Bioinformatics Resources for Crop Improvement

What is Bioinformatics?

Bioinformatics is broadly defined as the development and application of computational tools to acquire, store, organize, retrieve, and analyze large amount of biological information. The exponential growth of sequencing and genotyping technology and the parallel growth of bioinformatics and online biological resources can successfully be harnessed for innovative breeding and pathogen diagnostic approaches.

Why Bioinformatics?

The ever increasing amount of genomic resources necessitates that biological scientists be equipped with fundamental genomic analysis tools for 'Agri-genomics' driven crop improvement. The role of bioinformatics as a pivotal tool for molecular breeding is growing steadily, particularly in areas of crop improvement, plant protection, and enhancement of quality.

- It is an integral part of molecular biology and pivotal tool for molecular breeding; enables breeders to design cost-effective and efficient breeding strategies
- Key tool to harness genomics technologies for genetic engineering and pathogen characterization and diagnosis
- Enables innovative genome-wide marker (SSR, SNP) discovery
- Widely used for functional or Gene-targeted marker development
- Orphan crops with little DNA sequence information can be studied through comparative genomics methods

Bioinformatics at IITA Bioscience Center

Sequence retrieval and manipulation. Search for a sequence of interest begins with keywords, accession number, gene name, species name, etc. The Entrez search engine at NCBI, in addition to retrieving sequences, returns pre-computed lists of data elements such as related sequences, gene, protein, taxonomy, and others. Search can be performed in all databases or restricted to nucleotide in the drop down menu. The result can be displayed in different format or downloaded, the most common format being FASTA.

Sequence alignment. Sequence alignment is the prerequisite of virtually all forms of sequence analysis ranging from search, to assembly, and to phylogenetics. Various algorithms have been developed to produce optimal alignment. Two examples of widely used open access softwares, namely BioEdit (Hall, 1999), and MEGA (Tamura et al. 2007), or using web tools such as ClustalW maintained by EMBL-EBI.

Phylogenetics. Phylogenetic analysis is the basis of taxonomical and evolutionary studies. Phylogenetic analysis is performed to cluster multiple sequences based on genetic distances. A deluge of tools and web services can also be found online (e.g. http://evolution.genetics.washington.edu/phylip/software.html).

Similarity search

Sequence comparison is essential for understanding evolutionary relationship between genes. The most common and widely used similarity search tool is BLAST (Best Local Alignment Search Tool (Ye et al. 2006). BLAST is a set of programs used to compare a nucleotide or protein query sequence to all of the available sequence databases. NCBI and EBI provide many different types of BLAST. Further information can be found at NCBI-BLAST home page (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Primer design

Primer design. There are several applications in which primer designing is required for marker development. The most widely used program for primer designing is PRIMER 3.0 (http://frodo.wi.mit.edu/primer3/) with several versions of web interface. The web-site provides user-friendly web interface and user manual describing the underlying principle of the program.

Advanced Skills

Advanced Skills. The major areas of high-end bioinformatics include the development of databases and algorithms for multiple sequence alignment, analysis and annotation of various types of microarray platforms, high-density oligonucleotide chips, variety of mass spectrometry, and diverse platforms of next generation sequencing. Programming skill allows automation of such large scale and complex jobs.

Databases and Resources

In parallel with the advances in genomic technologies that are churning out astronomical data, numerous databases and tools are being developed, making it difficult, for an average biologist, to choose from. A good starting point could be the collection of molecular biology databases published in the journal Nucleic Acid Research (http://www.oxfordjournals.org/nar/database/c/).

Comprehensive and Integrated Biological Data

NCBI and EMBL provide access to a comprehensive and integrated collection of biological data worldwide. Such integrated system is particularly useful for beginners as it eliminates the need for formatting input/output data for various interrelated suites of applications.

Online Bioinformatics Learning Resources

Suggested Online resources for self-paced tutorials and other skill building opportunities:

1. NCBI Training and Tutorials: http://www.ncbi.nlm.nih.gov/guide/training-tutorials/
2. Free Bioinformatics Educational Resource EBI (http://www.ebi.ac.uk/2can/home.html)
3. Ensembl tutorials and other help documentation for Ensembl at:http://www.ensembl.org/info/website/tutorials/index.html
4. Open Access literature, books, manuals, lecture notes, slides (via Google scholar; Scirus.com; Lab pages; wikipedia; university web sites)