ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

P. Bertolazzi, G. Felici

LEARNING TO CLASSIFY SPECIES WITH BARCODES

R. 665 2007

Paola Bertolazzi – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: bertola@iasi.cnr.it.

G. Felici – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: felici@iasi.cnr.it.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

According to many field experts, species identification based on morphological keys needs to be supported with automated techinques based on the analysis of DNA fragments (referred to as "barcode"). The most successful results in this area are those obtained from a particular fragment of mithocondrial DNA, the gene cytochrome c oxidase I (COI).

We present the results obtained on a dataset where a number of COI fragments composed of 690 LOCI are used to describe 1700 individuals belonging to 150 different species. The method proposed exhibits high correct recognition rate on a 20% testing set (correct recognition approx. 99% when 30 features are used). It is able to provide compact formulas on the values (A,C,G,T) at the selected loci that synthetize the characteristic of each specie, a quite precious information for many tasks of interest for taxonomists.

1. Introduction

In this paper we consider an automatic data analysis method to support the classification of species. Species identification based on morphological keys is in fact showing its limitations for several reasons, among which the main are found in the lack of a sufficient number of taxonomists - and consequently of the expertise required to recognize the morphological keys - and in the fact that the chosen observable keys could not be present in a given individual, since they are effective only for a particular life stage or gender. Identification methods based on a small DNA subsequence are first proposed for least morphologically distinguished species like bacteria, protists and viruses Nanney82, Pace97 and then extended to higher organisms Brown99,Bucklin99.

In his first paper on this topic Hebert03 Hebert proposes a new technique, *DNA barcoding*, that uses a short DNA sequence from a standardized and agreed-upon position in the genome as a molecular diagnostic for species-level identification. He identifies a small portion of the mitochondrial DNA (mt-DNA), the gene cytochrome c oxidase I (COI), to be used as a taxon "barcode", that differs by several percent, even in closely related species, and collects enough information to identify the species of an individual. This molecule, prevoiusly identified by [Saccone *et al.*, 1999] as a good target for analysis, is easy to isolate and analyze and it has been shown Min07 that resumes many properties of the entire mt-DNA sequence. Since 2003 COI has been used by Hebert to study fishes, birds, and other species Hebert04a, Hebert04b; one of the most significant results concerns the identification of criptic species among insect parasitoids Hebert06. For sake of completeness we remind that another mt-DNA subsequence (gene), Cytochrome b, was proposed as a common species-level marker, while COI is specific for animal species Hebert07.

On the basis of these results the Consortium of Barcode of Life (CBOL)[1] was established in 2003. CBOL is an international initiative devoted to developing DNA barcoding as a global standard for the identification of biological species, and has identified data analysis issue as one of the central objectives of the initiative. In particular:

- optimize sample sizes and geographic sampling schemes, as barcodes are not easy to measure, and large samples are very expensive;
- consider various statistical techniques for assigning unidentified specimens to known species, and for discovering new species;
- stating similarity among species using character-based barcodes and identify what are the character based patterns of nucleotide variation within the sequenced region;
- identify small portion of the barcode that are relevant for species classification, as sequencing long molecules is expensive (shrinking the barcode).

In this paper we deal with the last two items. We propose a method that, given a sample, finds a small (30 sites) relevant portion of the COI sequence that allows to distinguish among the species that are present in the sample, and we provide a character based pattern for each species (i.e. a *logic formula*) that allows to correctly classify all the individuals of the sample and individual whose speciesis unknown.

The method, already described and applied in some other variants in previous work (Felici02, Felici05, Felici06, Bertolazzi08) is new for this problem; all the referred works analyze COI by first creating comprehensive profiles and then using the Neighbor Joining (NJ) method Saitou87 to obtain a phylogenetic tree, so that each species is identified as represented by a distinct, non overlapping cluster of sequences in the NJ tree.

Our method is comprised of two steps. The first step is feature selection, where the problem of selecting a small number of relevant features is formulated as an integer programming problem; a similar approach for feature selection has been adopted in [Bertolazzi *et al.*, 2008], but here we adopt a more compact model that can be solved exactly rather than with heuristics there used. The second step is the identification of the logic formulas that separate each class from all the others. Such task is accomplished using the *Lsquare* system for logic mining, originally described in [Felici *et al.*, 2002].

[1] http://www.barcoding.si.edu/

The overall method is a *supervised learning* method, based on identifying the features and the set of rules on a training data set and applying the model to a test data set (in the experiments described at the end of the paper, we have adopted $90-10$ and $80-20$ training-testing splits).

The main benefit of this method with respect to other more standard data mining approaches is its capability to provide compact classification rules that have a great semantic content, since they identify, for each specie, the sites of the molecule, the alleles that are characteristic of that specie, and the propositional logic formulas that link them.

The paper is organized as follows: in Section 2.1 we introduce the main notation and definitions used in the paper; in Section 2.2 we describe the features selection model adopted; in secion 2.3 we provide the reader with a synthetic description of the logic mining method *Lsquare* (futher details are found in the related literature). Then, in section 3 we describe the data set used and the results of the experiments. Final remarks are the topic of section 4.

2. System and Methods

2.1. Main Definitions and Notation

We introduce the terminology adopted in the paper. We assume that each individual is described by its barcode, that in turn is composed of a fixed number of m sites (690, in the case of COI). Each individual belongs to one and only one specie, or *class*. The data set is composed of n individuals, belonging to two or more classes; we refer to the individuals also as *element*, and to the sites of the barcode as *features*. The $i-th$ element of the data set is represented by the vector $f_i = (f_{i1}, f_{i2}, ..., f_{im})$, where $f_{ij} \in \{A, C, G, T\}$; the data matrix is represent by the sequence of vectors $f_1, f_2, ..., f_n$. Given this matrix representation of the data set, when appropriate the elements may also be referred to as *rows*, while the features as *columns*. The classification method adopted is basically a two-class separation method, in the sense that it identifies the logic formula that separate the elements of one class in the data set from the remaining elements of the data set (such elements may belong to one or more classes). When needed, we refer to the two classes to be distinguished as *class A* and *class B*.

2.2. Shrinking the Barcode

The identification of a subset of relevant features among a large set is a typical problem in Data Analysis and Data Mining, often referred to as *feature selection*. Among the different approaches, the idea of formulating the feature selection problem as a mathematical optmization problem where the number of selected features is to be minimized under some constraints has received some attention in the literature, and proven to be effective in many situations. In [Chang *et al.*, 2006] the authors adopt such an approach for the selection of TAG SNPs; the mathematical model adopted turns out to be a linear problem with binary variables whose structure is well known in the combinatorial optimization literature as the *set covering problem*. Several similar models where also treated in [Garey and Johnson, 1979], where large set covering models where proposed (a.k.a. the *test cover* problems). The main drawback of this approach, and of the many variants that have been then proposed, lays in the fact that it uses one constraint of the integer programming problem for each pair of elements of the data set that belong to different classes. Such fact implies a rapid growth of the dimension of the problem, and thus of its intractability, that then requires the use of non optimal solution algorithms (e.g., in [Bertolazzi *et al.*, 2008] an efficient GRASP heuristic is used to solve such large formulations).

In this paper we use a different approach that results in a more compact formulation and can be solved optimally in reasonable computation time for problem of interesting size (as the ones considered in the experiments). Such formulation is based on a very simple idea.

For the time being, we assume the items to belong to only two classes, class A and class B. Given a feature f_j, we define $P_A(j,k)$ and $P_B(j,k)$ be the proportion of individuals where feature f_j has value k (for $k \in (A, C, G, T)$) in sets A and B, respectively. If $P_A(j,k) > P_B(j,k)$ (resp. $P_B(j,k) > P_A(j,k)$), then the presence of f_j with value k is likely to characterize items that belong to class A (resp. B). To better qualify the strict inequality between $P_B(j,k)$ and $P_A(j,k)$ we introduce an additional parameter $\lambda > 1$, and then define, for each feature j and for each individual i in class A the vector d_{ij} as follows.

$$d_{ij} = \begin{cases} 1, & \text{if } f_{ij} = k \text{ and } P_A(j,k) \geq \lambda P_B(j,k); \\ 0, & \text{if } f_{ij} = k \text{ and } \lambda P_A(j,k) \leq P_B(j,k); \\ 1, & \text{if } f_{ij} \neq k \text{ and } \lambda P_A(j,k) \leq P_B(j,k); \\ 0, & \text{if } f_{ij} \neq k \text{ and } P_A(j,k) >\geq \lambda P_B(j,k); \end{cases}$$

While, for individuals i in class B, the value of d_{ij} will be:

$$d_{ij} = \begin{cases} 1, & \text{if } f_{ij} = k \text{ and } \lambda P_A(j,k) \leq P_B(j,k); \\ 0, & \text{if } f_{ij} = k \text{ and } P_A(j,k) \geq \lambda P_B(j,k); \\ 1, & \text{if } f_{ij} \neq k \text{ and } P_A(j,k) \geq \lambda P_B(j,k); \\ 0, & \text{if } f_{ij} \neq k \text{ and } \lambda P_A(j,k) \leq P_B(j,k); \end{cases}$$

In the practical application the parameter λ directly influences the density of the matrix composed of d_{ij} and can be adjusted to obtain a reasonable value for the density itself (say 20%).

According to this definition, we assume that the number of ones in vector $d_{.j}$ is positively correlated with the capability of feature f_j to discriminate between classes A and B. We would then like to select a subset of the features that exhibits, as a set, a good discriminating power for all the items considered, so that we may use more features combined together to build rules that perform a complete separation between A and B.

The purpose of the feature selection model is then to select a given and small number of features that guarantee a good discriminating power for all the elements of the data sets. This can be formally stated asking to select a given number of features (say, β) that maximize the minimum of the discriminating power over all the items.

We define the binary decision variable $x_j = \{0, 1\}$ with the interpretation that $x_j = 1$ (resp. $x_j = 0$) means that feature j is selected, (resp., is not selected). The binary integer optimization problem can then be defined as follows:

$$\begin{array}{lll} \max & \alpha & \\ s.t. & \sum_{i=1}^{m} d_{ij} x_j \geq \alpha & i = 1 \ldots n \\ & \sum_{j=1}^{m} x_j \leq \beta & \\ & x_j \in \{0, 1\} & j = 1 \ldots m, \end{array} \tag{1}$$

The optimal solution of the above problem would then select the β features that guarantee the largest discriminating power over all the elements in the data[2] (we note that β is a parameter of the problem, and not a variable).

The number of variables of the problem is given by the number of features (m), and the number of rows by the number of individuals (n), keeping the size of the problem in a linear relation with the size of the data. The problem is anyway difficult to solve, and for large sizes approximate solution methods may be needed; nevertheless, for cases where the number of features and the number of elements are in the thousands, optimal solutions can be obtained with commercial softwares (for the experiments described in this paper, ILOG CPLEX 10.0 was used). Once a optimal set of β features is selected, these are used by the logic mining tool *Lsquare* to extract the separating formulas, as described in the next section.

2.3. The Extraction of Separating Logic Formulas

Lsquare is a learning method that operates on data represented by logic variables and produces rules in propositional logic that classify the items in one of two classes. The appopriateness of *Lsquare* for species classification is motivated by the fact that it uses a logic representation of the description variables, that

[2]Despite the problem has been described with straight-forward arguments, it is easy to see how it amounts to identify the feature set that maximize the additive *class entropy* of its elements.

are to all extents logic variables, and of the classification rules, that are logic formulas in Disjunctive Normal From (DNF). Such property enables to analyze the classification results also from the semantic point of view, as the classification rules determined by the method express combination of the features that can be appreciated by domain experts and may bring to light useful knowledge in an easily understandable format.

The classification rules are determined using a particular problem formulation that amounts to be a well know and hard combinatorial optimization problem, the *minimum cost satisfiabilty problem*, or MINSAT, that is solved using a solver based on decomposition and learning techniques Truemper04. The DNF formulas identified have the property of being created by conjunctive clauses that are searched for starting from those that cover the largest portions of the training set. Therefore, they usually are formed by few clauses with large coverage (the interpretation of the trends present in the data) and several clauses with smaller coverage (the interpretation of the outliers in the training set).

More formally, the problem of finding a separating DNF for A and B is solved sequentially, identifying at each iteration a conjunctive clause that holds *True* for the largest non-separated subset of A and *False* for all B. Termination of the process is guaranteed by some property of the method (see [Felici *et al.*, 2002]). Each iteration is in turn formulated as a logic optimization problem, that we briefly describe here. Basic notions of propositional logic are given for granted and can be found in Felici02.

First, we expand the selected features into 4 different logic variables, each one associated with the presence or absence of one the 4 nucleotides in the given position. For example, v_{jA} with value *True* indicates that in position j is present nucleotide A, and *False* otherwise. For simplicity, assume that all these logic variables are sequentially indexed from 1 to M, and referred to as v_j. Thus, v_j= *True* for individual i means that, for that individual, a certain position exhibits a certain nucleotide.

Second, we formulate a MINSAT problem whose solution identifies one of the CNF clauses that will form the final DNF formula. To do this, we introduce two additional types of logic variables:

- p_j and q_j, linked with the v_js as follows: v_j is chosen in the clause with value *True* if $p_j = True$ and $q_j = False$; v_j is chosen in the clause with value *False* if $p_j = False$ and $q_i = True$, and v_j is not chosen in the clause if $p_j = q_j = False$;

- e_i, associated with each element i of class A, that are forced by the constraints to assume value *False* if the clause identified by the solution holds *True* for i, and *False* otherwise.

Also, define as A_i^+ the set of indices of the features that appear in element i of class A with value *True*; and, simmetrically define as A_j^- the set of indices of the features that appear in element i of class A with value *False*. Analogously define B_i^+ and B_i^-.

Consider now the following MINSAT problem, whose solution is determined by an assignment of the logic variables p_j and q_j and e_i such that all the logic constraints are satisfied and the sum of costs of the variables that hold *True* is minimized:

$$\begin{array}{ll} \min & \sum_{i \in A} e_i \\ & (\bigvee_{j \in B_i^+} q_j) \vee (\bigvee_{j \in B_i^-} p_j) \quad \forall i \in B \\ & \neg q_j \vee \neg p_j \quad \forall j \\ & \neg q_j \vee e_i \quad \forall i \in A, \forall j \in A_i^+ \\ & \neg p_j \vee e_i \quad \forall i \in A, \forall j \in A_i^- \end{array} \tag{2}$$

It can be verified that the solution of problem (3) identifies a CNF clause on the v_j and that the set $A' = \{i \in A | e_i = False\}$ is the largest portion of A that can be separated from B by a simple CNF clause. Using this information, we can formulate a second MINSAT problem where we select, amongst all separating clause that separate A' from B, the one that uses the least number of literals (e.g., the most compact clause):

$$\begin{array}{lll} \min & \sum_j p_j \sum_j q_j & \\ & (\bigvee_{j \in B_i^+} q_j) \vee (\bigvee_{j \in B_i^-} p_j) & \forall i \in B \\ & \neg q_j \vee \neg p_j & \forall j \\ & \neg q_j. & \forall i \in A', \forall j \in A'^+_i \\ & \neg p_j. & \forall i \in A', \forall j \in A'^-_i \end{array} \tag{3}$$

A more detailed description of the system and of its other components can be found in the related papers ([Felici *et al.*, 2002], [Felici *et al.*, 2005], [Felici *et al.*, 2006]); an efficient implementation of the algorithm can be downloaded at *www.leibnizsystems.com*.

3. Implementation and Discussion

The method has been tested on a data set provided by the Consortium of the Barcode of Life in the 2006 Conference[3]; it is composed of 1700 barcode fragments coming from individuals belonging to 150 different species; each fragment contains 690 sites (or nucleotides). The experiments have been conducted according to the following scheme.

- The data is split into training and testing data, adopting a proportion of 80% and 90% as training (the remaining being used for testing);
- For each species$s$:
  - a 2-class classification problem is defined, where class A contains the individuals of species$s$, and class B the individuals of the other 149 classes.
  - the training data is used to formulate the feature selection problem described in Section (2.2), and to identify the optimal set of features for different values of the parameter β (10, 20, and 30).
  - The *Lsquare* system is used to identify logic formulas based on the selected features, to separate the individuals in class A from those in class B.
- the formula for species$s$ is saved, and the above is iterate for all the 150 species.

At the end of this process, we have 150 logic formulas - one for each species- and apply these formulas to the individuals of the training and of the testing splits. If an individual is recognized as positive by the formula of species$s$, we declare its predicted class to be s and then verifies if such prediction is correct. When an individual is recognized as positive by more than one formula (or by none of them) we register such an event as a recognition error.

The scheme is then repeated for different random splits of the data in training and testing.

The results are summarized in Table 1, that reports a row for each of the experiments that have been conducted. The values of β (10, 20, or 30) and the corresponding value of α obtained from the optimal solution of the feature selection problem are listed in columns 1 and 2; column 3 contains the percentage of data used for testing (10% or 20%). In the last 2 columns are reported the percentage of error obtained on the training and on the testing data, respectively.

The overall error date decreases, as is to be expected, when a larger training set is used, due to the fact that the information used to extract the formulas is larger and the formulas are therefore more accurate. In the same way, we note that experiments with fewer features (where $\beta = 10$) are less precise than those with more features; to any extent, for the largest values of β used (30), the error rates are very small also when the testing data used is larger (20%). This means that the system is able to extract good formulas using only 30 of the 690 sites that are present in the barcode. Moreover, when comparing the error rates obtained on the training set with those obtained on testing we note a very little decay in the performances, thus highlighting the good generalization capabilities of the formulas and the important role of the barcode data for species discrimination.

[3] http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges2007/

Table 1: Optimal values and Recognition Rates

β	α	test%	values of error training	values of error testing
10	2	10	11.14%	11.38%
10	2	10	5.62%	8.98%
10	3	10	5.99%	7.19%
10	3	20	9.21%	10.17%
10	3	20	7.11%	9.75%
10	2	20	7.81%	8.05%
average			7.81%	9.25%
20	6	10	2.06%	2.40%
20	6	10	0.84%	1.20%
20	6	10	0.28%	0.60%
20	6	20	1.90%	2.12%
20	6	20	0.30%	1.27%
20	6	20	0.30%	1.27%
average			0.95%	1.48%
30	9	10	0.37%	0.60%
30	9	10	0.28%	0.60%
30	9	10	0.37%	0.60%
30	9	20	0.30%	1.69%
30	9	20	0.30%	0.85%
average			0.33%	0.87%

Table 2: Logic Formulas for Species 1 to 5

SPECIE	DIM	CLAUSE(S)
1	1	19 T 172 T
2	1	340 G 343 A 445 C
3	1	445 T 499 T 580 G
4	1	172 C 445 T 493 G 499 C 580 A
5	1	58 G 430 T

It is of interest to check the frequency by which the different features (barcode sites) appear in all the formulas that have been identified for the different random splits. We identify a group of sites that appear with particularly high frequency (i.e., are present in many of the formulas obtained by the method) that are likely to be those whose combination best expresses the difference among the 150 species considered: such sites are in position 580, 490, 346, 469, 544, 637, 331 of the barcode.

The logic formulas are indeed very compact, and very few of them are composed of more than one CNF clause; such clauses are composed of few (usually 3, but never more that 5) literals (i.e., combination of a feature and its value). In Table 2 we report a list of the separating formulas for the first 5 species of the 150 available for one of the experiments:

An example of the interpretation of the formulas in Table 2 is here given:

- First line of Table 2: *if nucleotide in position 19 of the barcode has value T and nucleotide in*

position 172 has value T, then the speciesis 1.

- Second line of Table 2: *if nucleotide in position 340 of the barcode has value G, nucleotide in position 343 has value A, and nucleotide in position 445 has value C, then the species is 2.*

4. Conclusions

In this work we have discussed the application of Data Mining methods for species classification. We consider the problem of the analysis of barcode - a particular fragment of mithocondrial DNA that has recently been identified as a potential collector of genetic information that is useful to discriminate among species. The method adopted is comprised of two main steps; the first is based on the compression of the barcode into a reduced set of very informative features using a particular integer programming formulation; the second consists in the application of a logic mining method - the *Lsquare* system - to identify separating formulas on the compressed data. The method has been proved to be practical, sufficienty fast and precise, exhibiting negligible error rates on training data and producing extremely compact separating formulas. Such latter feature plays a very important role in this type of applications as it results in consistent semantic value that can be used by field experts to enhance and complete their knowledge of the studied phenomenon - in this case, the relation between species taxonomies and the COI mithocondrial DNA.

Acknowledgements

We wish to thank Prof. David Schindel - Executive Secretary of the Consortium ofr the Barcode of Life - and Prof. Xavier Cabrera for their interest and their support in our research. To Livia Di Trani we are grateful for having introduced us this very interesting problem. We finally thank Paul Hebert, Cecilia Saccone, and Giuseppe Martini for their precious advice and feedback on this work.

References

[Bertolazzi *et al.*, 2008] Bertolazzi P., Felici G., Festa P., Lancia G.,(2006) Logic Classification and Feature Selection for Biomedical Data, *Computers & Mathematics with Applications*, 55-5, Pages 889-899.

[Brown *et al.*, 1999] Brown, B., Emberson, R.M., Paterson, A.M. (1999) Mitochondrial COI and II provide useful markers for Weiseana (Lepidoptera, Hepialidae) species identification, *Bull. Entomol.*, **89**, 287294.

[Bucklin *et al.*, 1999] Bucklin, A., Guarnieri, M., Hill, R. S., Bentley, A. M., Kaartvedt, S. (1999) Taxonomic and systematic assessment of planktonic copepods using mitochondrial COI sequence variation and competitive, species-specific PCR, *Hydrobiology*, **401**, 239254.

[Chang *et al.*, 2006] Chang C-J., Huang Y-T, Chao K-M.(2006) A greedier approach for finding tag SNPs, *Bioinformatics*, **22-6**, pages 685691.

[Felici *et al.*, 2002] Felici G., Truemper K (2002) A Minsat Approach for Learning in Logic Domains, *INFORMS JOC*, **1**.

[Felici *et al.*, 2005] Felici G., Truemper K.(2005) The Lsquare System for Mining Logic Data, *Encyclopedia of Data Warehousing and Mining*, **1**, .

[Felici *et al.*, 2006] Felici G., Sun F-S., Truemper K., Learning Logic Formulas and Related Error Distributions, in *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques*, G. Felici and E. Trintaphyllou eds., Springer 2006

[Felici *et al.*, 2006b] Felici G., de Angelis V., Mancinelli G., Feature Selection for Data Mining, in it Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques, G. Felici and E. Triantaphyllou eds., Springer 2006.

10.

[Garey and Johnson, 1979] ,Garey M.R.,Johnson, D.S., Computer and Intractability: a guide to the theory of NP- completeness, *Freeman, San Francisco.*

[Hajibabaei *et al.*, 2007] Hajibabaei M., Singer G. A. C. , Clare E. L. and Paul D. N. Hebert P. D. N. (2007) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring, *BMC Biology* , **5;24**.

[Hebert *et al.*, 2003] Hebert P. D. N. , Cywinska A, Ball S.L., deWaard J.R. (2003) Biological identifications through DNA barcodes, *Proc. R. Soc. Lond. B (2003)*, **270**, 313321.

[Hebert *et al.*, 2004a] Hebert P.D.N, Penton E.H, Burns J.M, Janzen D.H, Hallwachs W.(2004a) Ten species in one: DNA barcoding reveals cryptic species in the Neotropical skipper butterfly Astraptes fulgerator, *Proc. Natl Acad. Sci. USA.*, **101**, 1481214817.

[Hebert *et al.*, 2004b] Hebert P.D.N, Stoeckle M.Y, Zemlak T.S, Francis C.M. (2004b) Identification of birds through COI DNA barcodes, *PLOS Biol.* , **2**, 17.

[Min *et al.*, 1999] Min, X. J., Hickey, D. A.(2007) DNA barcodes provide a quick preview of mitochondrial genome composition, *PLoS ONE* , **2(3)**, e325.

[Nanney, 1982] Nanney, D. L. (1982) Genes and phenes in Tetrahymena *Bioscience* , **32**, 783740.

[Pace, 1997] Pace, N. R. (1997) A molecular view of microbial diversity and the biosphere *Science* , **276**, 734740.

[Saccone *et al.*, 1999] Saccone, C., DeCarla, G., Gissi, C., Pesole, G., Reynes, A. (1999) Evolutionary genomics in the Metazoa: the mitochondrial DNA as a model system, *Gene* , **238**, 195210.

[Saitou *et al.*, 1987] Saitou N., Nei M. (1987) The Neighbour-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4(4)**, 406-425.

[Smith *et al.*, 2006] Smith M. A., Woodley N. E., Janzen D. H., Hallwachs W., and Paul D. N. Hebert P. D. N. (2006) DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae), *PNAS* , **103**, 36573662.

[Truemper *et al.*, 2004] Truemper K.(2004) Design of Logic-Based Intelligent Systems, *Wiley-Interscience.*