Applied Mycology and Biotechnology

ELSEVIER SCIENCE

An International Elsevier Science Series

Volume 5 & 6: Genes, Genomics & Bioinformatics

List of Accepted Chapters

(Volume 5 and 6).

Genetic Recombination in Filamentous Fungi

Fred Bowring, Jane Yeadon and David Catcheside, School of Biological Sciences, Flinders University, PO Box 2100, Adelaide, 5001 AUSTRALIA (david.catcheside@flinders.edu.au).

Computational Methods in Genome Research: An Overview

G P S Raghava, Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, INDIA (raghava@imtech.res.in).

Fungal Site-Specific DNA Recombination as a Tool for Genomic

Manipulations

Makkuni Jayaram, Molecular Genetics and Microbiology, 1 University Station A5000, UT Austin, Autin, TX 78712-0162, USA (jayaram@icmb.utexas.edu).

EST Data Mining and Usage in Fungal Genomics

Peijun Zhang and Xiangjia Min, Centre for Structural and Functional Genomics, Concordia University, 7141 Sherbrooks Street West, Montreal, QC H4B1R6, Canada (jack@gene.concordia.ca).

DNA Repair and Genome Stability in Fungi

Steeve Harris, Plant Science Initiative, University of Nebraska, N234 Beadle Center, Lincoln, NE 68588-0660, USA (sharri1@unlnotes.unl.edu).

Gene Silencing as Tool for Identification of Gene Function in Fungi

Giuseppe Macino, Dipartimento di Biotecnologie Cellulari ed Ematologia, Sezione di Genetica Molecolare, Policlinico Umberto I°, Viale Regina Elena 324, 00161 Roma, Italy. (macino@bce.uniroma1.it).

Heterologous Gene Expression in Fungi

Helena Nevalainen, Department of Biological Sciences, Macquarie University, Sydney NSW 2109, Australia (hnevalai@rna.bio.mq.edu.au).

Genomic Diversity of Industrial Enzymes of Fungal Origin

Said S[1]. Pietro R[2]. and R Prade[3], [1]Faculdade de Ciências Farmacêuticas ,Universidade Estadual Paulista Júlio de Mesquita, Araraquara SP, Brazil; [2]Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto SP, Brazil; [3] Department of Microbiology, Oklahoma State University, Stillwater OK, USA.

An Overview of Genetics of Morphogenesis in Fungi

Scott Gold, Plant Pathology,University of Georgia, Miller Plant Science Bldg Athens, GA 30602-7274, USA (sgold@arches.uga.edu).

Gene Networks of Stress Response in Yeast

Irina L. Stepanenko, Laboratory of Theoretical Genetics, Institute of Cytology and Genetics, Lavrentyeva 10, Novosibirsk, 630090, Russian Federation (stepan@bionet.nsc.ru).

Molecular Genetics of Signaling in Fungi

Yong-Hawn Lee, Divison of Plant Pathology, Agricultural Biology, Seoul National University, Suwon, 441-744 KOREA (yonglee@plaza.snu.ac.kr).

Expression and Engineering of Hydrophobin Genes in Fungi

[1]Karin Scholtneijer and H.A.B. W.sten[2], [1]Department of Plant Biology, University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands; [2]Microbiologie, H.R. Kruytgebouw, Universiteit Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands. (h.a.b.wosten@bio.uu.nl).

Enumeration of Cellulose and Biomass Induced Genes from *Trichoderma* Using DNA Microarrays

*Elena Bashkirova and Randy Berka, **Novozymes Biotech, Inc., 1445 Drew Avenue, Davis, CA 95616-4880, USA (Rambo@novozymesbiotech.com).***

The *cfp* Gene of *Neurospora crassa*: a Tale of a Gene, Two Transcripts,

One Enzyme and a Bunch of Filaments

1E.D. Temporini, 2**H.D. Folco** and 3**Alberto Luis Rosa,** 1Division Of Plant Pathology and Microbiology, The Department of Plant Sciences, College of Agriculture and Life Sciences, Forbes Building 204, P.O. Box 210036, Tucson AZ 85721-0036 , USA (etempor@ag.arizona.edu); 2University of Edimburgh, Scotland; 3 Inst de Méd "Mercedes y Martín Ferreyra, INIMEC - CONICET, Lab. de Neurogenetica, Friuli 2434-Colinas de Velez Sarsfield, Cordoba, 5016 ARGENTINA (arosa@immf.uncor.edu).

Genetic Regulation of Carotenoid Synthesis in Fungi

J. W+stemeyer, Friedrich-Schiller-Universitat Jena, Lehrstuhl fuer Allgemeine Mikrobiologie und, Mikrobengenetik, Neugasse 24-D-07743 Jena, Germany (B5wojo@rz.uni-uni-jena-de).

Heterokaryosis as a Tool for Discovery and Expression of New Pharmaceuticals in Filamentous Fungi

Edward B. Cambareri and W. Dorsey Stuart, Neugenesis Corporation, Industrial Road, Suite J, San Carlos, CA 94070, USA (dstuart@neugenesis.com).

A Search for Developmental Gene Sequences in the Genomes of Filamentous Fungi

David Moore, Conor Walsh, Geoffrey D. Robson, School of Biological Sciences, The University of Manchester, 1.800 Stopford Building, Manchester M13 9PT, United Kingdom (david.moore@man.ac.uk; conor.walsh@stud.man.ac.uk; geoff.robson@man.ac.uk; URL: www.oldkingdom.com).

Bioinformatics Packages for Sequence Analysis

Yeisoo Yu and Sangdum Choi, Division of Biology 147-75, California Institute of Technology, Pasadena, CA 91125, USA (schoi@caltech.edu).

Computing and Genomics

Gautam B. Singh, Computer Science and Engineering, Oakland University Rochester, MI 48309, USA (singh@oakland.edu).

Phylogenetic Network Construction Approaches

1Vladimir Makarenkov and 2Pierre Legendre, 1Computer Science, Universite du Quebec a Montreal, Canada (makarenv@magellan.umontreal.ca) and 2Biological Sciences, Universite de Montreal, Canada (Pierre.Legendre@UMontreal.CA).

Comparative and Functional Genomics of Pathogenic Oomycetes

Sophien Kamoun and Mark Waugh, 1Department of Plant Pathology, Ohio State University

Ohio Agricultural Research and Development Center 1680 Madison Ave. Wooster, OH 44691, USA (Kamoun.1@osu.edu; http://www.ncgr.org); Bioinformatics Group, National Center for Genome Resources, 2935 Rodeo Park Drive East Santa Fe, NM 87505, USA (mew@ncgr.org http://www.ncgr.org).

Alternaria Genomics

Christopher B. Lawrence, Virginia Bioinformatics Institute, 1880 Pratt Dr., Bldg. XV (0477), Virginia Tech, Blacksburg, VA 24061-0477, USA (Lawrence@vbi.vt.edu).

Molecular Genetics and Genomics of *Colletotrichum*

Amir Sharon and Stan Freeman Dept. of Plant Pathology and Weed Research, ARO, The Volcani Center, Bet Dagan 50250, ISRAEL (freeman@volcani.agri.gov.il).

Genomics of *Aspergillus oryzae*

Tetsuo Kobayashi, Graduate School of Bioagricultural Sciences, Nagoya University, Chikusa, Nagoya 464-8601, Japan (koba@agr.nagoya-u.ac.jp).

Genomics of Phanerochaete

Dan Cullen, University of Wisconsin (dcullen@facstaff.wisc.edu).

Teaching in Fungal Genomics and Bioinformatics

Chuck Staben, School of Biological Sciences, University of Kentucky
101 Morgan Hall, Lexington, KY 40506-0225 **(**staben@pop.uky.edu; **URL:** biology.uky.edu).

Fungal Introns and Other Repetitive DNA Elements in Fungi

Jack Kennell (Department of Biological Sciences, Southern Methodist University, 220 Fondren Science, Dallas, TX 75275-0376, USA e-mail: jkennell@mail.smu.edu) or Stefanie Pöggeler (Department for General and Molecular Botany, Ruhr-University Bochum, 44780, Bochum, Germany; stefanie.poeggeler@ruhr-uni-bochum.de)

Other volumes in this series:
Volume 1: Agriculture and food Production (2001).
Volume 2: Agriculture and food Production (2002).

Volume 3: Fungal Genomics (2003).

Volume 4: Fungal Genomics (2004).

Bioinformatics Packages for Sequence Analysis

Yeisoo Yu* and Sangdun Choi†

***Yeisoo Yu**
Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, USA

†**Sangdun Choi**
Division of Biology 147-75, California Institute of Technology, Pasadena, CA 91125, USA

*Correspondence should be addressed to:
Sangdun Choi
Division of Biology, 147-75
California Institute of Technology
1200 East California Boulevard
Pasadena, CA 91125, USA
Tel: (626)395-8732, Fax: (626)796-7066
E-mail: schoi@caltech.edu

ABSTRACT

Research paradigms in modern biology are shifting from single gene to genome-wide scale. Two major contributors to this new trend are large-scale genome sequencing and bioinformatics. Recently, bioinformatics emerged as a new science field that provides computational tools for collecting and maintaining complex biological data. Along with an exponential accumulation of sequence data, many bioinformatics software and algorithms have been developed to assist the genome scale analysis. A comprehensive knowledge of those tools can help not only to understand gene functions and genome organizations but also to provide an opportunity to develop new tools that can answer many biological questions.

1. INTRODUCTION

The amount of sequence information available from public database is exponentially increasing. By 2003, 36.5 G bases of sequences representing 31 million entries were deposited into GenBank. The database is approximately 170 times larger compared to a decade ago. Advanced sequencing technologies and model organism genome projects, including human, were the major driving forces behind the sequence information explosion during the last decade. This genome data will provide fundamental information for biological and biomedical researches to extend our knowledge for better understanding gene functions and regulations of different model organisms from evolutionary related species.

Today's biological researches require parallel strategies to simultaneously gather, examine and integrate the large volumes of information. Biologists often face genome-wide or cross-genome analysis of genes of interest. Thus, without good data handling skill, researchers can not achieve their ultimate research goals. This is where biology requires informatics. Bioinformatics is a new science field that examines complex biological data on the basis of statistics and computer science. It provides a powerful tool for collecting, maintaining, distributing and analyzing huge amounts of genome data. It also contributes to give biological meaning in the data by discovering structural and functional relationships to explain biological phenomena.

Many sequence analysis tools have been developed and successfully used for interpreting genome data. As biologists, we are using one or more programs on a daily basis without knowing which software is more suitable to analyze data. In this chapter, we described several bioinformatics programs which are commonly used in genome sequencing to make sense of sequence assembly, similarity search, repeat identification and gene annotation. Of course, there are more programs available and we leave these challenges to you.

2. NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI)

NCBI was established in 1988 as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). Its mission is defined as development, distribution and maintenance of various molecular databases and computer software in order to support biological and biomedical studies at the molecular level. Regardless of the complicated NCBI structure, it is divided into two major aspects, in terms of data flow: sequence submission and retrieval.

2.1. Sequence Submission System

Submission program
GenBank is the sequence depository place, which provides two programs to support sequence submission, BankIt and Sequin.

BankIt (http://www.ncbi.nlm.nih.gov/BankIt/) is a web based sequence submission tool that can be used for depositing a few sequences when annotation is not complicated. Submission is accomplished in four steps: general submission information (contact information and release date), reference information (author, publication and citation information), source information (organism and source description), and input sequence (molecular type and sequence). BankIt does not require any special tools to submit sequences other than web browser and the submission directions are fairly easy to follow.
Sequin (http://www.ncbi.nlm.nih.gov/Sequin/index.html) is a stand-alone program to submit and update long complex sequences and annotation information. It runs on Macintosh, PC and UNIX operating systems and is available from NCBI Sequin ftp site (ftp://ftp.ncbi.nih.gov/sequin/) with documentation and instruction. Sequin has a restriction to read input files, thus submitters must prepare their sequences by following specific instructions (FASTA file is the standard format). Though more steps are involved in Sequin submission, it provides sophisticated tools to review and verify the sequence and annotation before submission. Submission is finished by sending the sequin output file (.sqn file) via e-mail to GenBank. A Sequin quick guide is available from Sequin web site at http://www.ncbi.nlm.nih.gov/Sequin/QuickGuide/sequin.htm.

GenBank division for submission
GenBank maintains databases according to the nature of the DNA sequence. Submitters have a choice of divisions to deposit their sequences to GenBank based on the source of sequences. It is currently categorized into 17 divisions listed in Table 1. Divisions of PRI, ROD, MAM, VRT, INV, PLN, BCT, VRL and PHG contain only sequences from specific organisms whereas EST, HTG, STS and GSS contain sequences generated by specific technologies from various organisms.

dbEST: Expressed sequence tags (EST) are short and single pass sequence from mRNA via cDNA (complimentary DNA) cloning procedure (Adams et al. 1991). It represents gene expression profiles in a specific cell, tissue and organ, or in a specific developmental stage in a normal or stressed growth condition. Currently 23 million entries are available from GenBank (release 081304; http://www.ncbi.nlm.nih.gov/dbEST/index.html).

dbSTS: Sequence tagged sites (STS, Olson et al. 1989) contain short, unique sequences on chromosomes or genomes used to generate genetic maps. About 374,000 STSs are available in GenBank (release 073004; http://www.ncbi.nlm.nih.gov/dbSTS/index.html).

dbGSS: Short, single pass sequences from genomic DNA origin are deposited in GSS (Genome Survey Sequence) division. Entries are comprised of genomic sequences from exon trapping, Alu PCR and end sequences of large insert genomic clones such as BAC, cosmid, fosmid and YAC (Venter et al. 1996; Mahairas et al. 1999; Siegel et al. 1999; Batzoglou et al. 1999). About 9.4 million entries are available from GenBank (release 081304; http://www.ncbi.nlm.nih.gov/dbGSS/index.html).

dbHTG: High-throughput genome sequences (usually called shotgun sequences) from large scale genome sequencing projects are deposited into HTG division (Ouellette and

Boguski 1997). Based on the degree of finishing, phase number is divided to 3 types: Phase1 submission means unfinished, and sequence contigs are not ordered. Phase2 sequences are also unfinished but sequence contigs are ordered. Phase3 is finished sequence achieved contiguity with less than 1 in 10,000 bases. Finished sequences are transferred to the organism specific databases (e.g., PRI, MAM, PLN, etc.)

WGS: Assembled contigs and annotation data from whole genome shotgun (WGS, Fleischmann et al. 1995; Venter et al. 1998) sequencing projects are submitted into GenBank. Nucleotide sequences are transferred to BLAST WGS and protein sequences go to BLAST non-redundant (nr) database. Scaffold or supercontig information can be submitted to GenBank with specific format (agp format) that contains contig orders and orientation information. Currently 125 WGS projects including human and mouse are listed in GenBank and detailed information can be found at http://www.ncbi.nlm.nih.gov/Genbank/WGSprojectlist.html.

2.2. Sequence Retrieval System

NCBI's Entrez (http://www.ncbi.nlm.nih.gov/Entrez/index.html) is an integrated database retrieval system. Its cross-reference system allows researchers to not only access nucleotide, protein or genome information but also related research articles and relevant records from 22 databases using text based query. Query search can be refined using Boolean Operators; "AND", "OR" and "NOT". Figure 1 is an example of Entrez search using text query as "callose synthase AND plant". The number of positive entries from the query is displayed and 25 nucleotide sequences with pull-down display menu are shown in Fig. 1.
Batch Entrez (http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?) is also a convenient tool to retrieve large number of sequence data at once. Batch Entrez can read a text file containing either GI or accession number in one entry per line and provide the sequences based on user preference of the data format. Detailed help documentation is also available at http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html.

3. SEQUENCE ANALYSIS TOOLS

3.1. Sequence Assembler

Two well known methods are applied to generate genome sequences. One is called clone by clone (CBC) approach and other is whole genome shotgun (WGS) method. Figure 2 shows the general procedure for both CBC and WGS methods. Both methods create many small overlapping sequences or reads, which are eventually assembled by computer software to build the sequence contigs. Clone by clone method sequences shotgun clones derived from minimum tiling BAC (bacterial artificial chromosome) or PAC (P1-derived artificial chromosome) clones. Progress is slow using CBC method but this method generates high quality data across the genome, even in highly repetitive regions, whereas, WGS generates sequences from shotgun clones derived directly from genomic DNA and sequences from various insert sizes of genomic clones (10-150kb). Progress is fast with WGS but it permits low quality data and mis-assembly in repeat regions.

Phred/phrap/consed package

Phred, phrap and consed package is UNIX based software most widely used in CBC shotgun sequence assembly and it provides the standardized quality assurance for many genome projects.

Software phred (Ewing et al. 1998; Ewing and Green 1998) calls bases by reading electropherogram or trace files as raw data and assign the quality value to each base. Quality value or phred value is a base call error probability and it is calculated by the formula:

$QV = -10 \times \log_{10}(p)$

where, p is the probability that the base call is an error. Phred value of 20 considers one base call error in 100 bp (p=0.01) and phred value 30 means one base call error in 1000 bp (p=0.001). Generally, a phred value greater than or equal to 20 is considered as high quality sequence. Phred generates FASTA formatted sequence, quality and PHD files that can be used to assemble sequences using phrap.

Software phrap (http://www.genome.washington.edu/UWGC/analysistools/Phrap.cfm) is a program for assembling shotgun sequences based on sequence overlap. Before assembly, vector sequences are removed from an individual read using program cross_match, which comes with the package. Three input files are required to run phrap; vector screened sequence, quality and phd (phred) files. Forward and reverse sequences along with chemistry are recognized by phrap when pre-defined naming convention is used. Manyreads and longreads versions of phrap and cross_match are recommended when more than 64,000 reads are assembled or a sequence longer than 64,000 bp is included in a single assembly. Software consed is required in order to display and edit the assembly by reading phrap output file (.ace file).

Consed (Gordon et al. 1998) is a program for viewing and editing phrap assembly in its finishing phase. It shows global assembly view with forward-reverse pairs, read depth and repeat match. Thus, this information can be easily used for finishing procedures. Independent from phrap assembly, Consed also allows breaking or joining contigs by comparing contigs. *In silico* digestions can be generated and compared with real digestions to verify the overall assembly in final finishing. For primer walk in finishing phase, Consed provides built-in primer picking function. Figure 3 shows examples of Consed screen shots. Autofinish (Gordon et al. 2001) is also a good finishing tool and is part of Consed program. It generates an experiment list focused on filling gaps, improving quality in low quality areas and determining the orientation of contigs.

Whole genome assembler

Unlike CBC assembly, whole genome assembly is a challenging procedure because it processes hundreds of thousands of reads at a time so repeat sequences possibly cause mis-assembly. Development for WGS assembler thus focuses on reducing computation time and resolving mis-assembly problems caused by repeat sequences in the genome.

Arachne (Batzoglou et al. 2002; Jaffe et al. 2003) uses quality score associated sequences for whole genome shotgun sequence assembly. It utilizes the forward and reverse pair information within a similar insert size library and removes low quality vector and other contaminated sequences in an initial assembly. It identifies potential repeats by clustering reads, excludes them from the assembly and merges overlapping read pairs to make contigs. Read pairs from larger insert clones are used later to build supercontigs. Arachne simulated assembly of the *Drosophila* genome showed about 98% coverage with N50 contig length of 324 kb and N50 supercontig length of 5143 kb. Mouse whole genome assembly was made possible with Arachne2. The contigs have an N50 length of 24.8 kb, whereas the supercontigs have an N50 length that is approximately 700-fold larger at 16.9 Mb.

PCAP (Huang et al. 2003) is a contig assembly program using parallel computer processors. First, it removes vector and low quality area from reads. Then, it uses BLAST2 to identify pairs of reads that contains potential overlaps. Identification of repetitive regions in reads is based on deep coverage by longer approximate matches. The score of every overlap is adjusted to reflect the depths of coverage for the two regions in the overlap. The consensus sequence of a contig is generated by constructing an alignment of reads in the contig. Human chromosome 20's assembly is simulated with PCAP and it showed N50 contig with scaffold length of 41 kb and 2Mb, respectively.

Phusion (Mullikin and Ning 2003) first groups sequence reads by determining the number of times that sequences of length k (called *k*-mer word) occur in the data, and eliminates reads representing highly redundant *k*-mer sequences. It generates reads list and matrix based on reads showing less repetitive or unique *k*-mer distribution. Phrap uses paired reads information along with the above information to assemble sequences. An iterative phrap with read pairs from different size insert clones will merge contigs and make supercontigs.

EST clustering

Expressed Sequence Tag (EST) is useful information in a sense that it is a profile of expressed gene sequences. It usually does not contain full length gene sequence because about 600 bp sequences are generated from 5' and 3' end of cDNA clones. At the same time EST permits low quality bases due to single pass sequencing and often times some sequences are highly redundant in certain genes. In order to overcome these disadvantages and collect more unique sequences (called UniGene), clustering ESTs is necessary. Phrap, TIGR assembler (Pop and Kosack 2004) and CAP3 are used to cluster or assemble EST data and among those, CAP3 developed by Xiaoqiu Huang (1999) is one of the popular programs.

CAP3 takes FASTA sequence file as an input for assembly, and two additional files, FASTA quality and forward and reverse pair information, can be used to correct assembly. It removes low quality regions from 5' and 3' end of sequences and detects overlaps between input reads followed by joining them to make contigs. Forward and reverse information (constraints) is used to correct assembly, and then it writes consensus sequences and quality value in each base. The output or assembly file (.ace file) can be

viewed using Consed program. Lucy is a program to prepare raw DNA sequences for EST or shotgun assembly. It removes low quality and vector sequences in raw data and provides high quality sequence and quality files for assembly (Chou and Holmes 2001).

3.2. Pairwise and Multiple Sequence Alignments.

Sequence alignment is a daily task of most biologists to find the relationship or similarity between biological sequences. Pairwise sequence alignment tries to find the optimal alignment in parts of sequences (local alignment) or in entire sequences (global alignment). In a global alignment, all of the nucleotides or amino acids in both sequences participate in the alignment, thus it is useful for aligning closely-related sequences. Local alignment finds and aligns related regions within sequences. It is more flexible than global alignment, which is useful to identify related regions that appear in a different order in two sequences.

Multiple sequence alignment is an extension of pairwise alignment to identify common regions within several sequences given as an input. This tool is mostly used for building phylogenetic trees and also creating sequence profiles which can be used to search distant related sequences in database.

BLAST

BLAST (Basic Local Alignment Search Tool) is the most popular local alignment program for similarity search and sequence alignment developed by NCBI (Altschul et al. 1990; Altschul et al. 1997). BLAST algorithm generates a list of short word match (default words size is 3 for protein and 11 for nucleotide) in query sequences and then database is searched for the occurrence of these words. The matching words are extended to local alignment between two sequences and extensions are continued until score is below the threshold. BLAST search can be performed at NCBI's BLAST server using web browser or at any local computer by installing the BLAST software (stand-alone BLAST). Stand-alone BLAST reduces searching time significantly by avoiding on-line communication and allows batch blast (submit multiple queries at once) against local databases downloaded from GenBank or created by user. Stand-alone BLAST document including installation can be found at NCBI's FTP site (ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blast.txt).

Several BLAST programs are available based on search purpose and also based on query and database relation. FASTA file format is a standard input requirement for all BLAST programs and database format. Using formatdb program is necessary in order to do local BLAST search.

Nucleotide search: BLASTN is used to compare nucleotide query sequences against nucleotide databases. NCBI provides several databases to compare query sequence against. Table 2 showed NCBI databases for BLAST search. MegaBLAST uses greedy algorithm (Zhang et al. 2000) to perform nucleotide search using word size of 28 as a default (while BLASTN is 11 in word size), which makes search 10 times faster in closely related sequences. Web MegaBLAST allows multiple queries search with FASTA format sequences or accession numbers.

Protein search: BLASTP program is used to compare amino acid sequence query against amino acid sequence database (Table 2) using BLOSUM62 (Henikoff and Henikoff 1992) as a default substitution matrix. More specialized protein search can be done with PHI-BLAST and PSI-BLAST (Altschul et al. 1997). Pattern-Hit Initiated (PHI)-BLAST is designed to search for proteins that contain a pattern in the query specified by the user. Position-Specific Iterated (PSI)-BLAST is designed to find very distantly related protein sequences using a PSSM (position-specific scoring matrix) generated from each progressive search. PSI-BLAST is the most sensitive BLAST program.

Translated query or database search -
BLASTX: Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. This program can be used to find potential translation products of an unknown nucleotide sequence.
TBLASTN: Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
TBLASTX: Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Blast 2 sequences (bl2seq): bl2seq (Tatusova and Madden 1999) is designed to directly compare two sequences without format database. Most programs mentioned above are used to align two query sequences such as BLASTN (1^{st} query- nucleotide and 2^{nd} query-nucleotide), BLASTP (protein and protein), BLASTX (nucleotide and protein) and TBLASTN (protein and nucleotide).

FASTA
Another local alignment tool is FASTA program developed by Pearson and Lipman (1988) which is available over the Web or by download. FASTA searches short sequences (called k turples, which are similar to words in BLAST) in query and database to identify ungapped alignments. Then, the alignments are tested and merged into a local alignment to find optimal local alignment based on threshold and score. FASTA provides programs similar to BLAST but it also performs global alignment which is not provided in BLAST. FASTA program is used to compare nucleotide query to nucleotide database or protein query to protein database (equivalent to BLASTN and BLASTP). FASTX/FASTY is the same as BLASTX in comparing translated DNA against protein database. TFASTX/TFASTY is used to compare protein query against translated nucleotide database (similar to TBLASTN). ALIGN performs global alignment between two protein or nucleotide queries. Current FASTA programs can be found at URL http://fasta.bioch.virginia.edu/.

BLAT
BLAT stands for BLAST-like alignment tool developed by Jim Kent (2002) and it is designed to effectively align EST sequence to genomic sequences. Its algorithm is similar to BLAST in which it finds words match and extends to high scoring pairs (HSP). However, BLAT makes index of database first then searches query. Moreover, it extends alignments on any number of perfect or nearly perfect hits and provides a large alignment. BLAT also gives unspliced alignment between RNA and nucleotide sequence

so that splicing site is easily detected in the alignment. Web search is available at http://genome.ucsc.edu/cgi-bin/hgBlat and stand-alone is also available at http://www.soe.ucsc.edu/~kent/exe/. Standard output is in tabular format (tab delimited .pls file) and BLAST pairwise output is also available.

CLUSTALW

Multiple sequence alignment is a tool to study closely related genes or proteins to find the evolutionary relationships between genes and to identify shared patterns among functionally or structurally related genes. A popular program for multiple sequence alignment is CLUSTALW (Higgins et al. 1994; http://www.ebi.ac.uk/clustalw/). Both progressive global and local alignments can be done in ClustalW. The user has the option to control parameters to make the best alignments (e.g., word size, matrix, gap open, extension, etc.). It also provides two guide phylogenic trees, cladogram (equal length of branched tree showing common ancestry) or phylogram (unequal length of branched tree showing evolutionary distances). Alignment can be further edited using Jalview program (Clamp et al. 2004; http://www.ebi.ac.uk/jalview/).

3.3. Gene Prediction

Gene prediction is one of the important subjects in genome projects. In eukaryotes, gene prediction and annotation is not a simple process mainly because of the various sizes of introns (uncoding sequences) located between exons (coding sequences). In addition, many genes have alternative splice variants. In other words, every eukaryotic gene shows different structures and length to be predicted.

Many gene prediction programs mainly categorized into three groups have been developed for genome wide annotation. The first group is using *ab initio* approach to predict genes directly from nucleotide sequences. Prediction programs in this group utilize statistical model to differentiate the promoter, coding or noncoding regions, and intron-exon junctions in genomic sequences. Hidden Marcov Models (HMMs) is a popular model and applied to make gene prediction programs such as GENSCAN (Burge and Karlin 1997; Burge and Karlin 1998), FGENESH (Solovyev et al. 1995; Salamov and Solovyev 2000), Grail (Xu et al. 1994), MZEF (Zhang 1997), HMMgene (Krogh 1997). The second group is a similarity based approach to identify gene structure using sequence alignment between genomic sequence and transcript (EST and cDNA) or protein databases. This approach is recently expanded to genomic sequence comparison (comparative approach) between evolutionary related species to identify functional regulatory elements which tend to be conserved through evolution. CRASA (Chuang et al. 2003), AAT (Huang et al. 1997) and AGenDA (Taher et al. 2003) belong in this group. The third group is combining the method of *ab initio* and similarity based approaches. GenomeScan (Yeh et al. 2001), GeneWise (Birney et al. 2004), FGENESH+ (Salamov and Solovyev, 2000), and Procrustes (Gelfand et al. 1996; Mironov et al. 1998) are available in this approach. Table 3 showed prediction programs listed above.

Splicing site prediction is important in order to choose the correct gene models in basis of accurate intron-exon boundaries. Many programs use either computational models based on consensus dimer sequence in donor sites, acceptor sites and branch point (about 30bp

upstream of acceptor site) or sequence alignment between transcript and genomic sequences, or both to predict splicing sites in genomic sequences. NetGene2 (Hebsgaard et al. 1996; http://www.cbs.dtu.dk/services/NetGene2/), GeneSplicer (Pertea et al. 2001; http://www.tigr.org/tdb/GeneSplicer/gene_spl.html) or SplicePredictor (Brendel and Kleffe 1998; http://bioinformatics.iastate.edu/cgi-bin/sp.cgi) is used for splicing site prediction.

tRNAScan-SE (Lowe and Eddy 1997) identifies transfer RNA genes in genomic sequences by searching conserved A & B box promoter sequence followed by progressively identifying various stem-loop structures. It provides tabular and secondary structure as the standard output. tRNA analysis is performed on-line at http://www.genetics.wustl.edu/eddy/tRNAscan-SE/.

Although some areas of the genome is relying only on *ab initio* or similarity based due to prediction failure or lack of experimental data, a combined approach generally increases the accuracy in gene annotation.

3.4. Repeat Identification

Repetitive sequences occupy a large portion of most eukaryotic genomes and are divided into tandem (including simple sequence repeats or SSRs) and interspersed repeats. Transposable elements (TEs), one of the interspersed repeats, are the most abundant in repeat family. Transposable elements are further classified into two groups: class I and class II. Class I TE is called retrotransposon and it transposes using RNA intermediate. This element encodes reverse transcriptase and other viral proteins (*gag* and *pol*). It is subdivided into LTR type (*gyp*sy and *copia* group) and non-LTR type (long interspersed repetitive elements [LINE] and short interspersed repetitive elements [SINE]) based on long terminal repeats (LTRs). Class II TE is transposon and DNA intermediate. It has terminal inverted repeats (TIRs) and encodes transposase that moves from one position to another in a genome. *P* elements in fruit fly and *Ac, Spm* and *Mu* elements in maize are well studied transposon (Girard and Freeling 1999; Wessler 2001).
Recently, another type of repeat element called miniature inverted-repeat transposable element (MITE) has been identified. MITEs are usually less than 600 bases in size and also have short terminal inverted repeats (TIRs). It can be divided into two groups based on TIRs and target site duplication; *Tourist* and *Stowaway* (Bureau and Wessler 1992; Bureau and Wessler, 1994; Jiang et al. 2003).

Repeat finding programs more often use a similarity based approach to find and annotate the repeat regions in genomic sequences, although Juretic and co-workers (2004) made some efforts to annotate TEs in rice genome with HMMER (Eddy 1998) basis of Hidden Marcov Models.

RepeatMasker

RepeatMasker (Smit, AFA and Green, P, RepeatMasker at http://repeatmasker.org) is a wildly used program to find interspersed repeats (LINEs, SINEs, LTRs and DNA elements), simple sequence repeats (SSRs) and low complexity regions in the sequences using similarity search against well defined repeat database. RepeatMasker uses

"cross_match" program as a default search engine which accounts for the high sensitivity but slow speed when long sequences are searched. WU-BLAST is added as an optional search engine to increase search speed. User defined repeat database can be used to search against in stand-alone RepeatMasker.

RECON
RECON (Bao and Eddy 2002) allows *de novo* detection and classification of repeat family in genomic sequences. RECON algorithm detects and groups repeats in the genome sequences by blast itself and clusters them to a repeat family using multiple sequence alignment. This approach helps repeat annotation by determining repeat boundaries in genomic sequences and also enables identification of new repeat elements. RECON is available from http://www.genetics.wustl.edu/eddy/recon/.

3.5. Other Programs

PipMaker
PipMaker (Schwartz et al. 2000) is a tool to align two sequences and generates a percent identity plot (PIP) and dot plot as output. It requires two FASTA format sequences and repeat information from RepeatMasker and exon information can be optionally submitted. MultiPipMaker allows submitting multiple FASTA sequences (up to 20) and perform multiple sequence alignment in order to analyze relationship among input sequences. PipMaker analysis can be performed at http://pipmaker.bx.psu.edu/pipmaker/. zPicture program (Ovcharenko et al. 2004; http://zpicture.dcode.org/) also provides more dynamic alignment and visualization of comparing two sequences.

rVISTA
Regulatory Vista (rVISTA, Loots and Ovcharenko 2004) is a computational tool to identify evolutionary conserved transcription factor binding sites (TFBSs) by multiple alignment of orthologous sequences followed by prediction of TFBSs using TRANSFAC database (Matys et al. 2003) collected from eukaryotic transcription factors. Output from zPicture can be transferred to rVISTA program without further modification. This program can assist in understanding the function of conserved non-coding sequence and identify the potential *cis*-regulatory elements in genomic sequences. rVIST2.0 is available at http://rvista.dcode.org/.

MUMmer
MUMmer (Delcher et al. 2002; Kurtz et al. 2004) is a tool that allows rapid alignment of two large nucleotide or protein sequences or even genome to genome for comparison. It uses suffix tree algorithm to find minimum 20bp exact matches as an alignment anchor and then extends the alignment to generate pair-wise alignment like BLAST. MUMmer is comprised of several programs and it is selected on the basis of relatedness of sequences participating in alignments, such as MUMmer for finding maximal exact match, NUCmer for aligning closely related sequence and PROmer for aligning far related sequences. Alignment outputs are converted to dot and percent identity plots using mummerplot and gnuplot. Software is available at ftp://ftp.tigr.org/pub/software/MUMmer/.

EMBOSS

EMBOSS (Rice et al. 2000) stands for The European Molecular Biology Open Software Suite and is developed for the molecular biology community. Currently, more than 100 programs are available in EMBOSS package grouped by analysis functions such as alignment, display, edit, enzyme kinetics, nucleotide, protein analysis and phylogeny. Many more applications will be added in the near future. It is difficult to describe all the programs here but EMBOSS is mainly used for sequence alignment, restriction map, CpG island analysis, primer design, sequence extraction, sequence retrieve from database, codon usage analysis, protein motif analysis and many more applications. It runs on UNIX environment with command line mode (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/download.html) and java graphical user interface version (JEMBOSS; Carver and Bleasby, 2003) is also available (http://www.rfcgr.mrc.ac.uk/Software/EMBOSS/Jemboss/download/).

4. MOLECULAR DATABASE

Each year, Journal Nucleic Acids Research provides collective database information covering various biological research areas. This year 548 biological databases are updated in 11 hierarchical classifications, which helps users easily find the database they need (Galperin 2004).

Pfam database: Pfam (Beatman et al. 2004) is a comprehensive collection of protein domains and families represented by multiple sequence alignment and profile-HMMs of SwissProt and TrEMBL protein data. It is divided into two groups, pfam-A and pfam-B. pfam-A is a collection of protein domains from manual multiple alignment, it being the case that pfam-B is automatic collection of conserved domain. More than 7,400 families are listed in current release (June 2004). Protein, nucleotide and keyword searches are provided using web service at pfam site (http://pfam.wustl.edu/).

SCOP database: The Structural Classification of Proteins (SCOP; Andreeva et al. 2004) is a collection of classified proteins on the basis of known protein structure. Proteins are classified in alpha, beta, alpha and beta, multi-domain, membrane, cell surface or small proteins. Currently, 40,450 domains classified 2327 families, 1294 superfamilies and 800 folds in release 1.65 (http://scop.mrc-lmb.cam.ac.uk/scop/).

RefSeq: Reference Sequence (RefSeq; Pruitt and Maglott 2001) is a non-redundant collection of DNA, RNA and protein sequences for major research organisms in NCBI. Each RefSeq entry uniquely represents a stable reference for gene identification, mutation analysis, polymorphism discovery and comparative analysis. It is manually curated and updated periodically. Current release (release 6) contains about 1.3 million entries representing 2,467 taxa and is available at http://www.ncbi.nlm.nih.gov/RefSeq/.

TIGR Gene indices: TIGR Gene Indices (Quackenbush et al. 2001) are assembled EST database based on public EST sequences. EST sequences provide expression gene profiles in a genome but because of the nature of the short, single pass sequence, usage is limited. To overcome this problem, assembly strategy is applied on public ESTs. Currently 77 Gene Indices covering animals, plants and fungi are available at

http://www.tigr.org/tdb/tgi/index.shtml and EST clustering methods are described at http://www.tigr.org/tdb/tgi/software/.

MetaCyc: MetaCyc (Krieger et al. 2004) is a multi-organism metabolic pathway and enzyme database primarily in microorganisms and plants. It provides the information of metabolic pathways along with compounds, enzyme and genes. Thus, it is useful in pathway prediction in annotated genome using software Pathway Tools (Karp et al. 2002). More than 500 metabolic pathways are collected in MetaCyc representing about 4,900 enzymatic reactions and database can be browed by enzyme, pathway and gene name as a query (http://metacyc.org/).

5. CONCLUSIONS

Knowing and utilizing appropriate programs is important to many biologists these days in order to investigate their biological interests. As a starting point for this procedure, we have described a handful of bioinformatics tools for analyzing sequence data. Two important topics among many items that we did not discuss are UNIX operating system and Perl scripting. Many sequence analysis programs are written and operated in UNIX environment. Therefore, understanding UNIX gives more power to extensively utilize analysis programs as well as efficiently handle input and output files. Perl is the most popular computer programming language used in genomics. Perl script allows identifying the patterns in huge datasets or text inputs and outputs. Independently on existing programs, Perl script can add more flexible ways to organize and analyze data.

We are living in a flood of genomic information today and this may be only the beginning of a bigger wave. As many organisms are sequenced, we realize that we need more and more sequence information to explain similarities and differences within genome and between genomes. With this consequence, biology and bioinformatics work closely together to design new algorithms and programs for analyzing genome. All these collective efforts will give us profound knowledge to better understand the diversity of living organisms.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR and Venter JC (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651-1656.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). Basic local alignment search tool. J Mol Biol 215:403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG (2004). SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32 Database issue:D226-229.

Bao Z and Eddy SR (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269-1276.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C and Eddy SR (2004). The Pfam protein families database. Nucleic Acids Res 32 Database issue:D138-141.

Batzoglou S, Berger B, Mesirov J and Lander ES (1999). Sequencing a genome by walking with clone-end sequences: a mathematical analysis. Genome Res 9:1163-1174.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP and Lander ES (2002). ARACHNE: a whole-genome shotgun assembler. Genome Res 12:177-189.

Birney E, Clamp M and Durbin R (2004). GeneWise and Genomewise. Genome Res 14:988-995.

Brendel V and Kleffe J (1998). Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. Nucleic Acids Res 26:4748-4757.

Bureau TE and Wessler SR (1994). Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6:907-916.

Bureau TE and Wessler SR (1992). Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell 4:1283-1294.

Burge C and Karlin S (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78-94.

Burge CB and Karlin S (1998). Finding the genes in genomic DNA. Curr Opin Struct Biol 8:346-354.

Carver T and Bleasby A (2003). The design of Jemboss: a graphical user interface to EMBOSS. Bioinformatics 19:1837-1843.

Chou HH and Holmes MH (2001). DNA sequence quality trimming and vector removal. Bioinformatics 17:1093-1104.

Chuang TJ, Lin WC, Lee HC, Wang CW, Hsiao KL, Wang ZH, Shieh D, Lin SC and Ch'ang LY (2003). A complexity reduction algorithm for analysis and annotation of large genomic sequences. Genome Res 13:313-322.

Clamp M, Cuff J, Searle SM and Barton GJ (2004). The Jalview Java alignment editor. Bioinformatics 20:426-427.

Delcher AL, Phillippy A, Carlton J and Salzberg SL (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30:2478-2483.

Eddy SR (1998). Profile hidden Markov models. Bioinformatics 14:755-763.

Ewing B and Green P (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186-194.

Ewing B, Hillier L, Wendl MC and Green P (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM and et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496-512.

Galperin MY (2004). The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res 32 Database issue:D3-22.

Gelfand MS, Mironov AA and Pevzner PA (1996). Gene recognition via spliced sequence alignment. Proc Natl Acad Sci U S A 93:9061-9066.

Girard L and Freeling M (1999). Regulatory changes as a consequence of transposon insertion. Dev Genet 25:291-296.

Gordon D, Abajian C and Green P (1998). Consed: a graphical tool for sequence finishing. Genome Res 8:195-202.

Gordon D, Desmarais C and Green P (2001). Automated finishing with autofinish. Genome Res 11:614-625.

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P and Brunak S (1996). Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Res 24:3439-3452.

Henikoff S and Henikoff JG (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915-10919.

Higgins DG (1994). CLUSTAL V: multiple alignment of DNA and protein sequences. Methods Mol Biol 25:307-318.

Huang X, Adams MD, Zhou H and Kerlavage AR (1997). A tool for analyzing and annotating genomic sequences. Genomics 46:37-45.

Huang X and Madan A (1999). CAP3: A DNA sequence assembly program. Genome Res 9:868-877.

Huang X, Wang J, Aluru S, Yang SP and Hillier L (2003). PCAP: a whole-genome assembly program. Genome Res 13:2164-2170.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC and Lander ES (2003). Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res 13:91-96.

Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR and Wessler SR (2003). An active DNA transposon family in rice. Nature 421:163-167.

Juretic N, Bureau TE and Bruskiewich RM (2004). Transposable element annotation of the rice genome. Bioinformatics 20:155-160.

Karp PD, Paley S and Romero P (2002). The Pathway Tools software. Bioinformatics 18 Suppl 1:S225-232.

Kent WJ (2002). BLAT--the BLAST-like alignment tool. Genome Res 12:656-664.

Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY and Karp PD (2004). MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 32 Database issue:D438-442.

Krogh A (1997). Two methods for improving performance of an HMM and their application for gene finding. Proc Int Conf Intell Syst Mol Biol 5:179-186.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C and Salzberg SL (2004). Versatile and open software for comparing large genomes. Genome Biol 5:R12.

Loots GG and Ovcharenko I (2004). rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res 32:W217-221.

Lowe TM and Eddy SR (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955-964.

Mahairas GG, Wallace JC, Smith K, Swartzell S, Holzman T, Keller A, Shaker R, Furlong J, Young J, Zhao S, Adams MD and Hood L (1999). Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. Proc Natl Acad Sci U S A 96:9739-9744.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S and Wingender E (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374-378.

Mironov AA, Roytberg MA, Pevzner PA and Gelfand MS (1998). Performance-guarantee gene predictions via spliced alignment. Genomics 51:332-339.

Mullikin JC and Ning Z (2003). The phusion assembler. Genome Res 13:81-90.

Olson M, Hood L, Cantor C and Botstein D (1989). A common language for physical mapping of the human genome. Science 245:1434-1435.

Ouellette BF and Boguski MS (1997). Database divisions and homology search files: a guide for the perplexed. Genome Res 7:952-955.

Ovcharenko I, Loots GG, Hardison RC, Miller W and Stubbs L (2004). zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. Genome Res 14:472-477.

Pearson WR and Lipman DJ (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85:2444-2448.

Pertea M, Lin X and Salzberg SL (2001). GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res 29:1185-1190.

Pop M and Kosack D (2004). Using the TIGR assembler in shotgun sequencing projects. Methods Mol Biol 255:279-294.

Pruitt KD and Maglott DR (2001). RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137-140.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 29:159-164.

Rice P, Longden I and Bleasby A (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277.

Salamov AA and Solovyev VV (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Res 10:516-522.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R and Miller W (2000). PipMaker--a web server for aligning two genomic DNA sequences. Genome Res 10:577-586.

Siegel AF, Trask B, Roach JC, Mahairas GG, Hood L and van den Engh G (1999). Analysis of sequence-tagged-connector strategies for DNA sequencing. Genome Res 9:297-307.

Solovyev VV, Salamov AA and Lawrence CB (1995). Identification of human gene structure using linear discriminant functions and dynamic programming. Proc Int Conf Intell Syst Mol Biol 3:367-375.

Taher L, Rinner O, Garg S, Sczyrba A, Brudno M, Batzoglou S and Morgenstern B (2003). AGenDA: homology-based gene prediction. Bioinformatics 19:1575-1577.

Tatusova TA and Madden TL (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174:247-250.

Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO and Hunkapiller M (1998). Shotgun sequencing of the human genome. Science 280:1540-1542.

Venter JC, Smith HO and Hood L (1996). A new strategy for genome sequencing. Nature 381:364-366.

Wessler SR (2001). Plant transposable elements. A hard act to follow. Plant Physiol 125:149-151.

Wessler SR, Bureau TE and White SE (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev 5:814-821.

Xu Y, Mural RJ and Uberbacher EC (1994). Constructing gene models from accurately predicted exons: an application of dynamic programming. Comput Appl Biosci 10:613-623.

Yeh RF, Lim LP and Burge CB (2001). Computational inference of homologous gene structures in the human genome. Genome Res 11:803-816.

Zhang MQ (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci U S A 94:565-568.

Zhang Z, Schwartz S, Wagner L and Miller W (2000). A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203-214.

Table 1. Sequence submission divisions in GenBank.

Division Abbreviation	Data source
PRI	Primate sequences
ROD	Rodent sequences
MAM	Other mammalian sequences
VRT	Other vertebrate sequences
INV	Invertebrate sequences
PLN	Plant, fungal and algal sequences
BCT	Bacterial sequences
VRL	Viral sequences
PHG	Bacteriophage sequences
SYN	Synthetic sequences
UNA	Unannotated sequences
EST	Expressed sequence tags
PAT	Patent sequences
STS	Sequence tagged sites
GSS	Genome survey sequences
HTG	High-throughput genome sequences
HTC	Unfinished high-throughput cDNA sequences

Table 2. List of GenBank databases for BLAST search.

	Database	Description
Nucleotide Search	Nr	Non-redundant nucleotide data of GenBank, [1]EMBL and [2]DDBJ not including EST, STS, GSS and HTGS database
	Month	Newly releases nucleotide sequences (within 30 days) in GenBank, EMBL and DDBJ
	dbEST	Non-redundant EST sequences in GenBank, EMBL and DDBJ
	dbSTS	Non-redundant STS sequences in GenBank, EMBL and DDBJ
	Mouse EST	Non-redundant mouse ESTs in GenBank, EMBL and DDBJ
	Human EST	Non-redundant human ESTs in GenBank, EMBL and DDBJ
	Other EST	Non-redundant ESTs in GenBank, EMBL and DDBJ excluding mouse and human ESTs
	Yeast	*Saccharomyces cerevisiae* genomic sequences
	E. coli	*Escherichia coli* genomic nucleotide sequences
	Vector	Vector sequence database
	Mito	Mitochondria sequence database
	Alu	Alu repeat sequence database
	GSS	Single pass genome survey sequences including BAC end sequences
	HTGS	Phase1 and 2 high throughput genomic sequence database
Protein Search	Nr	All non-redundant GenBank [3]CDS translations+[4]PDB+SwissProt+[5]PIR+[6]PRF
	Month	Newly released (in 30 days) or revised GenBank CDS translation, SwissProt and PIR
	SwissProt	The last major release of the SWISS-PROT protein sequence database
	Yeast	*Saccharomyces cerevisiae* protein sequences derived from yeast genome sequence
	E. coli	*Escherichia coli* translated coding sequences
	PDB	Protein sequences derived from 3D structure data

[1]EMBL: European Molecular Biology Laboratory, [2]DDBJ: DNA Data Bank of Japan, [3]CDS: Coding Sequence, [4]PDB: Protein Data Bank, [5]PIR: Protein Information Resource, [6]PRF: Protein Research Foundation.

Table 3. List of gene prediction programs. Ab initio program: FGENESH, Genscan, Grail, MZEF and HMMgene. Similarity based program: CRASA, AAT and AgenDa. Combined methods: GenomeScan, Procrustes and FGENESH+.

Program	URL and training set
FGENESH	http://www.softberry.com/berry.phtml; Human, Mouse, Fruit fly, Monocot, Dicot, *S. pombe*, *Neurospora*, Fish, Algae, *Aspergillus*
Genscan/ Genscan+	http://genes.mit.edu/GENSCAN.html Vertebrates, *Arabidopsis*, Maize
Grail	http://compbio.ornl.gov/Grail-1.3/ Human, Mouse, *Arabidopsis*, *Drosophila*, *E. coli*
MZEF	http://rulai.cshl.org/tools/genefinder/ Human, Mouse, *Arabidopsis*, Yeast
HMMGene	http://www.cbs.dtu.dk/services/HMMgene/ Human and other vertebrates, *C. elegans*
CRASA	http://crasa.sinica.edu.tw/bioinformatics/bioinformatics.html
AAT	http://genome.cs.mtu.edu/aat/aat.html
AGenDa	http://bibiserv.techfak.uni-bielefeld.de/agenda/
GenomeScan	http://genes.mit.edu/genomescan.html Vertebrates, *Arabidopsis*, Maize
Procrustes	http://www-hto.usc.edu/software/procrustes/wwwserv.html
FGENESH+	http://www.softberry.com/berry.phtml

Figure legend

Figure 1. An example of Entrez search.

Figure 2. Simplified procedures involved in clone by clone shotgun (CBC) and whole genome shotgun (WGS) methods.

Figure 3. Consed is a program for viewing and editing phrap assembly. A: Snapshots of consed assembly view, B: Aligned reads window, C: Trace window.

Figure 1. An example of Entrez search.

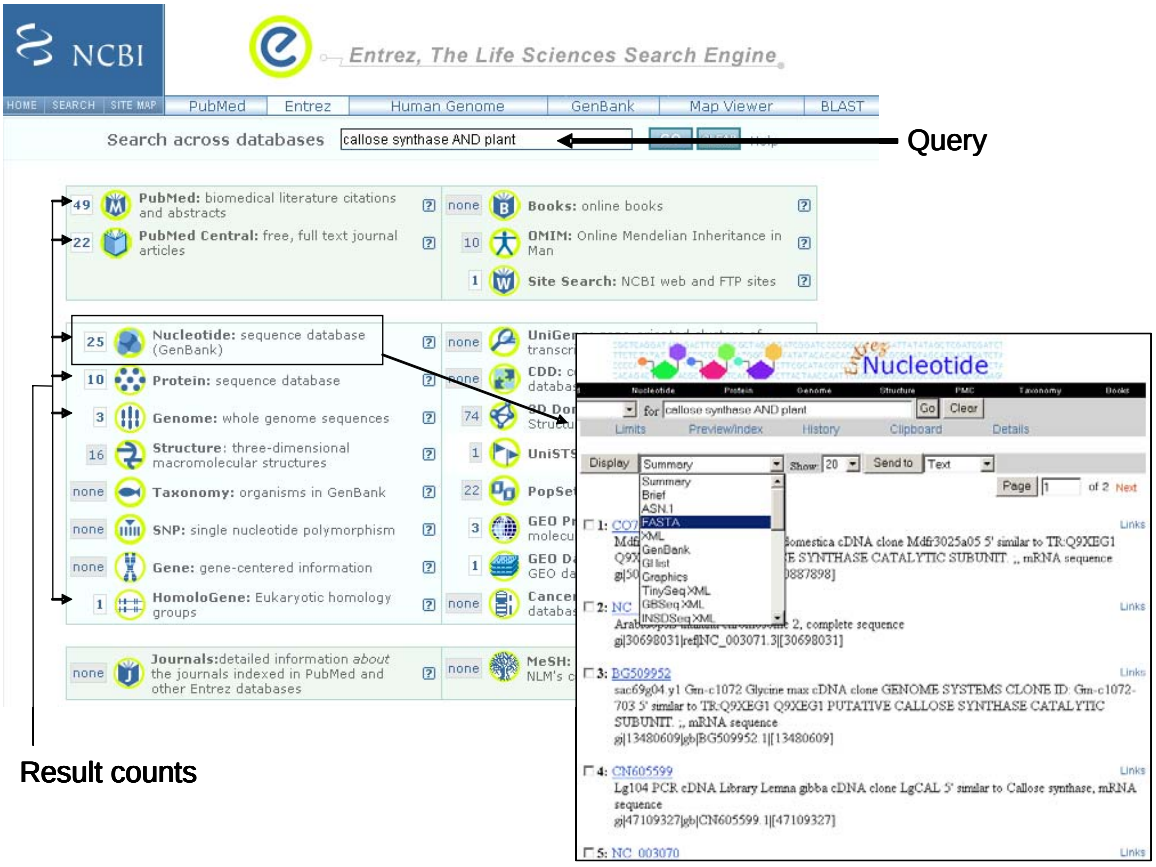

Figure 2. Simplified procedures involved in clone by clone shotgun (CBC) and whole genome shotgun (WGS) methods.

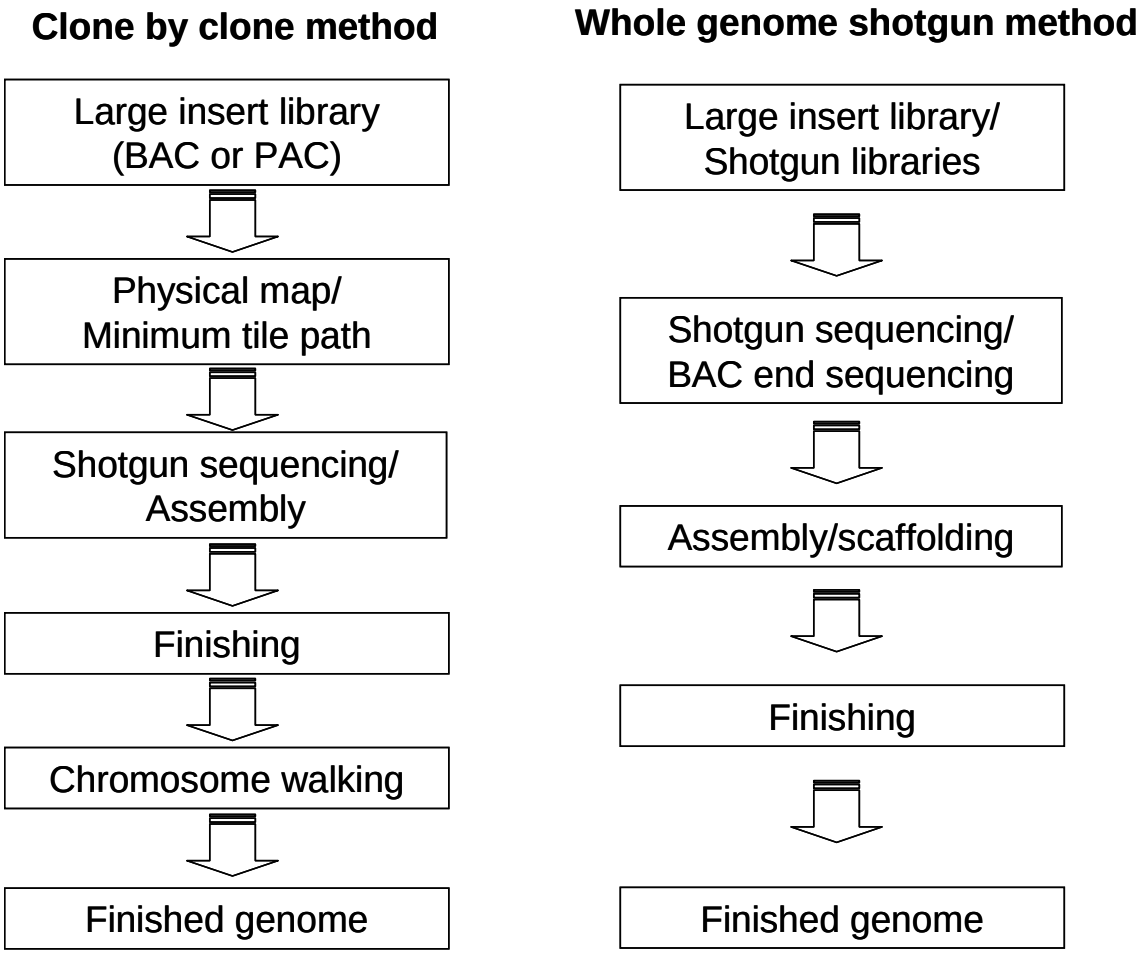

Figure 3. Consed is a program for viewing and editing phrap assembly. A: Snapshots of consed assembly view, B: Aligned reads window, C: Trace window.