntine and Jan Zytkow for their valuable comments and patient reviews of this work.

References

Allison, L., & Yee, C.N. (1990). Minimum message length encoding and the comparison of macromolecules. *Bulletin of Mathematical Biology, 52*, 431–453.

Babcock, Marla S., Olson, Wilma K., & Pednault, Edwin P.D. (1990). The use of the minimum description length principle to segment dna into structural and functional domains. In *Working Notes, AAAI Spring Symposium Series*, Stanford.

Bains, W. (1986). The multiple origins of human Alu sequences. *Journal of Molecular Evolution, 23*, 189–199.

Bell, T.C., Cleary, J.G., & Witten, I.H. (1990). *Text compression*. Englewood Cliffs, NJ: Prentice Hall.

Britten, R.J., Baron, W.F., Stout, D., & Davidson, E.H. (1988). Sources and evolution of human Alu repeated sequences. *Proceedings of the National Academy of Sciences of the United States of America, 85*, 4770–4774.

Chaitin, G.J., (1966). On the length of programs for computing finite binary sequences. *Journal of the Association for Computing Machinery, 13*, 547–569.

Cheeseman, P., Self, M., Kelly, J., Taylor, W., Freeman, D., & Stutz, J. (1988). Bayesian classification. In *Proceedings of the Conference of the American Association for Artificial Intelligence*. Los Altos, CA: Morgan Kaufmann.

Cheeseman, Peter, & Kanefsky, Bob. (1990). Evolutionary tree reconstruction. In *Working Notes, AAAI Spring Symposium Series*, Stanford.

Cover, Thomas & Thomas, Joy. (1991). *Elements of information theory.* New York: Wiley.

Duda, R.O., & Hart, P.E., (1973). *Pattern recognition and scene analysis.* New York: Wiley.

Farris, J.S. (1969). A successive approximations approach to character weighting. *Systematics and Zoology, 18,* 374–385.

Felsenstein, J. (1981). A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society, 16,* 183–196.

Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology, 57*(4), 379–404.

Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning, 2,* 139–172.

Gennari, J.H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence, 40,* 11–61.

Hamming, R.W. (1986). *Coding and information theory.* Englewood Cliffs, NJ: Prentice-Hall.

Hein, Jotun. (1990). Unified approach to alignment and phylogenies. *Methods of Enzymology, 183,* 626–645.

Hwu, H.R., Roberts, J.W., Davidson, E.H., & Britten, R.J. (1986). Insertion and/or deletion of many repeated dna sequences in human and higher ape evolution. *Proceedings of the National Academy of Sciences of the United States of America, 83,* 3875–3879.

Jiang, Tao, & Ming, Li, (1991). On the complexity of learning strings and sequences. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory* (pp. 367–371). San Mateo, CA: Morgan Kaufmann.

Jimenez-Montano, M.A. (1984). On the syntactic structure of protein sequences and the concept of grammar complexity. *Bulletin of Mathematical Biology, 46,* 641–659.

Jurka, J. (1989). Subfamily structure and evolution of the human L1 family of repetitive sequences. *Journal of Molecular Evolution, 29,* 496–503.

Jurka, J. & Milosavljević (1991). Reconstruction and analysis of human Alu genes. *Journal of Molecular Evolution, 32,* 105–121.

Jurka, J. & Smith, T. (1988). A fundamental division in the Alu family of repeated sequences. *Proceedings of the National Academy of Sciences of the United States of America, 85,* 4775–4778.

Kolmogorov, A.N. (1968). Three approaches to the quantiative definition of information. *International Journal for Computer Mathematics, 2,* 157–168.

Konagaya, Akihiko, & Yamanishi, Kenji. (1991). Stochastic decision predicates: A scheme to represent motifs. In *AAAI Workshop on AI Applications to Classification and Pattern Recognition in Molecular Biology,* Anaheim, California.

Kuhn, T.S. (1957). *The Copernican revolution.* Cambridge, MA; Harvard University Press.

LeQuesne, W.J. (1969). A method of selection of characters in numerical taxonomy. *Systematic Zoology, 18,* 201.

Losee, J. (1980). *A historical introduction to the philosophy of science.* Oxford: Oxford University Press.

Mayr, Ernst. (1961). Cause and effect in biology. *Science, 134,* 1501–1506.

Michalski, R.S., & Stepp, R.E., (1983). Automated construction of classifications: Conceptual clustering versus numerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5,* 396–410.

Milosavljević, Aleksandar. (1990). *Categorization of macromolecular sequences by minimal length encoding.* Ph.D. thesis, Computer Science Department, University of California at Santa Cruz.

Milosavljević, Aleksandar, Haussler, David, & Jurka, Jerzy. (1989). Informed parsimonious inference of prototypical genetic sequences. *Proceedings of the Second Workshop on Computational Learning Theory* (pp. 102–117). San Mateo, CA: Morgan Kaufmann.

Orloci, Laszlo. (1968). Information analysis in phytosociology: Partition, classification and prediction. *Journal of Theoretical Biology, 20,* 271–284.

Quentin, Y., (1988). The Alu family developed through successive waves of fixation closely connected with primate lineage history. *Journal of Molecular Evolution, 27,* 194–202.

Reichert, T.A., Cohen, D.N., & Wong, K.C. (1973). An application of information theory to genetic mutations and the matching of polypeptide sequences. *Journal of Theoretical Biology, 42,* 245–261.

Ridley, M. (1986). *Evolution and classification.* London and New York: Longman.

Smith, T.F., & Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology, 147,* 195–197.

Sober, E., (1988). *Reconstructing the past: Parsimony, evolution, and inference.* Cambridge, MA: MIT Press.
Solomonoff, R.J. (1964). A formal theory of inductive inference, Part I. *Information and Control, 7,* 1–22.
Vitanyi, P.M.B. & Li, M. Kolmogorov complexity and its applications. (Tehnical Report CS-R8901). Amsterdam: Centre for Mathematics and Computer Science, Amsterdam University.
Wallace, C.S. (1990). Classification by minimum-message-length inference. In *Working Notes, AAAI Spring Symposium on the Theory and Application of Minimal-Length Encoding.*
Wallace, C.S., & Boulton, D.M. (1968). An information measure for classification. *Computer Journal, 11,* 185–195.
Watson, J.D. (1987). *Molecular Biology of the Gene.* Reading, MA: Benjamin/Cummings.
Willard, C., Nguyen, H.T. & Schmid, C.W. (1987). Existence of at least three distinct Alu subfamilies. *Journal of Molecular Evolution, 26,* 180–186.

Received November 28, 1990
Final Manuscript May 26, 1992