BIOINFORMATICS TOOLS FOR CLASSIFICATION AND PREDICTION

Dinesh Kumar, Sarika and M A Iquebal
Indian Agricultural Statistics Research Institute
Library Avenue, PUSA, New Delhi 110012

1. Introduction

Molecular characterisation of genetic resources has been adding objectivity and rationality in decision making for conservation. Plant, animal, fish and microbial genetic resources are being characterised by various molecular markers, predominantly by microsatellite, AFLP and SNP covering both nuclear genome as well as mitochondrial genome. These molecular markers have inbuilt "molecular clock" entrained with evolutionary time scale having "pictures" or "signatures" of speciation and differentiation of dynamic germplasm in evolutionary pace and scale. Bioinformatics has not only revolutionised the germplasm characterisation, but had been proven as indispensable tool for molecular identification of species. Bioinformatics has become most powerful tool of taxonomy right from microbial meta-genome analysis of hitherto uncultured microbes, plant, animal and fish species identification. Advances in genome analysis technology are providing an unprecedented amount of information about animals, bacterial and viral organisms, and hold great potential for pathogen detection and identification. Here, a rational approach to the development and application of nucleic acid signatures is described based on SNP and STR nucleotides. Other bioinformatics tools for classification and prediction of such molecular data has also been discussed.

2. DNA barcoding of species and its origin

DNA barcoding is a taxonomic method that uses a short genetic marker in an organism's mitochondrial DNA to identify it as belonging to a particular species. It is based on a relatively simple concept: most eukaryote cells contain mitochondria and mitochondrial DNA (mtDNA) has a relatively fast mutation rate, which results in significant variance in mtDNA sequences between species and, in principle, a comparatively small variance within species. A 648-bp region of the cytochrome c oxidase subunit I gene (COI) was initially proposed as a potential 'barcode'.

The use of nucleotide sequence variations to investigate evolutionary relationships is not a new concept. Carl Woese used sequence differences in ribosomal RNA (rRNA) to discover archaea, which in turn led to the redrawing of the evolutionary tree, and molecular markers (e.g., allozymes, rDNA, and mtDNAvage). DNA barcoding provides a standardised method for this process via the use of a short DNA sequence from a particular region of the genome to provide a 'barcode' for identifying species. In 2003, Paul D.N. Hebert from the University of Guelph, Ontario, Canada, proposed the compilation of a public library of DNA barcodes that may be linked to named specimens. This library would "provide a new master key for identifying species, one whose power will rise with increased taxon coverage and with faster, cheaper sequencing".

2.1 Identification of birds by species bar code

In an effort to find a correspondence between traditional species boundaries established by taxonomy and those inferred by DNA barcoding, Hebert and co-workers sequenced DNA barcodes of 260 of the 667 bird species that breed in North America (Hebert et al. 2004a). It was found that every single one of the 260 species had a different COI sequence. 130 species were represented by two or more specimens. In all of these species, COI sequences were either identical or were most similar to sequences of the same species. COI variations between species averaged 7.93%, whereas variation within species averaged 0.43%. In four cases, there were deep intraspecific divergences, indicating possible new species. Three out of these four polytypic species are already split into two by some taxonomists. Hebert et al.'s (2004a) results reinforce these views and strengthen the case for DNA barcoding. They also proposed a standard sequence threshold to define new species, this threshold, the so-called "barcoding gap", was defined as 10 times the mean intraspecific variation for the group under study.

2.2 Delimiting cryptic species by DNA bar code

The next major study into the efficacy of DNA barcoding was focused on the neotropical skipper butterfly, *Astraptesfulgerator* at the Area Conservacion de Guanacaste (ACG) in north-western Costa Rica. This species was already known as a cryptic species complex, due to subtle morphological differences, as well as an unusually large variety of caterpillar food plants. However, several years would have been required for taxonomists to completely delimit species. Hebert et al. (2004b) sequenced the COI gene of 484 specimens from the ACG. This sample included “at least 20 individuals reared from each species of food plant, extremes and intermediates of adult and caterpillar color variation, and representatives” from the three major ecosystems where *Astraptesfulgerator*was found. Hebert et al. (2004b) concluded that *Astraptesfulgerator* consists of 10 different species in north-western Costa Rica. This highlights that the results of DNA barcoding analyses can be dependent upon the choice of analytical methods used by the investigators, so the process of delimiting cryptic species using DNA barcodes can be as subjective as any other form of taxonomy.

2.3 Identifying flowering plants by species DNA bar code

Kress et al. (2005) suggest that the use of the COI sequence “is not appropriate for most species of plants because of a much slower rate of cytochrome c oxidase I gene evolution in higher plants than in animals”. A series of experiments was then conducted to find a more suitable region of the genome for use in the DNA barcoding of flowering plants.

Three criteria were set for the appropriate genetic loci:

i. Significant species-level genetic variability and divergence
ii. An appropriately short sequence length so as to facilitate DNA extraction and amplification, and
iii. The presence of conserved flanking sites for developing universal primers.

At the conclusion of these experiments, Kress et al. (2005) proposed the nuclear internal transcribed spacer region and the plastid trnH-psbAintergenic spacer as a potential DNA barcode for flowering plants. These results suggest that DNA barcoding, rather than being a

'master key' may be a 'master keyring', with different kingdoms of life requiring different keys.

2.4 Strain identification of fungi

Pucciniagraminis, the causal agent of stem rust, has caused serious disease of small cereal grains (wheat, barley, oat, and rye) worldwide. *P. graminis* is the first sequenced representative of the rust fungi (Uredinales), which are obligate plant pathogens. The rust fungi comprise more than 7000 species and are one of the most destructive groups of plant pathogens. Stem rust of wheat has been a serious problem wherever wheat is grown and has caused major epidemics in North America. In 1999, a new highly virulent race TTKS (Ug99) of *P. graminis* was identified in Uganda, and since then has spread, causing a widening epidemic in Kenya and Ethiopia.

Bioinformatics can play very critical role in identification of species as well as strains and also its dynamics across globe. The plethora of data both from host and parasite generated by using latest molecular or biotechnological tools can easily be analysed by bioinformatics tools. The talk will focus on Ug99 race of *P. graminis*. How the genome of it can be used to track the movement of this fungal species and how the bioinformatics tools can be helpful in strain identifcation*P. graminis* including Ug99 identification.

3. DNA based signature of domestic species and animal breeds

Mitochondrial DNA markers have been proved to be successful in many species of domestic animals, being used especially for meat identification, poaching of wild animals, adulteration of dairy milk, dairy products(like cheese) of various domestic animal species.

The prevalent markers used for the breeds are almost STR but very recently the SNP based chip has proven its accuracy for breed signature along with details of admixture as well as very powerful for parentage and pedigree.

3.1 STR based signatures of breeds

A question has generally been asked at various scientific fora with regard to molecular characterization of breeds as to whether a livestock breed can be identified from a sample of blood, semen, hair, blood spot, carcass etc. Various attempts have been made in the last couple of years by the molecular geneticists of the world to answer this question. Some studies have succeeded in developing a technology for breed certification and breed-specific genetic/DNA signature in different breeds of cattle in Spain, Portugal and France; horses in Norway, sheep in Spain, and camel in Kenya. The degree of accuracy of certification of a breed in these studies was very high ranging between 95% to 99%.

Three methods viz (i) Frequency method (Paetkau et al., 1995), (ii) Bayesian method (Rannala et al, 1997) and (iii) Distance methods (Goldstein et al 1995) have been used for developing breed specific signatures. The Bayesian method has been reported to be more accurate with microsatellite data to the extent of > 99% confidence limits (Corander et al., 2003, Bustamante et al., 2003).

In the foreign countries, few attempts have been made to develop genetic signatures of some breeds of livestock in the recent past. For cases of doubtful breed identity where it becomes

difficult to assign an individual to a particular breed due to individual being an admixture of breeds, the studies have been made to develop breed hybrid index. The review of literature has therefore been made under two headings: (i) Development of breed-specific signatures/profiles and (ii) Development of breed hybrid index.

3.2 SNP chip based DNA signature of breeds

In Japan, Japanese Black and Holstein cattle are appreciated as popular sources of meat, and imported beef from Australia and the United States is also in demand in the meat industry. Since the BSE outbreak, the problem of false sales has arisen: imported beef has sometimes been mislabelled as domestic beef due to consumer concerns. A method is needed to correctly discriminate between Japanese and imported cattle for food safety. The SNP 50K based chip can discrimination markers between Japanese and US cattle. There is a report where five US-specific markers (BISNP7, BISNP15, BISNP21, BISNP23, and BISNP26) has been developed with allelic frequencies that ranged from 0.102 (BISNP15) to 0.250 (BISNP7) and averaged 0.184. The combined use of the five markers would permit discrimination between Japanese and US cattle with a probability of identification of 0.858. This result indicates the potential of the bovine 50K SNP array as a powerful tool for developing breed identification markers. These markers would contribute to the prevention of falsified beef displays in Japan (Suekawa*et al* 2010, Sasazaki*et al* 2011).

4. DNA based signature of plant variety, example-Basmati rice

Basmati rice has a typical pandan-like (*Pandanusamaryllifolius* leaf) flavour caused by the aroma compound 2-acetyl-1-pyrroline.Difficulty in differentiating genuine traditional basmati from pretenders and the significant price difference between them has led fraudulent traders to adulterate traditional basmati. To protect the interests of consumers and trade, a PCR-based assay similar to DNA fingerprinting in humans allows for the detection of adulterated and non-basmati strains. Its detection limit for adulteration is from 1% upwards with an error rate of ±1.5%. Exporters of basmati rice use 'purity certificates' based on DNA tests for their basmati rice consignments.It was developed at the Centre for DNA Fingerprinting and Diagnostics, Labindia, an Indian company has released kits to detect basmati adulteration. World's First Single-tube, Multiplex(co-amplify eight microsatellite loci) Microsatellite Assay-based Kit for Basmati Authentication.

The Basmati Verifiler™ Kit is the world's first product for establishing the authenticity of Basmati rice samples via a molecular assay. The kit uses a PCR amplification technique based on Simple Sequence Repeats (SSR) that provides the single most discriminating assay for Basmati genotyping.

5. DNA based bar-coded signature of fishes

Ward et al (2005) described in a paper the potential of cox1 sequencing, or ‘barcoding’, in to identification of fish species. In this study, two hundred and seven species of fish, mostly Australian marine fish, were sequenced (bar coded) for a 655 bp region of the mitochondrial cytochrome oxidase subunit I gene (cox1). Most species were represented by multiple specimens, and 754 sequences were generated. The GC content of the 143 species of teleosts was higher than the 61 species of sharks and rays (47.1% versus 42.2%), largely due to a higher GC content of codon position 3 in the former (41.1% versus 29.9%). Rays had higher GC than sharks (44.7% versus 41.0%), again largely due to higher GC in the 3rd codon

position in the former (36.3% versus 26.8%). Average within-species, genus, family, order and class Kimura two parameter (K2P) distances were 0.39%, 9.93%, 15.46%, 22.18% and 23.27%, respectively. All species could be differentiated by their cox1 sequence, although single individuals of each of two species had haplotypes characteristic of a congener. Although DNA barcoding aims to develop species identification systems, some phylogenetic signal was apparent in the data. In the neighbour-joining tree for all 754 sequences, four major clusters were apparent: chimaerids, rays, sharks and teleosts. Species within genera invariably clustered, and generally so did genera within families. Three taxonomic groups—dogfishes of the genus Squalus, flatheads of the family Platycephalidae, and tunas of the genus Thunnus—were examined more closely. The clades revealed after bootstrapping generally corresponded well with expectations. Individuals from operational taxonomic units designated as Squalus species B through F formed individual clades, supporting morphological evidence for each of these being separate species. This paper is still widely cited for DNA based fish signature.

6. Different bioinformatics tool for classification and prediction of molecular data

Advances in genome analysis technology are providing an unprecedented amount of information about animals, bacterial and viral organisms, and hold great potential for pathogen detection and identification. In this section, a rational approach to the development and application of nucleic acid signatures is described based on SNP and STR nucleotides.

Regardless of the origin of the SNPs (e.g., sequencing and public databases), once SNPs from a target organism and its nearest neighbours have been collected, it is necessary to identify those SNPs that will be useful for species and strain identification. The approach that has been taken is to use a database of SNP markers to enable phylogenetic analysis to identify evolutionary clades and the SNPs that define them. The need for large data storage capability, which facilitates data accessibility, automated SNP prediction (with reduction in manual intervention), signature delineation and facilitated complex query capability, has been recognized. Many databases exist as local resources, although some universal databases housing eukaryotic SNP data have been established (e.g., dbSNP). Such global databases have not been developed for microbial SNP data. Each database created for SNP discovery and phylogenetic analysis will have different content and different structure that are determined by the uses of the data. There is no single correct way to design a database but essential content is necessary not only to allow different polymorphism databases to communicate but to provide essential information for analysis of the data. Four essential core elements have been defined and include:

- ✓ A unique SNP identifier (allele)
- ✓ The data source (e.g., experimental or computational)
- ✓ The sequence flanking the allele and the allele(s)

Many databases have been created for the storage and analysis of eukaryotic SNP data, some are comprehensive or genomewide, and others are specialized or locus-specific. Both types of databases are essential. The comprehensive database will provide a genome-wide view of polymorphism, ideal for strain typing and identification. The locus-specific database will provide a more in-depth view of polymorphisms at a particular locus. A database should incorporate accurate information that can be used for downstream analyses and have the ability to integrate with other databases. Some additional information associated with SNPs should be implemented in the databases. A database and its associated pipeline should be able

to process and store data from a variety of sources, not only from a sequencing machine but external sequence databases (e.g., GenBank, dbEST). The database should track the organism and project to which a SNP belongs along with genome-, gene- and exon-specific information related to a SNP. A downstream analysis requires not just flanking sequences but also a reference sequence. Other information useful for quality assurance purposes and general data analysis include the algorithm by which a SNP was discovered and whether it was validated experimentally or not validated but computationally predicted and the method by which it was validated (e.g., genotyping assay or sequencing). The type of SNP should also be included (e.g., homozygous or heterozygous) along with the average allele frequency. Useful information, such as the position of the SNP relative to its reference sequence, contig or amplicon and whether the SNP is silent or pathogenic should be incorporated. To meet the needs of signature development, a relational database has been created to store information related to SNP discovery and downstream assay development. The information specific to SNP discovery and assay design is stored logically in database tables or entities enabling complex queries on SNPs and related data. Specifically, the SNP table includes, in addition to the SNP site alleles, the 5´ and 3´ flanking sequences for assay design. Information related to the gene, exon and project are stored to facilitate downstream analysis, such as population studies. Algorithm-specific rank values and method are included, which enable the investigator to assess the actual quality of each SNP. The SNP table is the central entity in the database. Associated with each SNP is a name where each SNP can have more than one name. Each SNP can also be associated with one or more reference sequences. Reference sequences have multiple purposes including:

- ✓ Serving as a template for PCR primer design
- ✓ Providing flanking sequence around a SNP
- ✓ Being included in a Phrap assembly to ensure an accurate assembly

Reference sequences also provide a starting point for functional annotation. The reference sequence has associated with it a name, GenBank accession or GI number, description and sequence. Amplicons are sequences used for SNP prediction. Associated with an amplicon is information, such as the name and description of each amplicon, primers used for its amplification and its expected size. Even though this database was designed for higher eukaryotes and their viruses, the data relationships will remain the same for prokaryotic SNP data. The SNP marker database serves as the repository of information required for downstream signature development and assay design activities.

Protocols and basic information of Bioinformatics tools which are important to search SNP, Sequence data analysis, STR data Analysis, and to develop SNP/STR based DNA signatures are shown below:

6.1GeneClass 2.0

The effectiveness of Single Nucleotide Polymorphisms (SNPs) for the assignment of various breeds of cattle and buffalo has already been investigated by analysing numerous SNPs. Breed assignment has been performed by comparing the Bayesian and frequency methods implemented in the STRUCTURE 2.2 and GENECLASS 2 software programs. The use of SNPs for the reallocation of known individuals to their breeds of origin and the assignment of unknown individuals has already been tested. Example isgiven with GeneClass2 in Buffalo having reference and unknown data of buffalo breeds (Figure 1 and Figure 2). The steps are as follows

Step 1: Download the GeneClass2 Software(Freely available at http://www.montpellier.inra.fr/URLB/geneclass/geneclass.html).
Step 2. Preparation of data files for reference and unknown samples.
Step 3. Open the main window of the software (Figure 1) and import both files.
Step 4.Choice of the parameters like Computational goal, Criteria for computation, Probability computation and Selection Criteria.
Step 5. By clicking on the start button we can see the result (Figure 2) and finally interpretation of the result can be drawn.

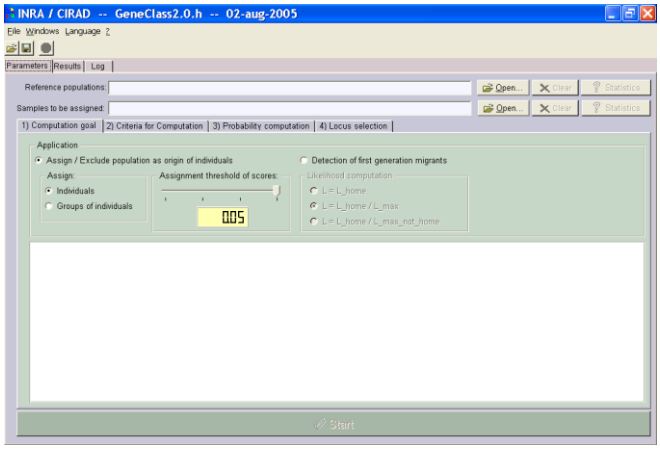
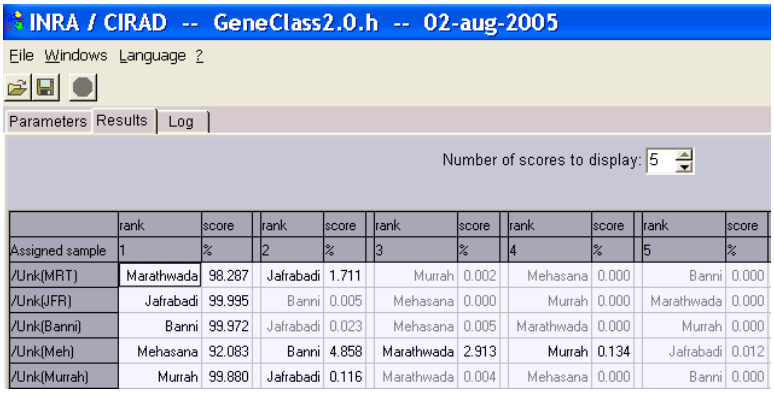


Figure 1. Main window of GeneClass2.0 Software



	rank	score	rank	score	rank	score	rank	score	rank	score
Assigned sample	1	%	2	%	3	%	4	%	5	%
/Unk(MRT)	Marathwada	98.287	Jafrabadi	1.711	Murrah	0.002	Mehasana	0.000	Banni	0.000
/Unk(JFR)	Jafrabadi	99.995	Banni	0.005	Mehasana	0.000	Murrah	0.000	Marathwada	0.000
/Unk(Banni)	Banni	99.972	Jafrabadi	0.023	Mehasana	0.005	Marathwada	0.000	Murrah	0.000
/Unk(Meh)	Mehasana	92.083	Banni	4.858	Marathwada	2.913	Murrah	0.134	Jafrabadi	0.012
/Unk(Murrah)	Murrah	99.880	Jafrabadi	0.116	Marathwada	0.004	Mehasana	0.000	Banni	0.000

Figure 2. Identification of 5 unknown breeds of Buffalo with reference data.

6.2 BioEdit

BioEdit is a mouse-driven, easy-to-use sequence alignment editor and sequence analysis tool. This tool can handle most simple sequence and alignment editing and manipulation functions that researchers are likely to do on a daily basis, as well as a few basic sequences analyses. For example alignment of different nucleotide sequence of various bacterial strains in Figure 1 and Figure 2. The steps are as follows:
File→Newalignment→Import→AccessaryApplications→ClustalWAlignment→Multiple Alignment (Figure 3) and to see the Alignment result View→ViewMode→Identity/similarity (Figure 4).

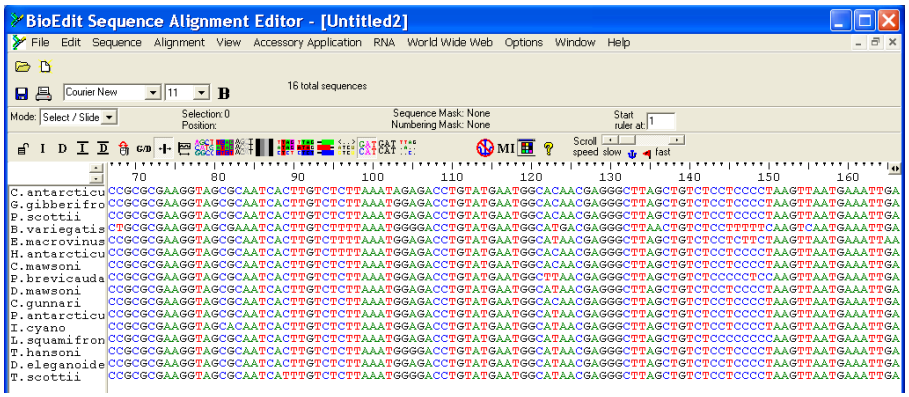


Figure 3. Nucleotide Sequence Data (16 Different Microbial strains) import in the main window

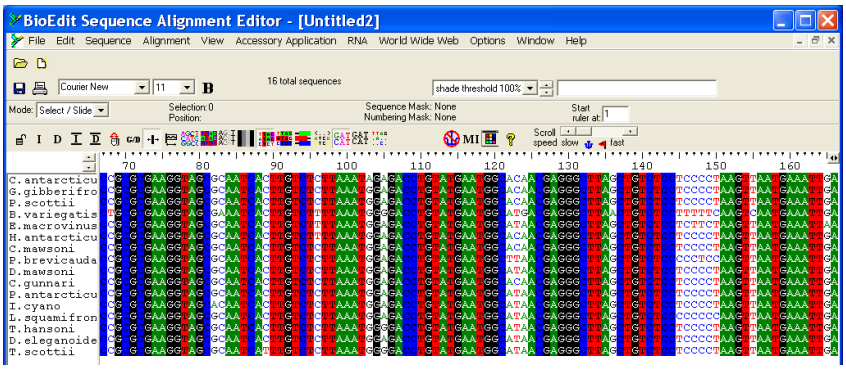


Figure 4. Alignment of all sequences showing nucleotide differences

6.3 Cleaver

Cleaver is an application for identifying restriction endonuclease recognition sites that occur in some taxa (Jarman, 2006). Differences in DNA fragment restriction patterns among taxa are the basis for many diagnostic assays for taxonomic identification; and are used in some procedures for removing the DNA of some taxa from pools of DNA from mixed sources. Cleaver analyses restriction digestion of groups of orthologous DNA sequences simultaneously to allow identification of differences in restriction pattern among the fragments derived from different taxa. Cleaver is freely available without registration from its website (http://cleaver.sourceforge.net/). The program can be run as a script for computers that have Python 2.3 and necessary extra modules installed. This allows it to run on Gnu/Linux, Unix, MacOSX and Windows platforms. Standalone executable versions for Windows and MacOSX operating systems are also available. The protocol for using the software is shown in Figure 5 and Figure 6.

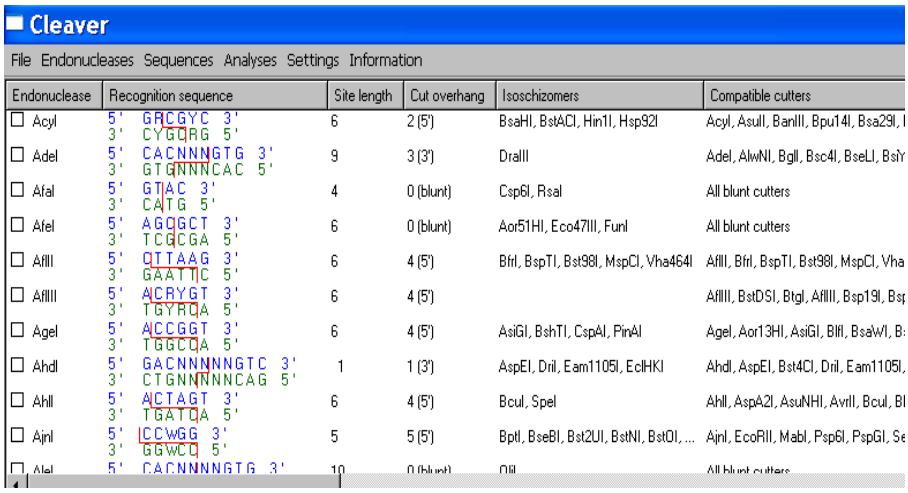


Figure 5. Main Window of Cleaver Software

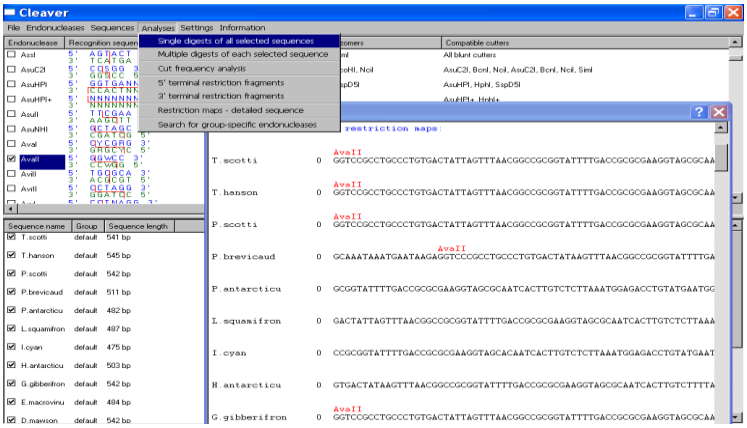


Figure 6.Restriction Map analysis of variable sequences of different Bacterial genomes using Cleaver software.

6.4 FastPCR

The FastPCRis an integrated tool for PCR primers or probe design, *in silico*PCR, oligonucleotide assembly and analyses, alignment and repeat searching (Figure 7)**.** The software utilizes combinations of normal and degenerated primers for all tools and for the melting temperature calculation are based on the nearest neighbour thermodynamic parameters.The "*in silico*" (virtual) PCR primers or probe searching or *in silico* PCR against whole genome(s) or a list of chromosome - prediction of probable PCR products and search of potential mismatching location of the specified primers or probes. Comprehensive primer test, the melting temperature calculation for standard and degenerate oligonucleotides, primer PCR efficiency, primer's linguistic complexity, and dilution and resuspension calculator. Primers (probes) are analyzed for all primer secondary structures including G-quadruplexes detection, hairpins, self-dimers and cross-dimers in primer pairs. FastPCR has the capacity to handle long sequences and sets of nucleic acid or protein sequences and it allowed the individual task and parameters for each given sequences and joining several different tasks for single run. It also allows sequence editing and databases analysis. Efficient and complete detection of various types of repeats developed (for DNA based signature) and applied to the program with a visualisation.

The program includes various bioinformatics tools for analysis of sequences with GC or AT skew, CG content and purine-pyrimidine skew, the linguistic sequence complexity; generation random DNA sequence, restriction analysis and supports the clustering of sequences and consensus sequence generation and sequences similarity and conservancy analysis.

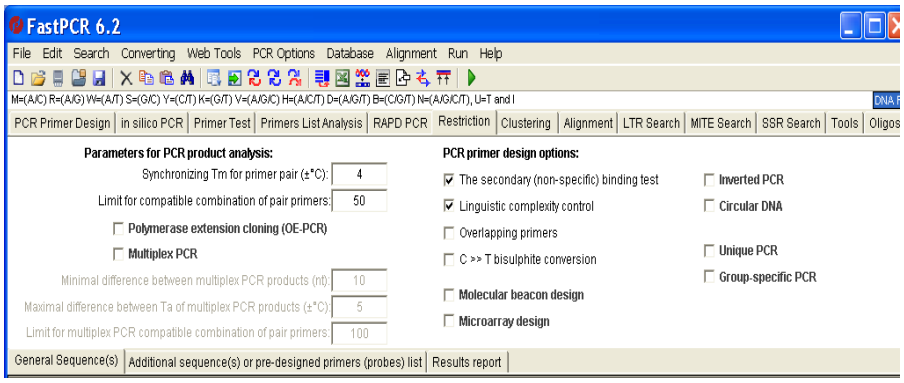


Figure 7.Main Window of FastPCR software.

For SSR search or any other analysis just we need to prepare data file in notepad file and import in the main window. As per our need we can import the data and analyse by clicking on Run/SSR search/Primer list analysis etc. option looking in main window.

References

Bustamante, CD., Nielsen, R. and Hartl, DL.(2003). Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data.*Theoretical Population Biology*.**63**: 91-103.

Corander, J.,Waldmann, P. and Sillanpaa, MJ.(2003). Bayesian analysis of genetic differentiation between populations.*Genetics*.**163**: 367-374.

Goldstein, DB., Linares, AR.,Cavalli-Sforza, LL. and Feldman, MW. (1995). Genetic absolute dating based on microsatellites an origin of modern humans. *PNAS USA*.**92**: 6723-6727.

Hebert, PDN., Penton, EH., Burns, JM., Janzen, DH. andHallwachs, W. (2004a). Ten Species in One: DNA Barcoding Reveals Cryptic Species in the Neotropical Skipper Butterfly Astraptesfulgerator. *Proc. Natl. Acad. Sci. USA***101(41)**: 14812-14817.

Hebert, PDN.,Stoeckle, MY.,Zemlak, TS. and Francis, CM. (2004b). Identification of Birds Through DNA Barcodes. *PLoS Biol*. **2(10)**: 1657-1663.

Jarman.(2006). Cleaver: software for identifying taxon specific restrictionendonuclease recognition sites. Bioinformatics Advance Access (http://bioinformatics.oxfordjournals.org/content/early/2006/06/20/bioinformatics.btl330.full.pdf.)

Kress, WJ.,Wurdack, KJ., Zimmer, EA.,Weigt, LA. and Janzen, DH. (2005). Use of DNA Barcodes to Identify Flowering Plants.*Proc. Nat. Acad. Sci. USA*,**102(23):** 8369-8374.

Paetkau, D., Calvert, W., Stirling, I. and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears.*Molecular Ecology*.**4**: 347-354.

Rannala, B. and Mountain, JL. (1997). Detecting immigration by using multi locus genotypes *PNAS, USA*. **94**: 9197-9221.

Sasazaki S., Hosokawa D., Ishihara R., Aihara H., Oyama K., Mannen, H. (2011).Development of discrimination markers between Japanese domestic and imported beef.*Animal Science Journal*,**82(1):**67-72.

Suekawa, Y., Aihara, H., Araki, M., Hosokawa, D., Mannen, H., Sasazaki, S. (2010). Development of breed identification markers based on a bovine 50K SNP array .*Meat Science,***85(2),** 285–288.