CHARACTERISTIC RESTRICTION ENDONUCLEASE CUT ORDER FOR CLASSIFICATION AND IDENTIFICATION OF FUNGAL SEQUENCES

Rajib Sengupta, Dhundy Bastola and Hesham Ali*

College of Information Science and Technology, University of Nebraska, Omaha, NE 68182, USA
**Email: hali@unomaha.edu*

Restriction fragment length polymorphism (RFLP) of chromosomal DNA is one of the powerful molecular tools used in the fingerprinting of microorganism and epidemiological studies. In a wet-lab setting, pattern-based classification of organism using RFLP begins with the digestion of DNA with one to two restriction enzymes, which is followed by gel electrophoresis. This wet-lab approach may not be practical when the experimental data set includes a large number of genetic sequences and a wide pool of restriction enzymes to choose from. Alternatively, the RFLP process must be simulated in-silico using computation intensive sequence alignment methods. In this study we introduce a novel concept of Enzyme Cut Order- a biological property- based characteristic of DNA sequences which can be defined and analyzed computationally without any alignment algorithm. In this alignment-free approach a similarity matrix is developed based on the pairwise Longest Common Subsequences (LCS) of the Enzyme Cut Orders. The choice of an ideal set of restriction enzymes used for analysis is augmented by using genetic algorithms, and the target sequences include the internal transcribed spacer regions of rDNA from fungi. The results obtained from this approach show that the organisms that are related phylogenetically form a single cluster and successful grouping of phylogenetically close or distant organisms is dependent on the choice of restriction enzyme used in the analysis. This novel alignment-free method, which utilizes the Enzyme Cut Order and restriction enzyme profile, is a reliable alternative to local or global alignment-based classification and identification of organisms

1. INTRODUCTION

Construction of phylogenies is one of the central activities of biologists for the reconstruction of the history of life and to understand biology in light of evolution [1]. Therefore, it is not surprising to see classification of organism as one of the major biological activities. The molecular approach to classification and identification of organisms requires comparison of genetic sequences. The detection of similarities between two or more sequences is often the first step in identification of relevant features in the DNA sequences or their translated amino acid sequences. Existing computational approaches involved in the identification of organisms based on this feature primarily include pairwise local or multiple sequence alignment. These approaches can be broadly categorized into two groups: the similarity-based approach where pairwise similarities are used in clustering the sequences, and the, pattern-based approach, where regional similarities are translated into statistically significant character patterns. Generally, the former method is commonly applied to DNA sequences while the later is used in the classification of proteins. However, in the biology laboratory, RFLP and Southern Blotting are still used and widely accepted methods in molecular identification and phylogenetic studies. This approach requires the sequences to be cut into several fragments with the help of restriction endonucleases, which are proteins that recognize particular sequences of nucleotide (called the restriction site and generally 4 to 8 bases long) and cut the double stranded DNA molecule at restriction site. Variations in the position of these sites along the DNA, among the sequences being analyzed, will lead to digested products of varying lengths. Following a high-resolution gel electrophoresis of the digested product, the fragment-patterns are visually compared to determine the similarity among the sequences. The inherent biological property of the restriction enzyme to selectively recognize a 4 to 8 base-long nucleotide sequence has been used in in-silico RFLP and the fragment data have been computationally analyzed [2]. While such an approach is useful and has been

* Corresponding author.

implemented to identify a set of restriction enzymes appropriate for high resolution analysis of complex microbial communities, this inherent property of the DNA has not been used in comparing genetic sequences at a higher level (coarse granularity). For this purpose, a pairwise alignment of the individual nucleotide base (fine-granularity) is typically performed.

Clustering sequences with their pairwise similarities is useful when the sequences are (a) closely related, (b) identical in size and (c) can be aligned over their entire length without the introduction of gaps. Such multiple sequence alignments are constructed by the method known as progressive sequence alignment [3-5] which is a computationally intensive process, particularly with large data sets. Additionally, the rule of progressive sequence alignment only allows gaps to be added or enlarged and disallows its movement or removal. Since gaps are interpreted as evolutionary events in molecular phylogeny, misaligned sequences have no useful biological information. Therefore, use of biological features derived from DNA sequences that can be modeled computationally is a desirable alternative to multiple alignment based analysis. Such an approach will allow us to utilize 'coarse-grain-features' as opposed to the 'fine-grain-features' represented by individual bases within the nucleotide sequences.

In this paper we have introduced a novel concept of Enzyme Cut Order- a restriction enzyme-based characteristic of DNA sequences. This method maps the restriction sites for a set of restriction enzymes on the sequences being analyzed and determines the longest common subsequences (LCS) among each pair of enzyme cut order as the similarity score of the corresponding sequences. The similarity matrix obtained from the pairwise LCS is then used in the clustering of these sequences. This new approach to utilize biological features derived from DNA sequences directly in the analysis of a large set of sequence data is a valuable contribution in the classification of organisms based on genetic sequences.

1.1. Previous-Work

Analysis of restriction endonuclease-derived fragmentation patterns obtained from RFLP [2, 6, 7] has been shown to be particularly useful for high resolution analysis of highly complex microbial communities. Since the late 1990s, the RFLP technique has become popular in ectomycorrhizal molecular ecology studies, which was done on specific regions (16s rRNA for bacterial community [8] and ITS nrDNA for fungal community [9]). For the fungi community the TRFLP technique that was initially developed by [10] is presently getting employed in several fungal ecology studies [9, 11, 12] for molecular fingerprinting. This is a derivative of RFLP in which species-specific DNA regions are selectively amplified by PCR with fluorescently labeled primers and subsequently digested with restriction enzymes. The terminal restriction fragments (TRFs) are thus tagged and their sizes are measured with extreme precision by using capillary electrophoresis DNA analyzers [13].

As the laboratory methods of RFLP and TRFLP gained popularity, simulation of lab methods in-silico has been attempted. TRFLP program (TAP) [8] shows that RFLP can be successfully used in analyzing microbial-community with the 16s rRNA sequence. Computer-simulated restriction analysis has also been carried out to analyze ctomycorrhizal Fungi [9]. However, these RFLP in-silico analyses use restriction fragment size for initial similarity, followed by pairwise alignment (DNAMAN program by Lynnon Biosoft). Additionally, these analyses make use of specific enzymes (HhaI, MspI, RsaI for Bacteria, TaqI, HaeIII, HinfI, AluI, RsaI, MspI for Fungi) . Thus, it requires manual intervention from biologists.

In the RFLP/TRFLP in-silico methods that have been proposed [8, 9, 14] the fragment length is measured from a particular site (e.g.: in [9] all TRF lengths are measured downstream of the 18S start codon (TCATTA)) and the ordering of the enzyme cuts are not considered at all - as such the effect of multiple enzymes in conjunction to each other and its effect on the RFLP pattern is ignored. Therefore, the RFLP data obtained by using these methods are not useful for sequence classification. As mentioned by [15] the connection of T-RFLP data to phylogenetic sequence information is difficult. To overcome this limitation, all these methods include a second step where sequence alignment is carried out for further analysis.

2. MATERIALS AND METHODS

In this section we will first discuss the concepts and definitions we are proposing in this paper. This discussion is then followed by overview of methods, data collection and discussion on the computational methods and algorithms including the application of this method in identification/classification of sequences.

2.1. Concepts and Definitions

Sub-headings should be typeset in boldface and capitalize the first letter of the first word only. Section number to be in boldface roman.

2.1.1. *Enzyme Cut Order*

The Enzyme Cut Order (ECO) for a DNA sequence (S) for a particular set of restriction enzymes [Ez] is defined as a string (array) of enzyme names (represented as numeric id) in the order each enzyme (ez € Ez) cuts the sequence. In the following example the restriction enzyme cut sites for HaeIII and AccII are shown by ↓ and the enzyme cut order is = [2, 1, 2..]:

TTTTACGC↓GCCCTCGAGG↓CCACCCTGG↓CCA……GAG

ENZ ID	ENZ NAME	CUT SITE	CUT POS
1	HaeIII	GGCC	2 (GG | CC)
2	AccII	CGCG	3 (CGC|G)

2.1.2. *Enzyme Cut Order Similarity*

We analyzed the enzyme cut orders on several test sequences. The restriction enzymes that are commonly used in laboratory for fungi (TaqI -209, HaeIII -108, HinfI-165, AluI -33, RsaI -22, MspI -109) are used to derive the Enzyme Cut Order on different groups of sequences. Table 1 below shows that similar organisms have a similar Enzyme cut order.

Table 1. Showing Similar Enzyme Cut Order among similar organisms.

Acc. Id	Organism	Enzyme Cut Order
OPL416069	*Oligoporus placentas*	33,108,108,209,165,209, 165,165,33,33
OPL249267	*Oligoporus placentas*	33,108,108,209,165,209, 165,165,33
AY310442	*Nectria haematococca*	108,108,209,165,209,165,209, 108,109,108,33,33,108,209,165
AY188918	*Nectria haematococca*	108,108,209,165,209,165,209, 108,109,108,33,33,108,209,165
AY138847	*Nectria mauritiicola*	165,109,108,109,109,209,165,20 9, 165,209,108,109,108,109,108, 209,109,108,209,165

2.1.3. *Enzyme Cut Order Similarity Score*

The similarity score between two Enzyme Cut Orders includes two components: a) how many enzymes are similar and b) the order in which these similar enzymes occur. The similarity score will be higher if we find many similar enzymes appearing in the same order among two Enzyme Cut Orders. Mathematically the similarity score can be obtained from the Longest Common Subsequence (LCS) among two strings, where the strings are the Enzyme Cut Orders in question. Therefore the length of Longest Common Subsequence (LCS) between two Enzyme Cut Orders (E1 and E2) of two corresponding sequences (S1 and S2) is considered as the Enzyme Cut Order Similarity Score between E1 and E2. As the Enzyme Cut Order is mapped with only one sequence and vice-versa (one sequence will have only one enzyme cut order) for a given enzyme set, the Enzyme Cut Order Similarity Score between E1 and E2 is considered to be the similarity score between S1 and S2.

2.2. Overview of the Method

The flow-chart in Fig 1 displays the overall method. First all necessary data, sequence, taxonomy information and restriction enzyme information are collected, curated and loaded in the local database. Next the Enzyme Cut order for each sequence is constructed and the similarity score algorithm, which is a dynamic programming algorithm for calculating the Longest Common Subsequence among every pair of Enzyme Cut Orders, is executed. The similarity score is used to construct the similarity matrix, which is clustered, and the cluster is analyzed for its phylogenetic accuracy. Finally at the end of this process we introduce the genetic algorithm to find the optimal enzyme set for a particular dataset. The optimal enzyme set is defined as the minimal size enzyme set , which shows increased phylogentic resolution on the Similarity matrix obtained from the enzyme cut order for a given dataset.

2.2.1. *Data Collection*

Internal Transcribe Spacer sequences for all the members of fungi were collected from the Genbank sequences downloaded via ftp. In the Genbank, sequences for fungi are categorized within the plant group. Therefore, a parser was written in Perl (that also utilize bio-Perl package) to read through the gbpln-files and selectively identify the records containing fungi sequences of our interest. In-silico PCR was performed on these sequences, where ITS specific primers were used in the regular expression to collect all the ITS entries in the genbank repository. Using these approaches, a total of 3005 ITS sequences for fungi were collected and entered into a local Postgres database. In addition to the ITS target sequences, every record consisted of accession and GI numbers and taxonomic information.

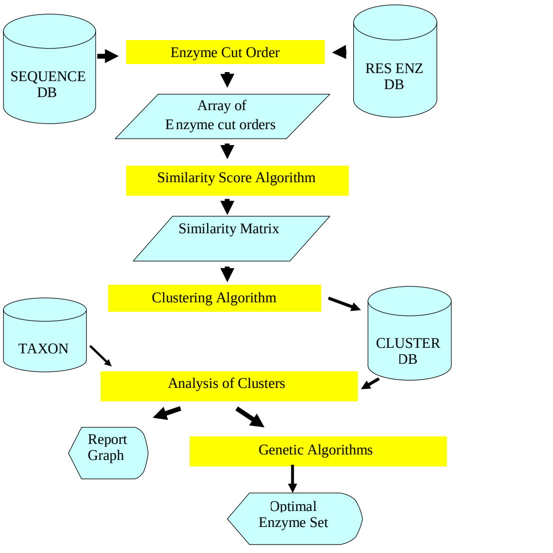

Fig. 1. Flow-Chart representing the high-level step by step method for proving that Enzyme Cut Order can be an alternative characteristic of Fungal DNA Sequence classification and identification

The latter was parsed from the “Organism description” of the Genbank entries (or OrgName_Lineage in XML format). Upon closer examination we discovered that the available taxonomy data were highly variable in terms of various classification categories. Therefore, we developed rules to computationally curate this taxonomic information before entering it into our sequence database. According to this rule, classification categories included the Kingdom, Division, Class, Order, Family, Genus, Species and when available, the strain names. As shown in Table 2, we employed simple suffix rule and the position of the taxonomical description to decide the first four taxonomical categories namely the division, class, order and family.

Table 2. Taxonomic classification categories and the decision rules that were implemented in the curate-algorithm

Suffix	Position	Tax Category
-cota	2nd and After Fungi	Division
-etes	Between Division and Order	Class
-ales	Between Class and Family	Order
-ceae	Before the known Genus	Familv

For Genus and Species, the <Org-ref_taxname> tag was used and verified with <BinomialOrgName> tag and end of the <OrgName_lineage> tag. As a result, we had two sets of data where the first (DB1) included all records with complete taxonomic information and the second (DB2) with partial taxonomic information. For the purpose of this current study only the records in DB1 were used. A list of restriction endonucleases was obtained from REBASE [16]. The Type II Restriction Enzymes that are commercially available were downloaded in bionetc format. These records were parsed and also maintained in a relational database. Each restriction enzyme entry consisted of unique identifier called Enzyme ID, enzyme name and the recognition sequence. Whenever appropriate, recognition sequences containing bases other than A, T, G and C were interpreted as per IUB ambiguity code [17]. Additional fields in the database included special features of the restriction enzymes including cut position from the start of the recognition sequence, isochizmers, and prototypes that could be used as needed. Table 3 shows some of the records for restriction enzymes

Table 3. Restriction Enzyme Information

Id	Enzyme Name	Recognition Sequence	Recognition Sequence with IUB conversion	Cut Position (From start)
3	AatI	AGG^CCT	AGGCCT	4
5	AccI,FbII, XmiI	GTMKAC	GT[AC][GT] AC	6
6	AcdI,MvnI, BstFNI,Bst UI	CGCG^	CGCG	4

2.2.2. *Data set construction*

Three sets of sequence data, namely AspCan, All9Genus, and AllFungi, were constructed. The

AspCan data set consisted of sequences from the genus *Aspergillus* and *Candida*, which are two of the most common medically important fungi. The All9Genus consisted of sequences from nine different genera that were randomly chosen and the AllFungi data set included all sequences from our local database (DB1), which had complete classification categories. Additionally, to evaluate the effect of the size and type of restriction endonucleases, different sets of restriction enzymes (Ez) were chosen with the following properties: (1) Enzymes that cut at least one of the sequences from the given sequence data. (2) Enzymes that cut 50% of the sequences of the given sequence data. (3) Enzymes that cut all the sequences at least once. (4) Commonly used restriction enzymes in a biology laboratory working with the RFLP of fungi and (5) Random enzyme sets (consisting a mixture from the sets listed previously).

2.2.3. *Similarity Matrix*

A similarity matrix or a complete weighted graph G_{Ez} was created for each enzyme set [Ez], such that each node represented a sequence and the weight between two nodes was the enzyme cut order similarity score (SS) of the corresponding enzyme cut orders. So, G_{Ez} = = (V,E) where v Є V is the sequence and e Є E is the edge between two sequences v1 and v2. The weight of e = $|e_{v1,v2}|$ = Enzyme Cut Order Similarity Score of the corresponding enzyme cut orders of v1 and v2 = The length of Longest Common Subsequence (LCS) between two Enzyme Cut Order of the corresponding sequences.

2.2.4. *Clustering and Data analysis*

Different clustering algorithms were employed including the Hierarchical and Maximum gap-based exclusive clustering. Additionally, the similarity-based hierarchical clustering, a new clustering algorithm more suited for phylogenetic problems, was used. The results obtained from the clustering were evaluated using the taxonomic information extracted from the Genbank records. The sensitivity and the positive predictive value were two important evaluation parameters and are defined as follows for a particular taxon in a group X:

Sensitivity (S) = TP / (TP + FN), and
Positive Predictive (PP) = TP / (TP + FP), where
True Positive (TP) = Count of the taxons in X
False Negative (FN) = Count of the taxons in DB1, excluding in X
False Positive (FP) = Count of other taxons which are not in X
TP+FN = Total count of the taxon in the entire DB1
TP+FP = Total counts of sequences in the group X

2.2.5. *Optimal Enzyme set using the Genetic Algorithm*

Initially the Enzyme sets were chosen with various properties as described in 2.2.2. Finally, the genetic algorithm is used to determine an unbiased way to deduce the optimal enzyme set for a given set of sequences. The optimal enzyme set is defined as the minimum size enzyme set which produces perfect clustering. This algorithm was implemented with different parameters (crossover rate, mutation rate and population size) and different genetic algorithm methods including roulette wheel, tournament, and random selection with uniform, one-point and two-point crossover. Each execution of the genetic algorithm returned the least size enzyme set(s) that generated the best score from fitness function. The Fitness Function is based on the expected and actual count of an organism in the cluster. The score was quantitatively determined in terms of Sensitivity and Positive Predictive Value and based on the taxonomic information obtained from Genbank. In other words, if for an enzyme set E1 we have x numbers of perfect groups, and for enzyme set E2 we have y numbers of perfect groups, and if x > y then E1 is more genetically suitable then E2. This enzyme set was used as the seed for the next run of genetic algorithm until the result converged or the fitness function score got too low

2.3. Sequence Identificaiton

The Restriction Enzyme-based clustering and the Genetic Algorithm to determine the best enzyme set for a given set of sequences can be used conjunctively to identify DNA sequences.

2.3.1. *Method*

The input(s) to the process are a) a master database of known, curated sequences whose taxonomical information is known and fully classified and b) a database of restriction enzymes and their recognition sequence. Based on the hypothesis of this paper, the organism of the unknown sequence will be the closest organism in the phylogenetic tree with the organism of the sequence from the master database whose enzyme cut order is most similar (Enzyme cut order similarity score) to the enzyme cut order of the unknown sequence. As for each enzyme set the enzyme cut order will be different and thus the similar sequence(s) will be different. One approach is to use the brute force method by trying each and every enzyme set and finding out the highest Enzyme cut order similarity score. If there are N number of enzymes then there is a possibility of 2N Enzyme sets and thus this problem will become a non-polynomial problem. So we reduce the computation time and complexity by a) creating a smaller set of known similar sequences with higher similarity score and b) deducing the best set of restriction enzymes to be used.

The algorithm is executed as follows:

- *Step1.* The unknown DNA sequence is input using a web browser/text file, say S. The user also can choose a subset of the sequence database (DB) to reduce the search space. The subset is defined by the taxonomical hierarchy.
- *Step2.* From the restriction enzyme database all the restriction enzymes that cut this sequence are identified ([E]). Then [E] is applied to the input sequence (S) to create the Enzyme Cut Order of sequence S (ECOs).
- *Step3.* Apply [E] on all the sequences of the database DB to create a set of Enzyme Cut Orders ([ECODB]).
- *Step4.* Obtain all the Enzyme Cut Order Similarity Score among ECOs and [ECODB] and choose the sequences and the corresponding Enzyme Cut Orders [ECODB-SEL] which have the topmost percentile (user entered percentile) similarity scores
- *Step5.* Execute Genetic Algorithm iteratively on [ECODB-SEL] using [E] to get the best enzyme set [E SEL], which is a subset of [E]
- *Step6.* Using [ESEL] create an Enzyme Cut Order similarity matrix for the unknown sequence with all the sequences from [ECODB-SEL]. Apply hierarchical clustering to create a dendogram and the unknown sequence will be grouped with the most similar sequence.

Output: The most similar sequence's organism will have the highest possibility to be the organism of the unknown sequence

3. RESULTS AND DISCUSSION

Unlike the alignment based approach, our goal in the present study was to abstract the problem at a higher level by considering enzyme recognition sequence (4 to 8 bp long patterns) instead of individual characters in the sequence while also trying to consider the effect of multiple enzymes. The thought behind the approach is to tag or fingerprint a long DNA sequence (>600bp) with several small, distinct words (enzyme recognition sequence) created by multiple restriction enzyme digestion and then finding out the similar tagging patterns among the DNA sequences. It was our hypothesis that phylogenetically related organism has similar Enzyme Cut Order. The more similar the tagging patterns, the closer are the corresponding sequences and organisms in the Phylogenetic tree.

Table 4. Summary of records in the sequence database (DB1)

Division	Class	Order	Family	Genus	Species
Basidiomycota	3	22	47	142	551
Ascomycota	11	20	49	137	443
Glomeromycota	1	2	2	2	10
Zygomycota	1	2	3	4	21
Chytridiomycota	1	1	1	1	1

The genBank parser was able to identify 3005 sequences based on the presence of given ITS-specific forward and reverse primers in the in-silico PCR step. However, upon close examination of these records, 701 sequences were not suitable for our work because they did not have all the necessary classification categories. A summary of the records grouped by division is presented in Table 4.

We initially constructed several Similarity Matrices with few test sequences (5 to 20 test sequences) to evaluate the effect of the number and choice of restriction enzymes for clustering. An example of the similarity matrix for enzyme set with 6 enzymes (TaqI -209, HaeIII -108, HinfI-165, AluI -33, RsaI -22, MspI -109) for 7 test sequences is shown in Table 5. The highest value (largest LCS) for each sequence is bold and italized and it is evident that the sequences from similar organisms have the largest similarity score (LCS). But it is also noticed that the difference in the score is not very high. For example, the maximum score for *Oligoporus* is 10 among similar sequences while the next one is 8 with *Nectria* and 7 with *Lirula*, so the gap in the score with the other organisms is 2 or 3. We created other similarity matrices with different sets of enzymes for the same 7 test sequences.

Table 5: Similarity Matrix created using only 6 enzymes (TaqI, HaeIII, HinfI, AluI, RsaI, MspI) on the following 7 sequences.

SqI	Organism	1	2	3	4	5	6	7
1	*Nectria mauritiicola*	0	12	12	11	11	6	6
2	*Nectria haematococca*	12	0	15	9	9	8	8
3	*Nectria haematococca*	12	15	0	9	9	8	8
4	*Lirula macrospora*	11	9	9	0	18	7	7
5	*Lirula macrospora*	11	9	9	18	0	7	7
6	*Oligoporus placentas*	6	8	8	7	7	0	10
7	*Oligoporus placentas*	6	8	8	7	7	10	0

Table 6: Similarity Matrix created using 57 enzymes on the same 7 sequences.

SeqId	Organism	1	2	3	4	5	6	7
1	*N. mauritiicola*	0	103	104	75	73	57	57
2	*N. haematococca*	103	0	128	72	70	62	62
3	*N. haematococca*	104	128	0	72	70	60	60
4	*L. macrospora*	75	72	72	0	123	64	64
5	*L.macrospora*	73	70	70	123	0	64	64
6	*O. placentas*	57	62	60	64	64	0	118
7	*O. placentas*	57	62	60	64	64	118	0

The choice of the enzyme as well as the number of enzymes used shows significant variation in the similarity matrix. We obtained the matrix table 6 for the 7 test sequences using 57 enzymes. In this matrix we have a higher gap between the highest score and the next score. Example: For *Oligoporus* the similarity score is 118 among similar sequences, while with non-similar sequence the next highest similarity score is 64 with *Lirula* macrospora. Thus it shows that by using a different set of enzyme the similarity score is clearly distinctive and high among similar organisms. We have three important observations from the above test matrices

- Enzyme Cut Order can be used for sequence classification.
- The Length of Common Subsequence of Enzyme Cut order can be a good enough measure for Sequence Classification
- For different enzyme sets the distance between LCS length differs and there is an optimal enzyme set for each set of sequences.

In the next step we constructed data sets consisting of sequences from the genera, *Aspergillus* and *Candida* (AspCan DB). For the medium data set, we included sequences from nine different genera (All9Genus DB). All the sequences consisting of complete taxonomic information were included in the large data set (AllFungi DB). These sequences were digested with six sets of enzymes (E1 – E6).

The results obtained from the analysis (Table 7) on the nine genus database (All9Genus) show that the perfect cluster, which is defined as a cluster where the observed size of the cluster is equal to the expected size, is a function of both size and type of enzyme sets used in the analysis. With enzyme set E4 (65 enzymes) we obtained the best result, i.e., 21 of 26 species were perfectly clustered, meaning that the observed size of the cluster is equal to its expected size (no outliers or noises) and Specificity = 1 (100%) and Positive Predictive Value = 1 (100%). The record is highlighted in Table 7.

We also use NJPlot to plot the clustering results as a dendogram and compared the results with the tree obtained from the traditional alignment-based approach followed by tree drawing. The phylogenetic resolution of the trees is almost similar except one particular sequence. Refer to the tree obtained from All9Genus database in Fig. 3 (full page diagram at end of the paper), in which we have sequences from two divisions, Ascomycota and Basidiomycota. The target sequences are from Ascomycota along with two sequences of Basidiomycota working as an outgroup. The sequences are perfectly clustered with the two Basidiomycota sequences grouped together in respect to the Ascomycota group. In the next taxonomy levels all the classes (Dothideomycetes, Chaetothynomycetes, Sordariomycetes), orders (Dothideales, Chaetothyriales, Phyllachorales) and families (Herpotrichiellaceae, Phyllochoraceae) of Ascomycota are grouped perfectly (100% or Perfect Clustering) except for one sequence (Accession id: AF451907, Colletotrichum Truncatum). In the tree obtained by using our approach this sequence was not grouped with any other sequence, thus identified as an outlier and a potential problem. In the tree obtained from Neighbor joining, this sequence is clustered but has been clustered with totally different organisms (taxonomic ranks do not match). This shows that this sequence is wrongly identified in Genbank and our method is pointing out the same.

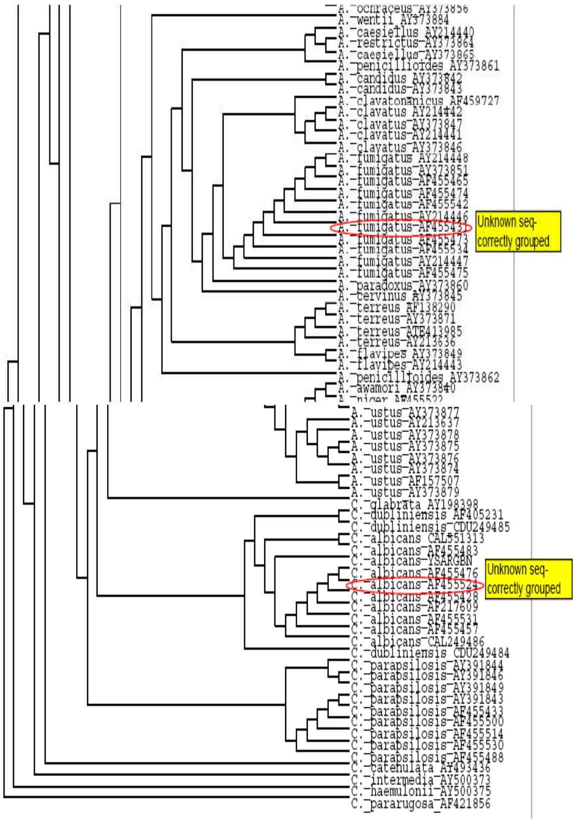

Fig 2: Identification of Sequence. We ran thru our sequence identification process for two sequences, AF455524 and AF455431. The first sequence is identified with *Candida albicans* and the second sequence with *Aspergillus fumigatus*. As per Genbank both of the sequences are identified correctly

As the last step we executed GA on the AspCan dataset and All9Genus dataset. The result is shown in Table 8. This result is not confirmed from laboratory tests, but these enzymes are known enzymes that are used for Fungi RFLP analysis in the laboratory.

To prove that the above approach of sequence identification is valid, we have chosen all the fungal sequences (Table 4) from Genbank which have full, unambiguous taxonomy information as the known sequence database (2304 sequences out of 3005 sequences). From these 2304 sequences we considered 10 sequences as unknown sequences. We carried out our test on these 10 se- quences and found the clustering of these sequences. All 10 have been clustered correctly as per their taxonomy. In Fig. 2 two of those sequences are shown as clustered and identified correctly

Table 7: Effect of enzyme property on phylogenetic resolution at the species level of All9Genus data set containing 97 sequences.

DB	Enzyme Set	Number of enzymes	Species (26)
All9Genus	E1	217	18
All9Genus	E2	57	18
All9Genus	E3	4	15
All9Genus	**E4**	**65**	**21**
All9Genus	E5	86	15
All9Genus	E6	33	15

In conclusion, this study showed that Restriction Enzymes data can be modeled and used computationally for sequence analysis and classification. In this effort we have introduced a new property of DNA sequences based on order of the cut by multiple restriction enzymes on the sequences, namely Enzyme Cut Order. We have also shown that this property can be translated to a similarity score as the length of the Longest Common Subsequence between two enzyme cut orders. The resulting similarity matrix shows high phylogenetic resolution while clustered. Thus this approach can be considered as an alternative method compared to computational intensive alignment methods or RFLP in-silico methods, followed by alignment. From a broad perspective our approach is different in the following areas:

- Instead of using the Restriction Fragment Length and pairwise alignment as used in earlier studies, we only use the Enzymes Cut Orders on the sequences.

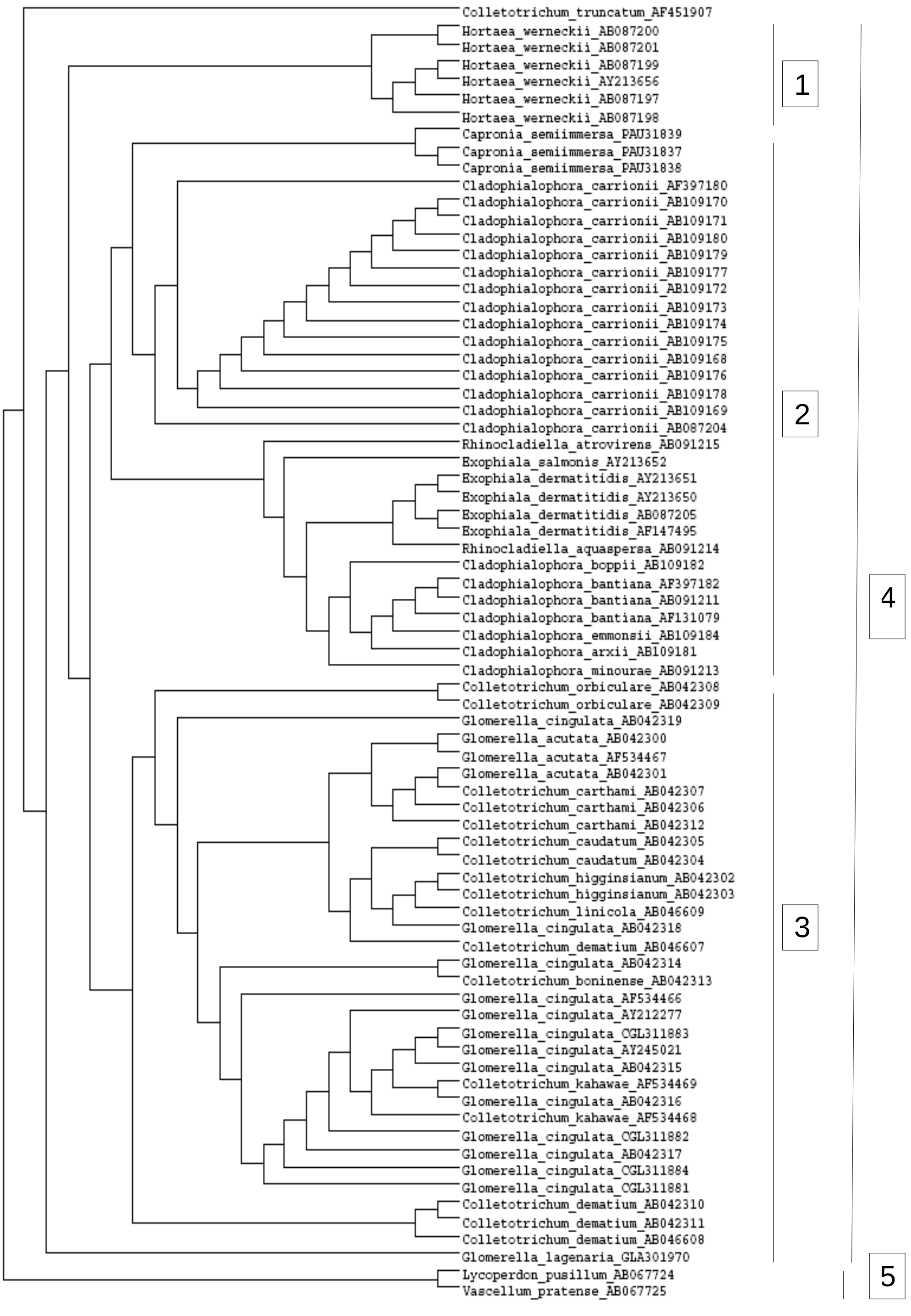
Colletotrichum_truncatum_AF451907
Hortaea_werneckii_AB087200
Hortaea_werneckii_AB087201
Hortaea_werneckii_AB087199
Hortaea_werneckii_AY213656
Hortaea_werneckii_AB087197
Hortaea_werneckii_AB087198
Capronia_semiimmersa_PAU31839
Capronia_semiimmersa_PAU31837
Capronia_semiimmersa_PAU31838
Cladophialophora_carrionii_AF397180
Cladophialophora_carrionii_AB109170
Cladophialophora_carrionii_AB109171
Cladophialophora_carrionii_AB109180
Cladophialophora_carrionii_AB109179
Cladophialophora_carrionii_AB109177
Cladophialophora_carrionii_AB109172
Cladophialophora_carrionii_AB109173
Cladophialophora_carrionii_AB109174
Cladophialophora_carrionii_AB109175
Cladophialophora_carrionii_AB109168
Cladophialophora_carrionii_AB109176
Cladophialophora_carrionii_AB109178
Cladophialophora_carrionii_AB109169
Cladophialophora_carrionii_AB087204
Rhinocladiella_atrovirens_AB091215
Exophiala_salmonis_AY213652
Exophiala_dermatitidis_AY213651
Exophiala_dermatitidis_AY213650
Exophiala_dermatitidis_AB087205
Exophiala_dermatitidis_AF147495
Rhinocladiella_aquaspersa_AB091214
Cladophialophora_boppii_AB109182
Cladophialophora_bantiana_AF397182
Cladophialophora_bantiana_AB091211
Cladophialophora_bantiana_AF131079
Cladophialophora_emmonsii_AB109184
Cladophialophora_arxii_AB109181
Cladophialophora_minourae_AB091213
Colletotrichum_orbiculare_AB042308
Colletotrichum_orbiculare_AB042309
Glomerella_cingulata_AB042319
Glomerella_acutata_AB042300
Glomerella_acutata_AF534467
Glomerella_acutata_AB042301
Colletotrichum_carthami_AB042307
Colletotrichum_carthami_AB042306
Colletotrichum_carthami_AB042312
Colletotrichum_caudatum_AB042305
Colletotrichum_caudatum_AB042304
Colletotrichum_higginsianum_AB042302
Colletotrichum_higginsianum_AB042303
Colletotrichum_linicola_AB046609
Glomerella_cingulata_AB042318
Colletotrichum_dematium_AB046607
Glomerella_cingulata_AB042314
Colletotrichum_boninense_AB042313
Glomerella_cingulata_AF534466
Glomerella_cingulata_AY212277
Glomerella_cingulata_CGL311883
Glomerella_cingulata_AY245021
Glomerella_cingulata_AB042315
Colletotrichum_kahawae_AF534469
Glomerella_cingulata_AB042316
Colletotrichum_kahawae_AF534468
Glomerella_cingulata_CGL311882
Glomerella_cingulata_AB042317
Glomerella_cingulata_CGL311884
Glomerella_cingulata_CGL311881
Colletotrichum_dematium_AB042310
Colletotrichum_dematium_AB042311
Colletotrichum_dematium_AB046608
Glomerella_lagenaria_GLA301970
Lycoperdon_pusillum_AB067724
Vascellum_pratense_AB067725
1
2
3
4
5

Fig 3: Dendogram showing the clustering of All9Genus database. The line noted with 4 and 5 shows first level of classification (Division : Ascomycota (4) and Basidiomycota (5)) which is clustered perfectly. The line noted with 1, 2 and 3 shows next taxonomy level of clustering (Class: Dothideomycetes (1), Chaetothynomycetes (2), Sordariomycetes(3)) with perfect clustering

- Instead of using one or two specific enzymes, sets of random enzymes are picked and the best enzyme set is determined that gives the maximum phylogenetic specificity.
- The longest common subsequence (LCS) is used not on the sequence but on the enzyme cut order.
- The approach is totally computer-oriented and requires minimum intervention from the biologists: thus it can be employed on large number of sequences.

Table 8: Obtaining optimal enzyme set for a Sequence set using Genetic Algorithm and Enzyme Cut Order property

DB	Optimal Enzyme Set (Minimum Number Of enzymes with high Clustering score)
AspCan	AatI, Hin6I, HpaII, CviRI
All9Genus	AccII, AspCNI, HaeIII,HpaII, MseI

Until today we have not found any one best efficient solution (laboratory or computational) for sequence classification, clustering and/or identification. The problem has been complicated by the sheer size of data available. We propose the use of this new property of sequence, which marries the laboratory method and computational method of sequence analysis. Due to the large size of DNA sequence databases and several possible combinations (2N where N>300) of restriction enzymes, we employ computation techniques and algorithms to prove the concept within polynomial time complexity. In the process we also show that the Longest Common Subsequence of Enzyme Cut Order is a good measure for sequence similarity. We envision that this property will open up a new domain for sequence analysis and classification, particularly in the profiling study of microbiome communities where the use of TRFLP is a widely used community profiling technique. Additionally, the use of ITS sequence, which has been previously used in the identification of bacterial organism in the laboratory setting [18] would be a recommended target sequence for such studies. Using the approach described here we will determine useful set of restriction enzyme that will return much better results than randomly chosen enzyme sets in computer simulations, and in vitro TRFLP profiling of fungal ITSs from both the known and unknown fungal species.

Acknowledgments

This research project was made possible by the NIH grant number P20 RR16469 from the INBRE Program of the National Center for Research Resources.

References

1. Snel B, Huynen MA, Dutilh BE: Genome trees and the nature of genome evolution. *Annu Rev Microbiol* 2005, 59:191-209.
2. Massol-Deya AA, Whallon J, Hickey RF, Tiedje JM: Channel structures in aerobic biofilms of fixed-film reactors treating contaminated groundwater. *Appl Environ Microbiol* 1995, 61(2):769-777.
3. Kruspe M, Stadler PF: Progressive multiple sequence alignments from triplets. *BMC Bioinformatics* 2007, 8:254.
4. Paten B, Herrero J, Beal K, Birney E: Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics* 2009, 25(3):295-301.
5. Schwartz AS, Pachter L: Multiple alignment by sequence annealing. *Bioinformatics* 2007, 23(2):e24-29.
6. Enwall K, Hallin S: Comparison of T-RFLP and DGGE techniques to assess denitrifier community composition in soil. *Lett Appl Microbiol* 2009, 48(1):145-148.
7. Ubeda JF, Fernandez-Gonzalez M, Briones AI: Application of PCR-TTGE and PCR-RFLP for intraspecific and interspecific characterization of the genus Saccharomyces using actin gene (ACT1) primers. *Curr Microbiol* 2009, 58(1):58-63.
8. Dunbar J, Ticknor LO, Kuske CR: Phylogenetic specificity and reproducibility and new method for analysis of terminal restriction fragment profiles of 16S rRNA genes from bacterial communities. *Appl Environ Microbiol* 2001, 67(1):190-197.
9. Edwards IP, Turco RF: Inter- and intraspecific resolution of nrDNA TRFLP assessed by computer-simulated restriction analysis of a diverse collection of ectomycorrhizal fungi. *Mycol Res* 2005, 109(Pt 2):212-226.
10. Liu WT, Marsh TL, Cheng H, Forney LJ: Characterization of microbial diversity by determining terminal restriction fragment length

polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* 1997, 63(11):4516-4522.

11. Buchan A, Newell SY, Moreta JI, Moran MA: Analysis of internal transcribed spacer (ITS) regions of rRNA genes in fungal communities in a southeastern U.S. salt marsh. *Microb Ecol* 2002, 43(3):329-340.

12. Klamer M, Roberts MS, Levine LH, Drake BG, Garland JL: Influence of elevated CO(2) on the fungal community in a coastal scrub oak forest soil investigated with terminal-restriction fragment length polymorphism analysis. *Appl Environ Microbiol* 2002, 68(9):4370-4376.

13. Avis PG, Dickie IA, Mueller GM: A 'dirty' business: testing the limitations of terminal restriction fragment length polymorphism (TRFLP) analysis of soil fungi. *Mol Ecol* 2006, 15(3):873-882.

14. Moeseneder MM, Arrieta JM, Muyzer G, Winter C, Herndl GJ: Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. *Appl Environ Microbiol* 1999, 65(8):3518-3525.

15. Widmer F, Hartmann M, Frey B, Kolliker R: A novel strategy to extract specific phylogenetic sequence information from community T-RFLP. *J Microbiol Methods* 2006, 66(3):512-520.

16. Roberts RJ, Vincze T, Posfai J, Macelis D: REBASE--enzymes and genes for DNA restriction and modification. *Nucleic Acids Res* 2007, 35(Database issue):D269-270.

17. Liebecq C (ed.): Compendium of Biochemical Nomenclature and Related Dcouments, Second Edition edn: Portland Press; 1992.

18. Mohamed AM, Kuyper DJ, Iwen PC, Ali HH, Bastola DR, Hinrichs SH: Computational approach involving use of the internal transcribed spacer 1 region for identification of Mycobacterium species. *J Clin Microbiol* 2005, 43(8):3811-3817.