國立交通大學

生物資訊及系統生物研究所

碩士論文

建立基因體計畫的計算流程:序列重組、基因註解與重建代謝路徑

Establishing a Computational Pipeline of Genome Projects: Sequence Assembly, Gene Annotation and Metabolic Pathway Reconstruction

研 究 生：黃至昶

指導教授：黃憲達　教授

中 華 民 國 九 十 九 年 七 月

建立基因體計畫的計算流程:序列重組、基因註解與重建代謝路徑

Establishing a Computational Pipeline of Genome Projects: Sequence Assembly, Gene Annotation and Metabolic Pathway Reconstruction

研 究 生:黃至昶 Student:Chih-Chang Huang

指導教授:黃憲達 Advisor:Hsien-Da Huang

國 立 交 通 大 學

生物資訊及系統生物研究所

碩 士 論 文

A Thesis

Submitted to Institute of Bioinformatics and Systems Biology

College of Biological Science and Technology

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

In

Bioinformatics and Systems Biology

July 2010

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 九 年 七 月

建立基因體計畫的計算流程:序列重組、基因註解與重建代謝路徑

學生:黃至昶 指導教授:黃憲達

國立交通大學生物資訊及系統生物研究所碩士班

摘要

西元2003年人類計畫完成,為爾後的生物研究帶來龐大的資源。近幾年 next-generation sequencing 技術的發展大量的降低定序成本及時間使得物種的定序更佳的容易,因此各物種的基因體定序也開始蓬勃發展，而且在這些基因體中也蘊藏了大量的研究資源，這些基因體計畫的分析及註解也將是更加急迫的需要，因而需要一個有系統的計算流程。這個流程針對不同的定序技術產生的序列會使用不同的序列重組工具，整合 ab initio 及 evidence-based 這兩種基因預測的方式來更準確的預測基因,此外還會根據基因註解的資訊來重建物種的代謝路徑。

這個基因體計畫計算流程可以重組各類不同定序技術產生的序列以及提供基因體註解的服務有:基因註解以及重建代謝路徑。這個流程將會在高產量的基因體註解中佔有一席之地。

Establishing a Computational Pipeline of Genome Projects: Sequence Assembly, Gene Annotation and Metabolic Pathway Reconstruction

Student：Chih-Chang Huang Advisor：Hsien-Da Huang

Institute of Bioinformatics and Systems Biology, National Chiao Tung University

Abstract

Human Genome Project had been completed in 2003. It provides gigantic resources for biological research. In recent years, next generation sequencing technique dramatically reduces the sequencing cost and time. Thus, completely sequencing new organisms will be popular and universal, and the genomes of these organisms also include huge research resources. The demands of comprehensive genomic annotation will be more urgent and necessary. Thus, it is necessary a computational pipeline. In order to assembly complete genome sequences, this pipeline uses several assembly tools which designed for assembling traditional sequencing and next generate sequencing raw data. It also integrates ab initio and evidence-based gene prediction approaches to predict genes. In

addition, this pipeline can reconstruct metabolic pathways from the gene annotation results. This computational pipeline can assemble sequencing data from various platforms and provide the service of genomic annotation including: gene annotation and metabolic pathway reconstruction. This computational pipeline can be a crucial part of pipeline in the high throughput genomic annotation.

致　謝

首先，我要感謝指導教授黃憲達教授在這兩年期間的細心指導，讓我對生物資訊這個領域從一知半解到更加的了解、精進，在您身上我學到不只是做研究的態度及方法更學到了處事的態度是我在未來方向的一個榜樣。

在這兩年期間感謝實驗室學長們對我在研究上面的指點及包容，同時也感謝在每位這篇論文研究的小組成員，沒有你們就不會有這篇論文的產生，與你們每一次的討論都讓我有更多的想法來推進研究的進度，實驗室的同仁們，感謝你們兩年來的照顧，這兩年期間的回憶將永遠記在我心中。

最後感謝家人對我的支持，在這段時間我幾乎很少回家，或者是很短暫的時間，感謝你們對我的支持、體諒以及經由電話的噓寒問暖，讓我不在家中也能感到家裡的溫暖。

我之所以能夠完成論文順利拿到碩士學位都是在身邊每個人的幫助，無論是研究上還是生活上都由衷的感謝你們，在這兩年雖然有辛苦但也過得很快樂，我很高興能來到 ISBLAB 我將不會忘記你們大家，謝謝你們。

國立交通大學 生物資訊及系統生物研究所

整合系統生物學實驗室 研究生 黃至昶

謹誌於交通大學　2010 年七月

Content

List of Figures

List of Tables

Chapter 1 Introduction

The techniques of next generation sequencing (high throughput sequencing) have made the acquisition of genomic sequences more affordable and easier. These techniques have enabled the large scale investigations of novel genomes especially for commercial, medical, and model organisms. Hence, a well integrated software platforms of genome annotation from next generation deep sequencing data is emerged, and this platform can reduce the time of learning individual packages and provide a sketch of genome in a very short time. Based on the sketch of genome annotation, we can quickly compare with well annotated genomes and excludes the well annotated genomic sequence. After eliminate the well annotate homologous genomic regions, the remaining parts might exists abundant genes with novel functions. In other words, we can organize the advanced researches in advance for the unexplored genomic region. Nevertheless, the prototype of novel genome also provides the basic biological understanding because our platform not only provides the function of assembly deep sequencing data but also includes the gene prediction and metabolic pathway reconstruction. In short, we construct a platform for quickly sketching of a novel genome including sequence assembly, gene prediction and metabolic pathway reconstruction, based on deep sequencing data.

1.1 Background

1.1.1 Next-generation sequencing

Next-generation sequencing is a high throughput sequencing technology. Next-generation sequencing platforms include the Genome Sequencer from Roche 454 Life Sciences (www.454.com), the Solexa Genome Analyzer from Illumina (www.illumina.com), the SOLiD System from Applied Biosystems (www.appliedbiosystems.com). Those platforms are characterized by highly parallel operation, higher yield, simpler operation, much lower cost per read, and shorter reads[1]. Next-generation sequencing produces large amounts (typically millions) of short DNA sequence reads of length between 25bp and 400bp. These reads are shorter than the traditional Sanger sequence reads of length between 500bp and 1000bp. The summary of sequencing technology show on Table 1.1

Table 1.1 The comparison of different next generation sequencing platforms.

Platform	Read length	Mb per run	Time per run	Mb per day	Cost per Mb
Sanger	1000 bp	-	-	~2 Mb	~$500
Roche 454	250 bp	100 Mb	7 hr	~350 Mb	~$60
Illumina	32~40 bp	1300 Mb	3 days	~400 Mb	~$2
SOLiD	35 bp	4000 Mb	7 days	~500 Mb	~$2

1.1.1.1 Roche 454

The first next-generation platform was GS20 developed by Roche 454 Life Sciences using pyrosequencing technology. Genomic DNA splice to smaller fragment, ligate adaports into the ends of fragment and amplified by emulsion PCR. After amplification, the DNA bound beads are placed into picotiter-plate wells with sequencing enzymes such as DNA polymerase, ATP sulfurylase, and luciferase. During the sequencing, the four DNA nucleotides are added into the well. When a nucleotide which added into the well complement to the template strand. That nucleotide will generate a light signal that detected and recorded by CCD (charge-couple device) (Figure 1.1). The performance of GS20 was over 20 million base pairs in over 4hour. The GS20 was replaced during 2007 by the GS FLX model, capable of producing over 100 million base pairs of sequence in a similar amount of time. There are other alternative sequecing platform which are Solexa Genome Analyser technology and the SOLiD[2].

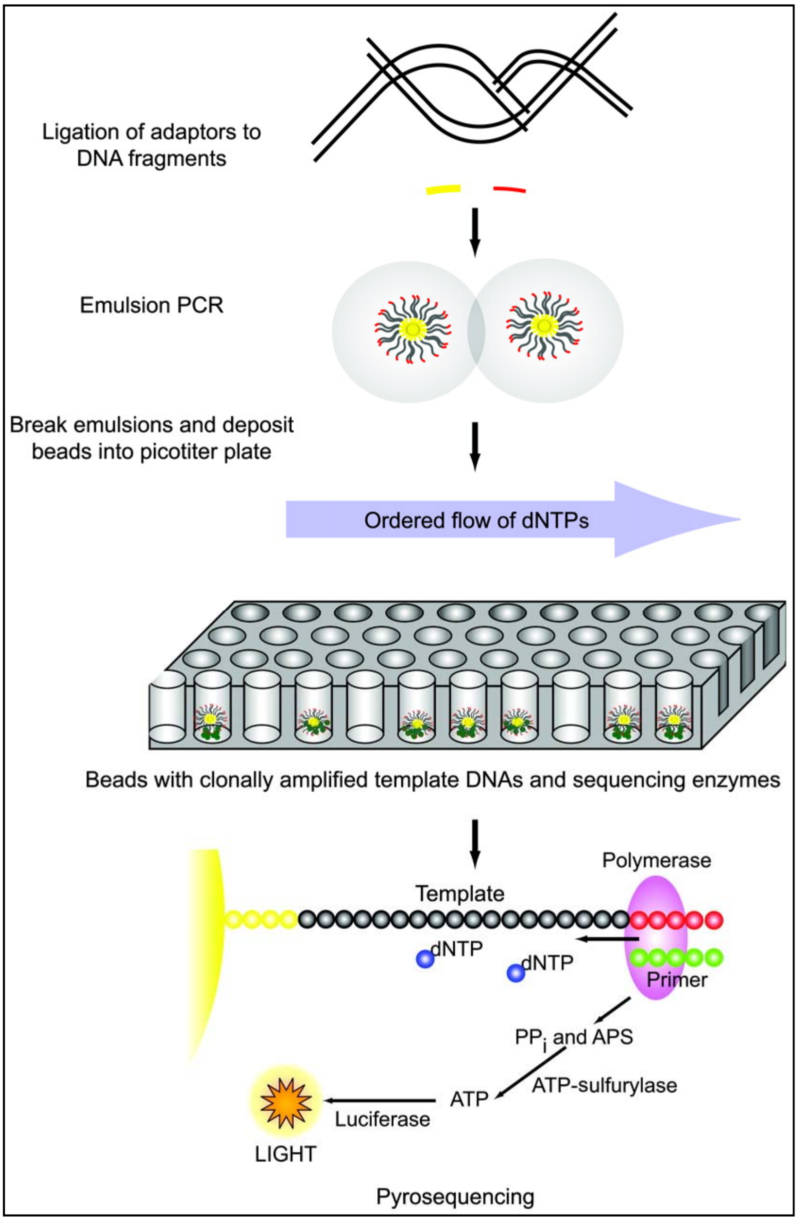


Figure 1.1 Roche 454 GS FLX sequencing[1][3]

---

[1]The picture is copy from Next-generation sequencing: from basic research to diagnostics Figure 1 Roche 454 GS FLX sequencing [3]

1.1.1.2 Illumina Solexa

The Solexa Genome Analyzer system was developed by Solexa using reversible terminator chemistry technology (Figure 1.2) and now owned by Illumina. This is the first of the massively parallel short-read platforms. Sequencing templates are immobilised on a flow cell surface, and solid phase amplification creates clusters of identical copies of each DNA molecule. Sequencing then uses four proprietary fluorescently labelled nucleotides to sequence the millions of clusters on the flow cell surface. These nucleotides possess a reversible termination property, allowing each cycle of the sequencing reaction to occur simultaneously in the presence of the four nucleotides [2].

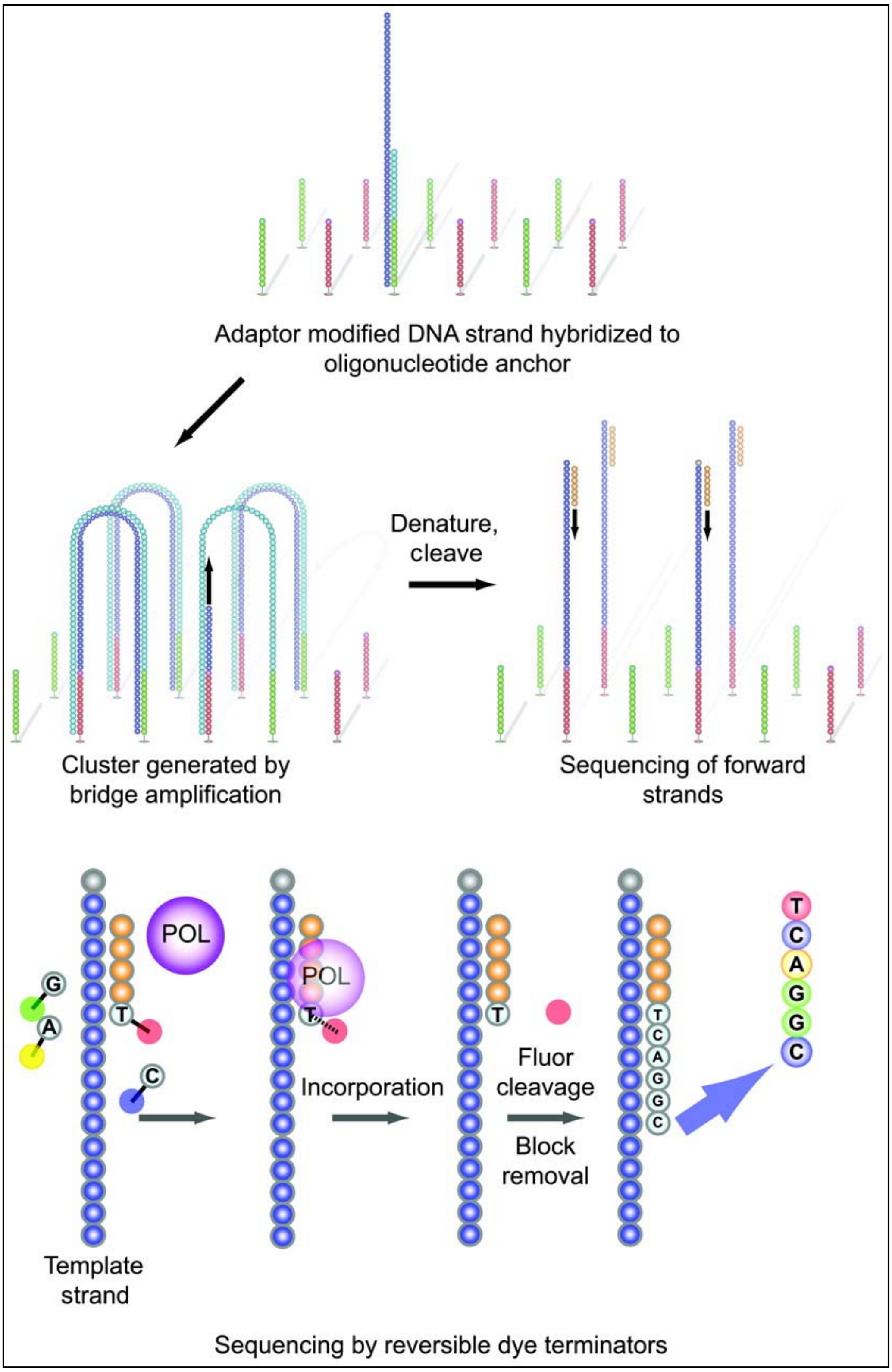


Figure 1.2 Illumina Genome Analyzer sequencing[2]

---

[2] The picture is copy from Next-generation sequencing: from basic research to diagnostics Figure 2 Illumina Genome Analyzer sequencing [3]

1.1.1.3 SOLiD

The SOLiD (Supported Oligonucleotide Ligation and Detection) System was developed by Applied Biosystems. Certain elements of the platform are directly analogous to features of both the 454 and Illumina systems. Template amplification is by emulsion PCR as the 454 platform and template is applied at high density to a flow cell as Illumina[4]. The feature of the SOLiD platform is the ligation-based sequencing chemistry (Figure 1.3).

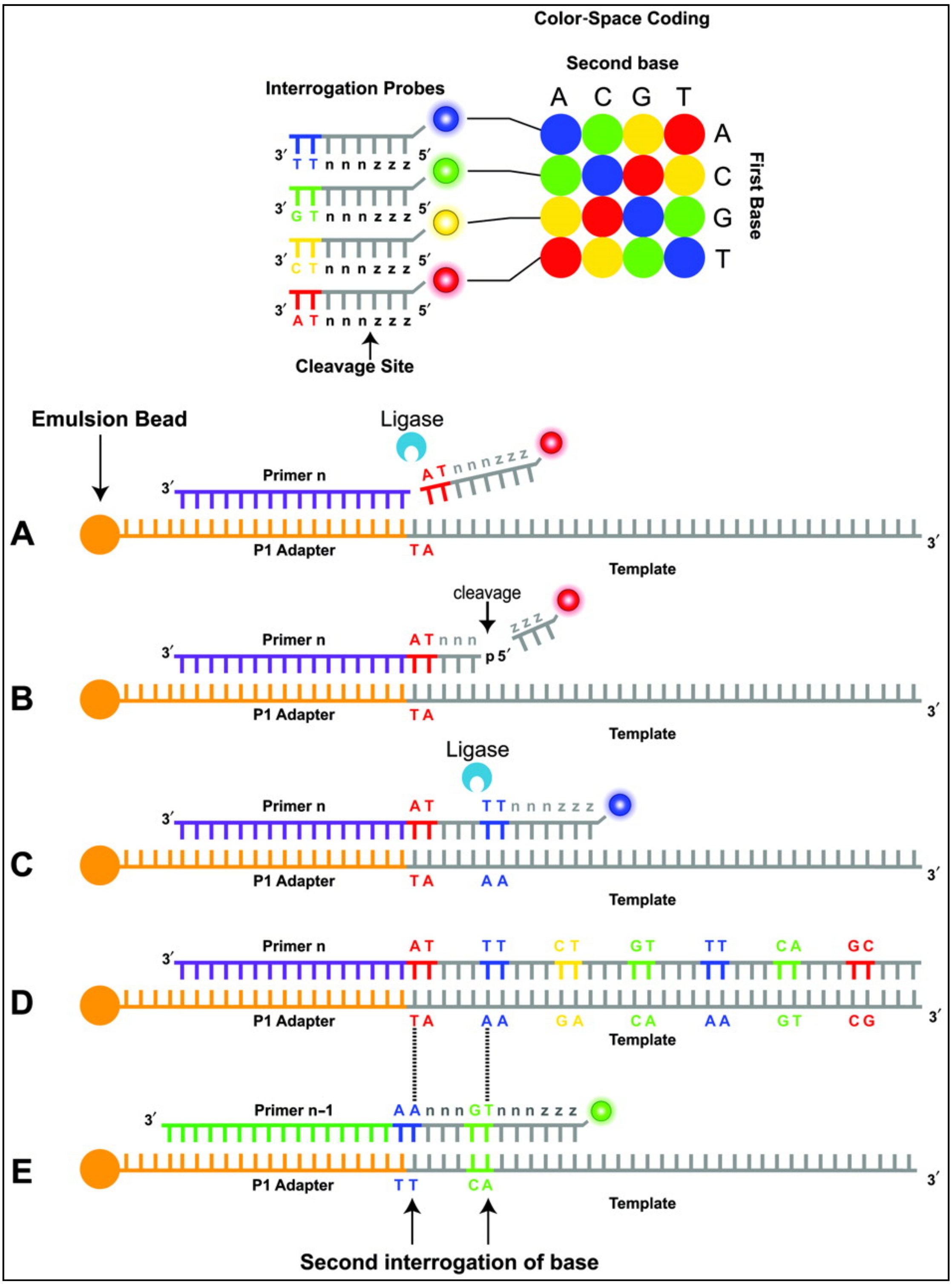


Figure 1.3 Applied Biosystems SOLiD sequencing by ligation[3]

---

[3] The picture is copy from Next-generation sequencing: from basic research to diagnostics Figure 3 Applied Biosystems SOLiD sequencing by ligation [3]

The next-generation sequencing platforms have characteristic error profiles. Error profiles can include enrichment of base call error toward the 3' ends of reads, compositional bias for or against high-GC sequence, and inaccurate determination of simple sequence repeats[1]. 454 error rate is approximate 0.1%, Illumina and SoLiD are approximate 1%.

1.1.2 Sequence Assembly

Sequencing assembly is a process to group reads into contigs and contigs into scaffolds. Reads are sequence fragments sequenced by sequencing platforms. Contigs (contiguous sequence) are a set of overlapping reads that represent a continuous region of DNA sequence. The scaffolds are the contigs order and orientation and the sizes of the gaps between contigs (Figure 1.4).

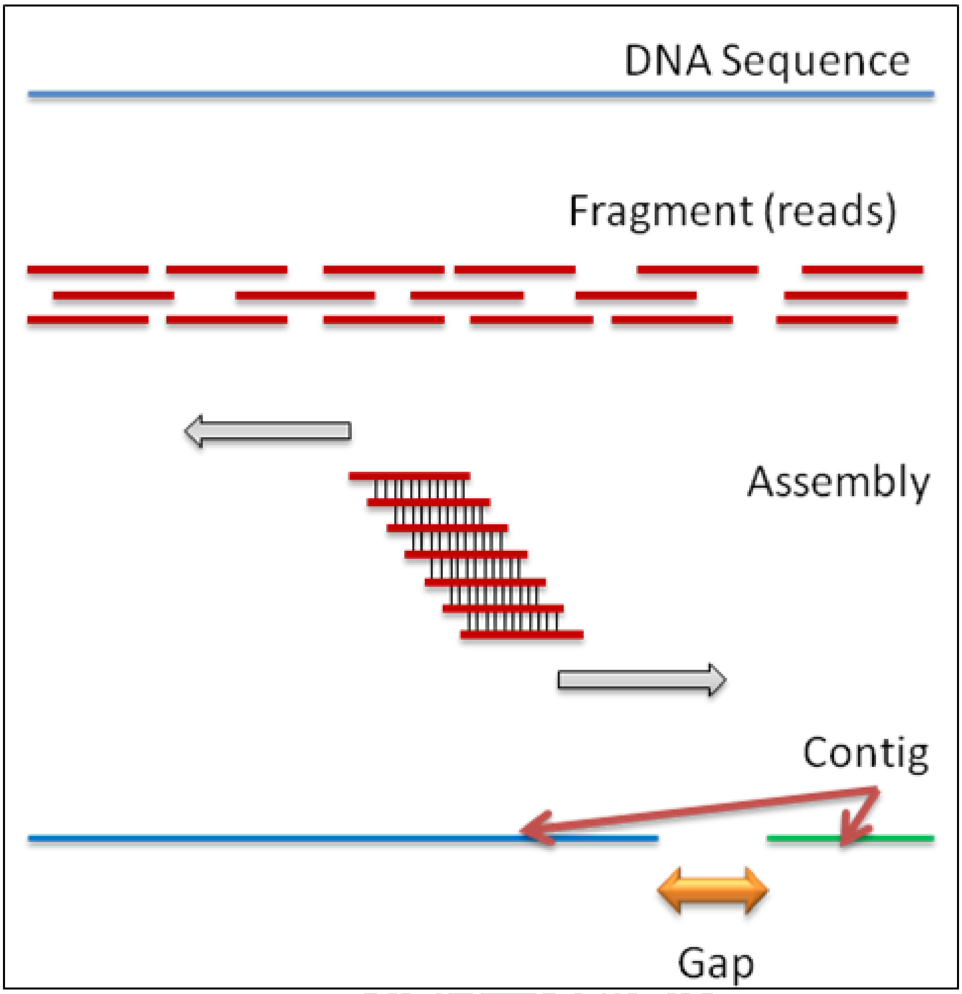


Figure 1.4 Sequence assembly flow

The goal of whole-genome shotgun assembly is to represent each genomic sequence in one scaffold. However, this is not always possible. One chromosome may be represented by many scaffolds or a single scaffold.

A challenge of assembly is that solve repeat region problem (Figure 1.5). In assembly process compute the overlap region between reads, the similar repeat region may confuse the order of nearby regions. For example in Figure 1.5, the repeat region X cause region A to D has A,C,B to D and A,B,C to D two ambiguous assembly sequence.

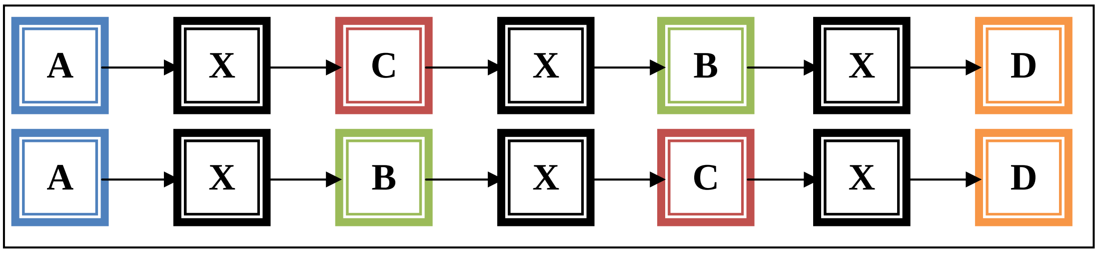


Figure 1.5 Repeat region problem

1.1.2.1 Assembly of next-generation sequencing

The challenges in assembly of next-generation sequencing are more difficult than traditional sequencing in shorter reads, high coverage data and error rate. The length of shorter reads may shorter than repeat region and many reads in repeats will have only one or no different bases. That will cause more ambiguous overlap in assembly. High coverage data compute the overlap between reads is more complex. Sequencing error will cause worse assembly accuracy and next-generation sequencing has higher error rate than traditional sequencing. Traditional assembly algorithm cannot assemble next-generation sequencing reads well. Thus, there are new assembly algorithm is developed in recent years.

1.1.2.2 Approaches of assembly

There are two basic approaches for sequence assembly: overlap-layout-consensus and de Bruijn graph.

1.1.2.2.1 Overlap-layout-consensus approach

This approach computes pair-wise overlap between all reads and reflect to overlap graphs. Reads represent nodes and overlaps represent edges.

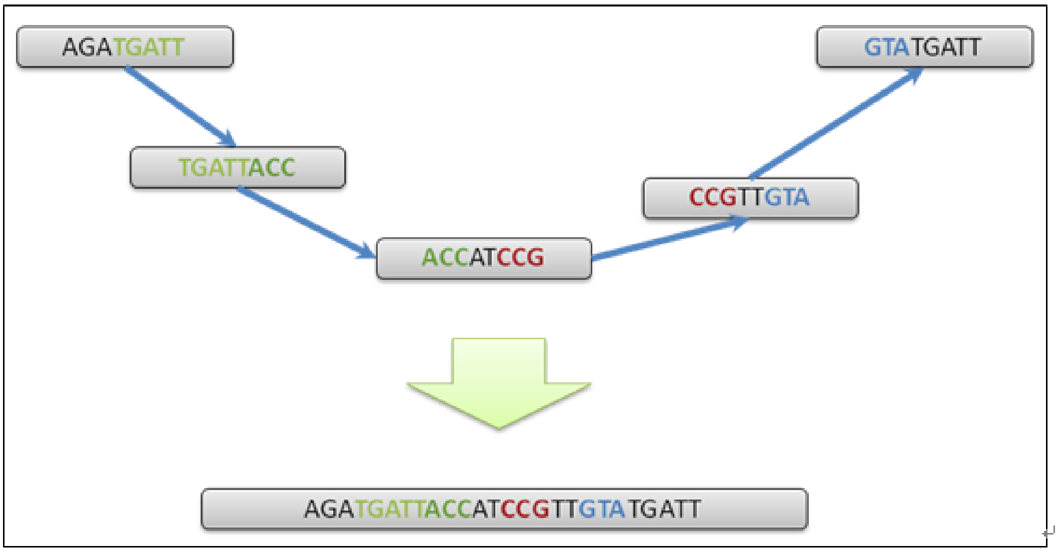


Figure 1.6 Overlap graph

Overlap-layout-consensus approach was used in Sanger sequencing assembly. The assemblers used this approach include: CAP3[5], PCAP[6], ARACHNE[7], *phrap*[8], Celera[9] and etc. Roche 454 assembly may also use this approach. The Newbler[10] is developed using overlap-layout-consensus approach by 454 life sciences.

1.1.2.2.2 De Bruijn graph

Because overlap-layout-consensus approach computes the short and high coverage next-generation sequencing data increase most computational time, most assembler for next-generation sequencing use de Bruijn graph. De Bruijn graph reduce computational complexity by splice reads to fragments with k length as k-mer and have k-1 length overlap between k-mers.

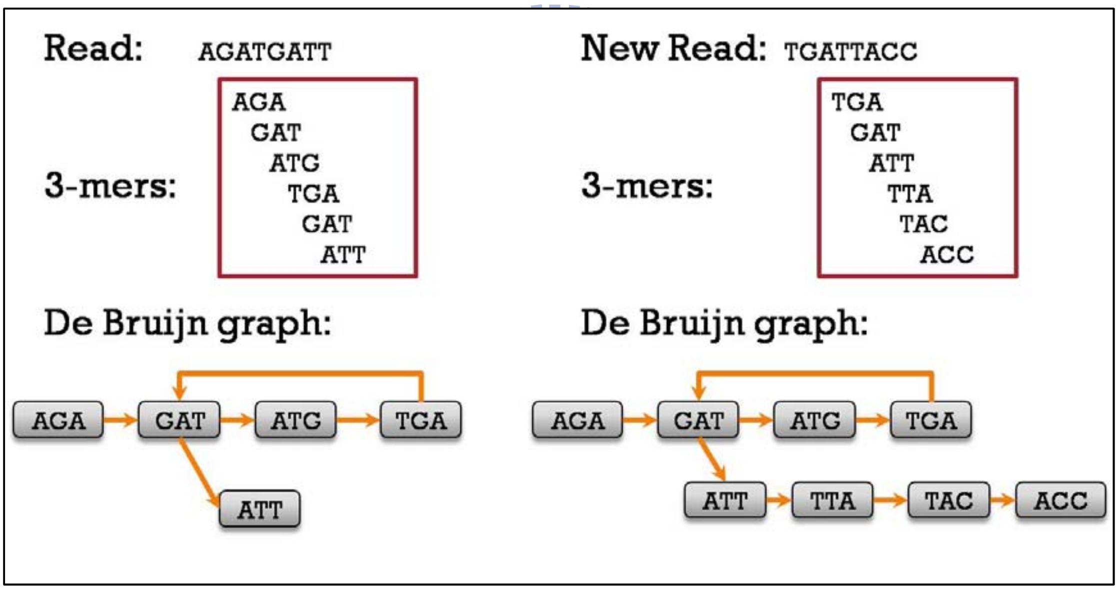


Figure 1.7 De Bruijn Graph for read with k-mer = 3

The assemblers used this approach include: Velvet[11], Abyss[12], ALLPaths[13], SOAPdenovo[14] and etc.

1.1.3 Gene annotations

Gene annotation is the process to identify gene location on the genome sequence and biological information of these genes.

1.1.3.1 Gene prediction approaches

1.1.3.1.2 *Ab initio* approach

Ab initio approach is prediction gene depend on the signal of protein-coding gene (start codon, stop codon, donor site, acceptor site, promoters and poly-A tail) and properties of protein-coding gene (exon, intron, intergenic region and UTRs). These features of gene show in Figure 1.8.The most *ab initio* gene prediction software is designed by hidden Markov model (HMM). Some example of *ab initio* gene prediction software include: AUGUSTUS[15], Fgenesh[16], GENSCAN[17], GeneMark.hmm[18], GlimmerHMM[19], and etc.

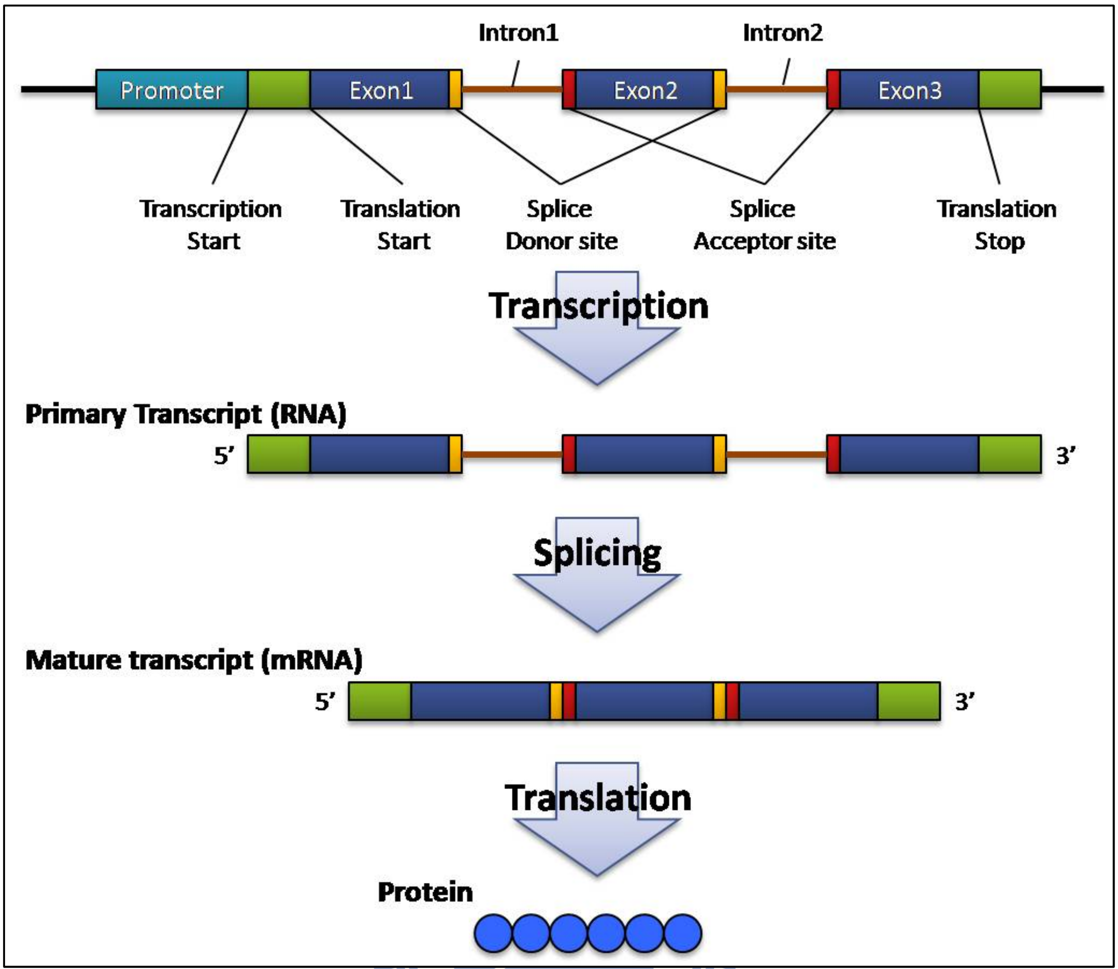


Figure 1.8 Gene structure

1.1.3.1.1 Evidence-based approach

The evidence-based is search gene on target sequence form known sequence of an mRNA, EST or protein product. BLAST[20] is a widely used software designed for this approach.

1.1.4 Metabolic pathway

Metabolic pathways are series of chemical reactions occurring in a cell. The molecules called substrates that are at the beginning of the reaction, and the enzyme changes these into different molecules as products. Those products may as substrates enter new reactions, and a series of reactions construct the metabolic pathways. Pathways are important to maintain the homeostasis of an organism. Some important metabolic pathways are: glycolysis, anaerobic respiration, citric acid cycle (krebs cycle), oxidative phosphorylation, pentose phosphate pathway, fatty acid oxidation, urea cycle.

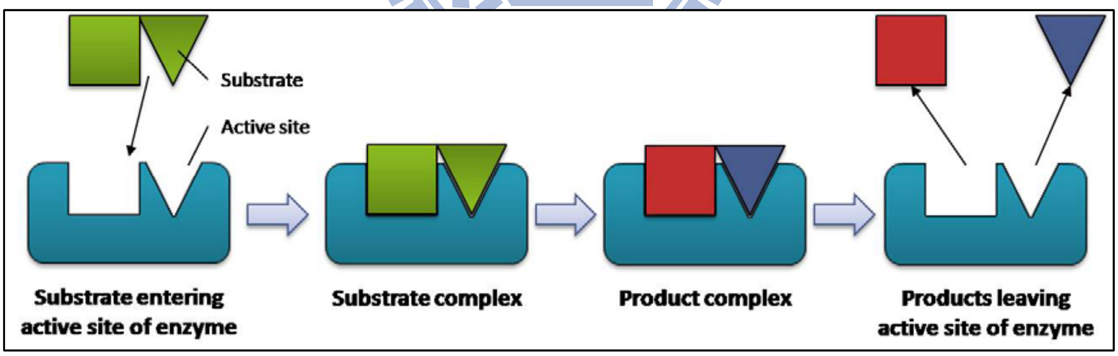


Figure 1.9 Enzyme reaction

Pathway is a method to understanding the role of gene in their larger biological context as effects of mutations, drug interventions and changes in gene regulation[21] .

1.2 Motivation

Human genomic project was initiated in 1990 and was completed in 2003. Around three billion USD were devoted into this project. Nowadays, the next-generation sequencing can sequence large genomic sequence with very low costs. Next-generation sequencing techniques provide an alternative method of massively study the genomes of novel organism and initiate the studies of understanding the genome variants of different individuals. These techniques have enabled the large scale investigations of novel genomes especially for commercial, medical, and model organisms.

Consequently, the sequence assembly and genome annotation are required for investigation the results of next-generation sequencing data. Therefore, a genome annotation platform for next-generation sequencing might reduce the time of annotating the novel genome and provide a quick profiling of the genomes.

1.3 The Specific Aim

For the explosive genome projects, a computational pipeline which includes sequence assembly, gene annotation and metabolic pathway reconstruction is necessary. This pipeline can assemble various sequencing platform and different platform combined data. It integrates several ab initio and evidence-based approach prediction tools to provide more accuracy prediction, and identify function of these genes. Finally, this pipeline reconstructs the metabolic pathway of the sequencing organisms.

Chapter 2 Related Works

2.1 Sequence assembly tools

Table 2.1 lists several sequence assembly tools: phrap[8] and CAP3[5] developed to assemble traditional sequencing data, and phrap supported Roche 454 sequencing data in later version. SSAKE and Velvet developed to assemble next generation sequencing data. SSAKE was one of first short reads assembly tools, but it did not support Roche 454 data. Velvet used a difference form overlapping assembly approach to implement as de bruijn graph, and it is now widely used short reads assembly tool.

Table 2.1 The comparison of assembly tools

Assembly tools	Approach	Supporting sequencing technology	Supporting paired-end	Reference
Phrap	Overlapping	Sanger, 454	No	[8]
CAP3	Overlapping	Sanger	No	[5]
SSAKE	Overlapping	Illumina	Yes	[22]
Velvet	De bruijn graph	454, Illumina	Yes	[11]

2.1.1 Phrap

Phrap was developed by Prof. Phil Green to provide rapid comparison, alignment, and assembly of large sets of DNA sequences. Phrap does pairwise alignment and search to a sequence region that match a designated length. Phrap extended the alignment form the match region with follow score. Matching residues receive a reward of +1, mismatches get a penalty of −2, gap opening residues a penalty of −4 and gap extension residues a penalty of −3. In this way PHRAP aligns the data into contigs[8].

2.1.2 CAP3

CAP3 is an assembly tool for traditional sequencing data. The assembly algorithm consists of three major phases (Figure 2.1). In the first phase, 5' and 3' poor regions of each read are identified and removed. Overlaps between reads are computed. False overlaps are identified and removed. In the second phase, reads are joined to form contigs in decreasing order of overlap scores. Then, forward–reverse constraints are used to make corrections to contigs. In the third phase, a multiple sequence alignment of reads is constructed and a consensus sequence along with a quality value for each base is computed for each contig. Base quality values are used in computation of overlaps and construction of multiple sequence alignments[5].

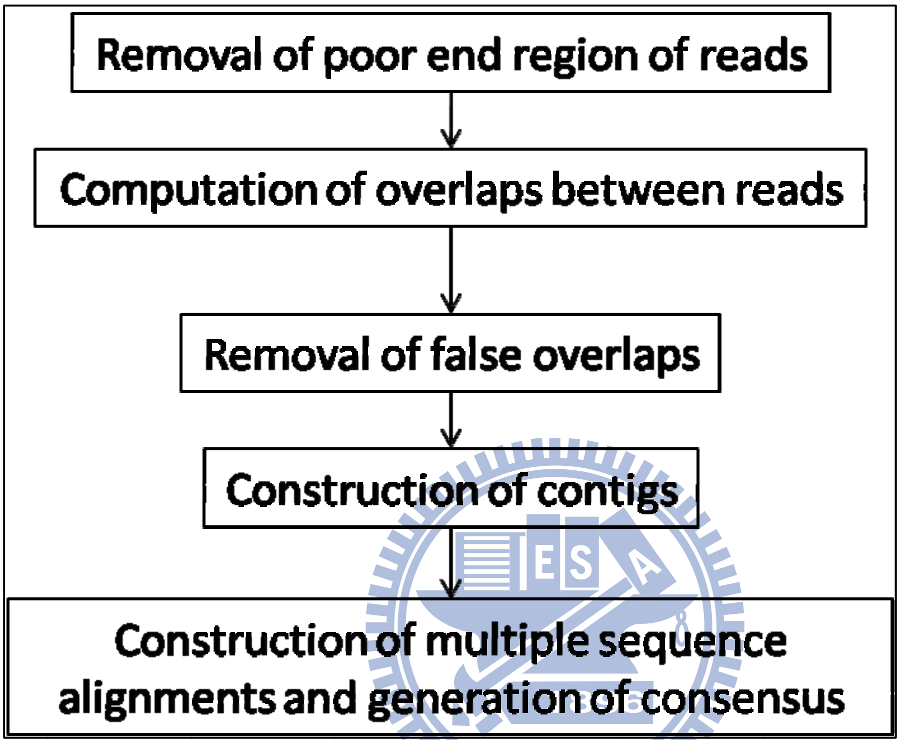


Figure 2.1 Major phases of CAP3 algorithm

2.1.3 SSAKE

SSAKE was one of first short reads assembly tools. SSAKE was for aggressively assembling millions of short nucleotide sequences by progressively searching through a prefix tree for the longest possible overlap between any two sequences. SSAKE is designed to help leverage the information from short sequence reads by stringently assembling them into contiguous sequences that can be used to characterize novel sequencing targets[22].

2.1.4 Velvet

Velvet was one of popular short read de novo assemblers. It was designed based on de bruijn graph and efficiently to both eliminate errors and resolve repeats. The first step is construction de bruijn graph with k-mer (Figure 2.2). After the initial graph constructed, simplifying it as possible without loss of information. Simplification iteratively chains of blocks are collapsed into single blocks and reduce the complexity of initial graph (Figure 2.3). There were "tip" and "bubble" error in simplified graph. Tip error was a chain of nodes that is disconnected on one end. Velvet iteratively removes tips from the graph under these two criteria: length and minority count. A tip will be removed if it is shorter than 2000. "minority count" be defined as starting from that node, going through the tip is an alternative to a more common path (Figure 2.4). A bubble error was two paths redundant if they start and end at the same nodes and contain similar sequences. Velvet removes bubbles with Tour bus algorithm. The tour bus algorithm detects redundant of paths using breadth-first search, and uses a combination of copy number and topographical information to remove the erroneous edges[11] (Figure 2.5).

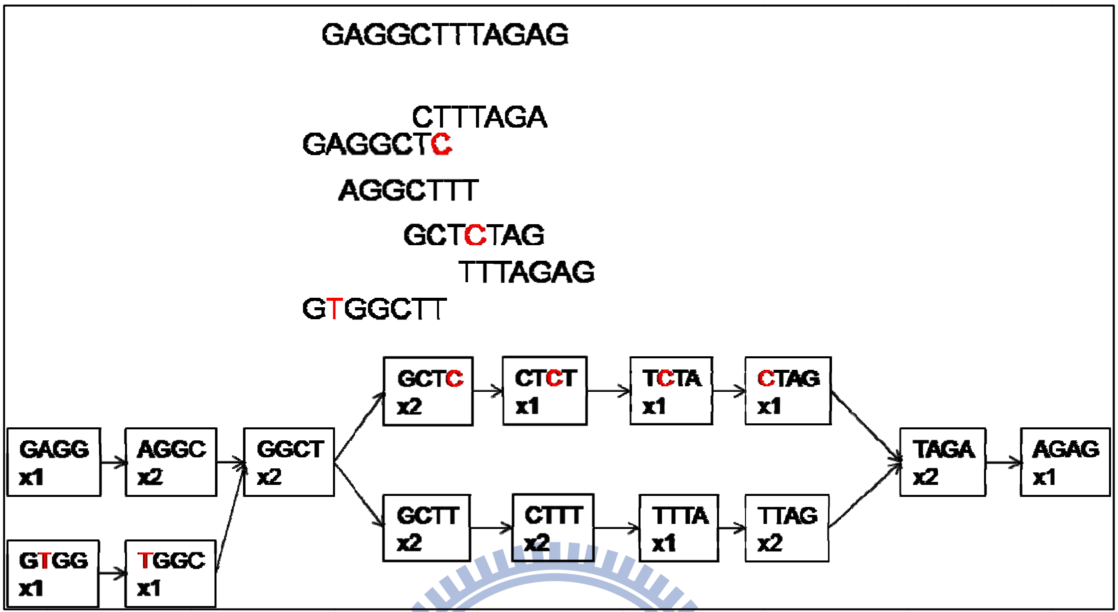


Figure 2.2 Initial de brujin graph

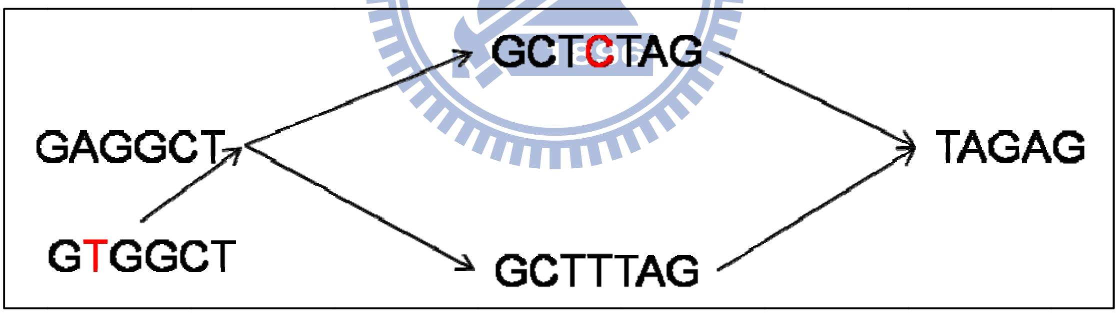


Figure 2.3 Simplification graph

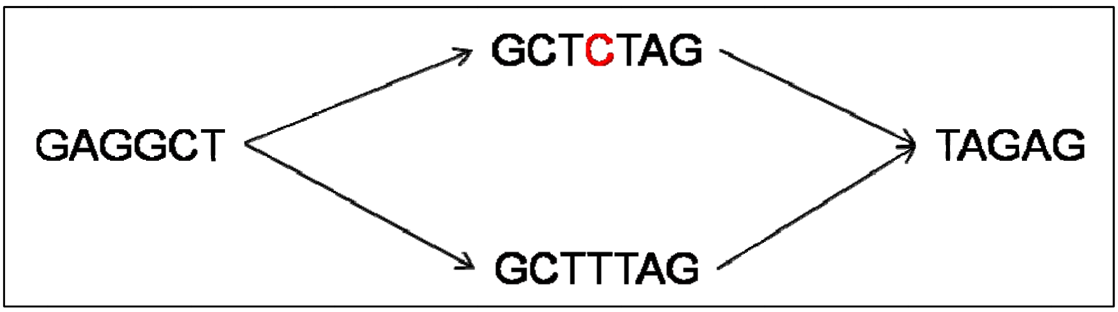


Figure 2.4 Remove tips

Figure 2.5 Remove bubble

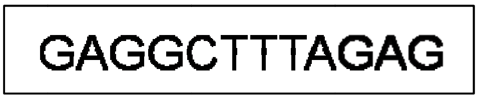


Figure 2.6 Re-simplification graph

2.2 Gene prediction tools

These gene prediction tools which list in below are based on ab initio approach using different Generalized Hidden Markov Model (GHMM) design. GENSCAN was one of first gene prediction tools using GHMM, and it is a very popular gene prediction tool, now. GlimmerHMM incorporates splice site models and utilizes Interpolated Markov Models for the coding and noncoding models. AUGUSUT use a donor splice site model to model intron lengths. SNAP is similar to GENSCAN and adaptable to a number of organisms.

2.2.1 GENSCAN

GENSCAN was a Generalized Hidden Markov Model approach ab initio gene prediction program developed by Chris Burge and Samuel Karlin. Novel features of the program include the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occurring on either or both DNA strands. GENSCAN is shown to have substantially higher accuracy than existing methods when tested on standardized sets of human and vertebrate genes, with 75 to 80% of exons identified exactly. Figure 2.7 is the HMM model, each circle or diamond represents a functional unit (state) of a gene or genomic region: N, intergenic region; P, promoter; F, 5' untranslated region (extending from the start of transcription up to the translation initiation signal); E_{sngl}, single-exon (intronless) gene (translation start to stop codon); E_{init}, initial exon (translation start to donor splice site); E_k ($0 \leq k \leq 2$), phase k internal exon (acceptor splice site to donor splice site); E_{term}, terminal exon (acceptor splice site to stop codon); T, 3' untranslated region (extending from just after the stop codon to the polyadenylation signal); A, polyadenylation signal; and I_k ($0 \leq k \leq 2$), phase k intron[17].

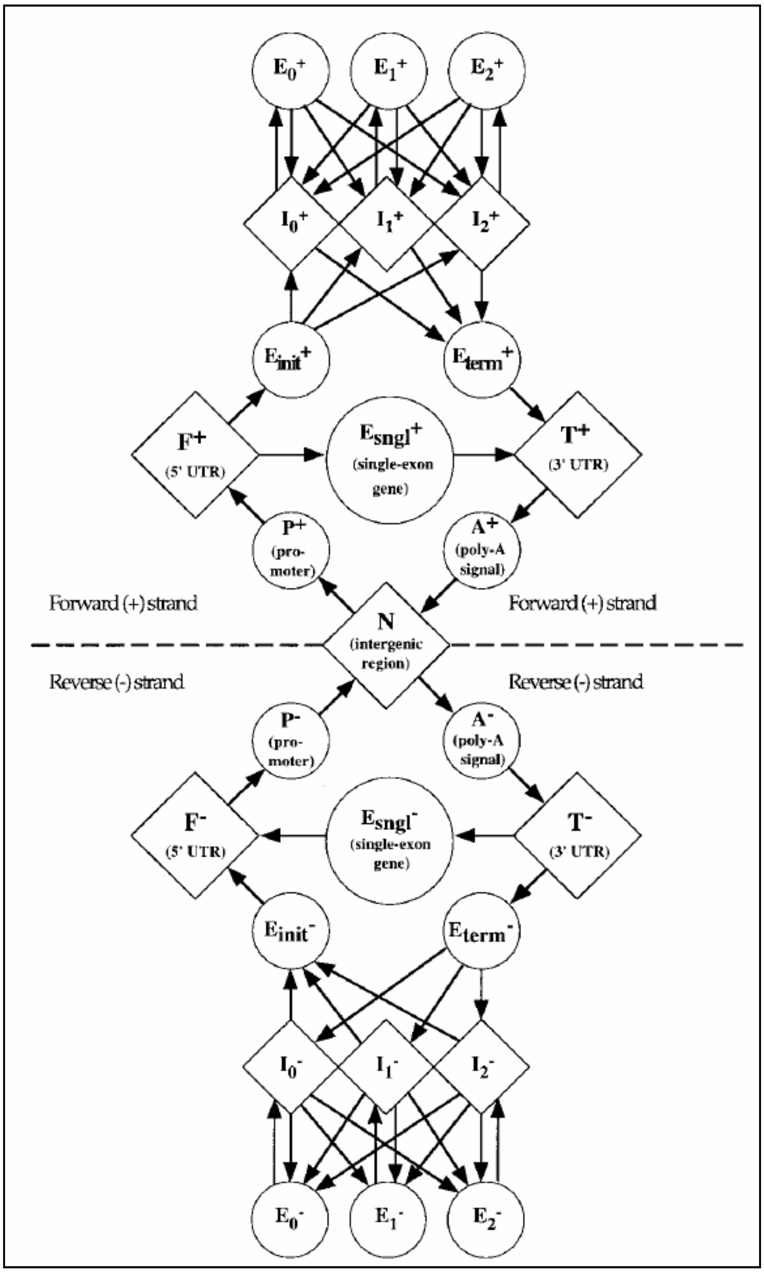


Figure 2.7 HMM model of GENSCAN[4]

2.2.2 GlimmerHMM

GlimmerHMM is a gene finder based on a Generalized Hidden Markov

[4] The picture is copy from Prediction of complete gene structures in human genomic DNA Figure 3 [17]

Model. Although the gene finder conforms to the overall mathematical framework of a GHMM, additionally it incorporates splice site models adapted from the GeneSplicer program and a decision tree adapted from GlimmerM. It also utilizes Interpolated Markov Models for the coding and noncoding models. Currently, GlimmerHMM's GHMM structure includes introns of each phase, intergenic regions, and four types of exons (initial, internal, final, and single). Figure 2.8 is the HMM model, the dashed line in the middle separates the positive strand and negative strand portions of the model. Each state in the GHMM is implemented as a separate submodel, such as a weight array matrix or an IMM (interpolated Markov models)[19].

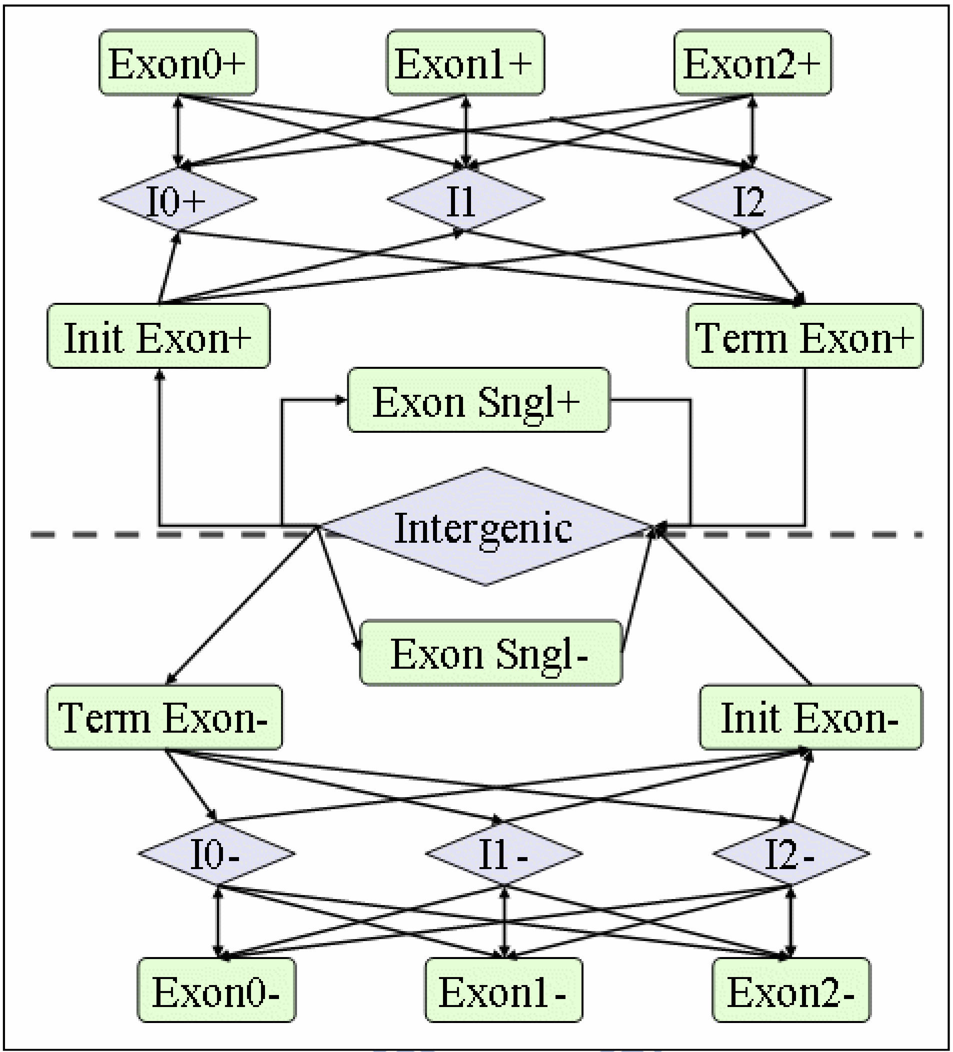


Figure 2.8 The HMM model of GlimmerHMM[5]

2.2.3 AUGUSUT

AUGUSUT was a Generalized Hidden Markov Model approach ab initio gene prediction program developed by Mario Stanke and Stephan Waack. The program is based on a Hidden Markov Model and integrates a number of known methods and submodels. It employs a new way of modeling intron lengths. It use a new donor splice site model, a new

[5] The picture is copy from http://www.cbcb.umd.edu/software/GlimmerHMM/

model for a short region directly upstream of the donor splice site model that takes the reading frame into account and apply a method that allows better GC-content dependent parameter estimation[15].

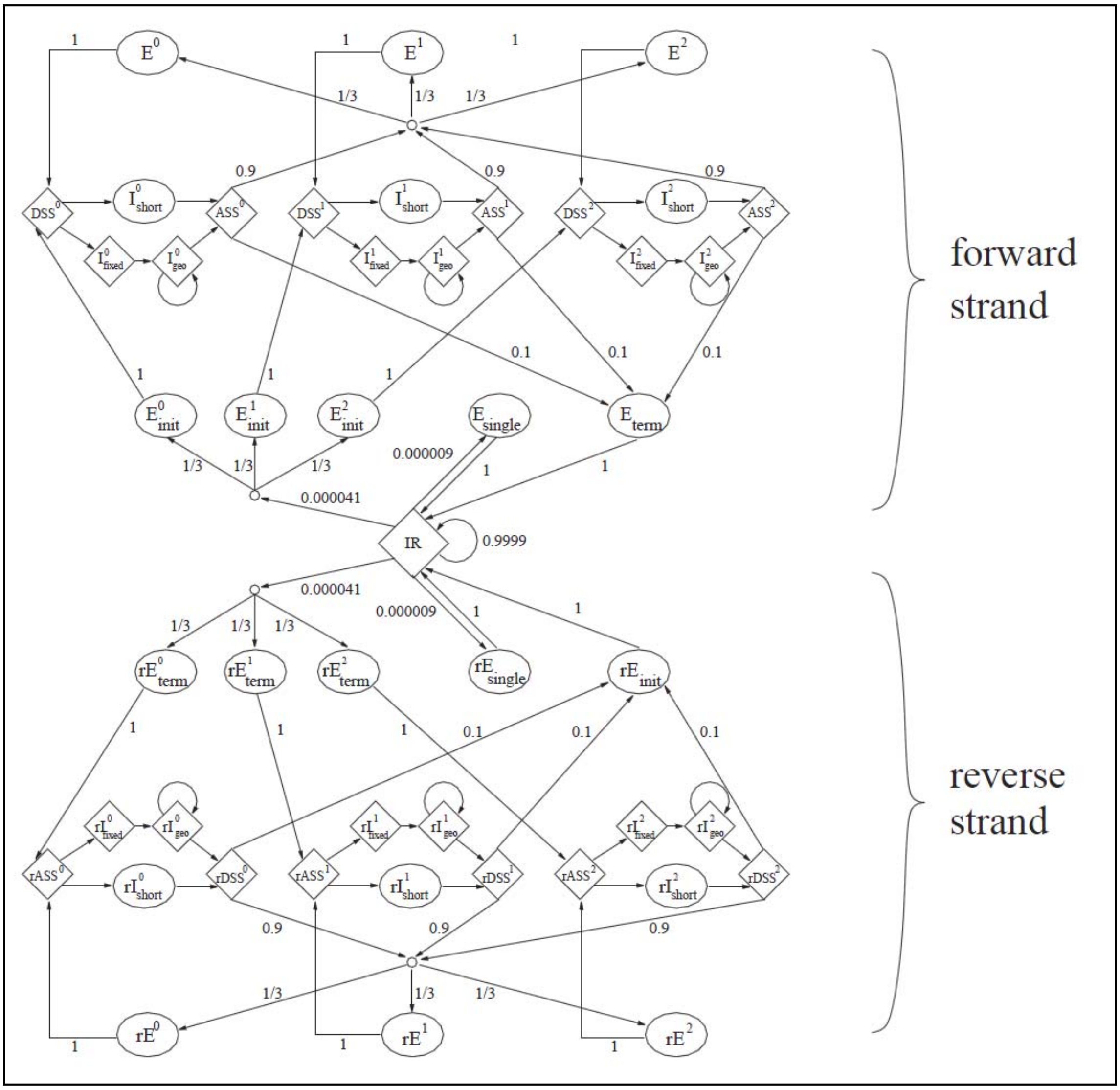


Figure 2.9 The HMM model of AUGUSUT[6]

6 The picture is copy from Gene prediction with a hidden Markov model and a new intron submodel Figure 1 [15]

2.2.4 SNAP

SNAP is similar to GENSCAN and other generalized hidden Markov model (HMM) gene finders, but unlike many, it is easily adaptable to a number of organisms and its source code is freely available[23].

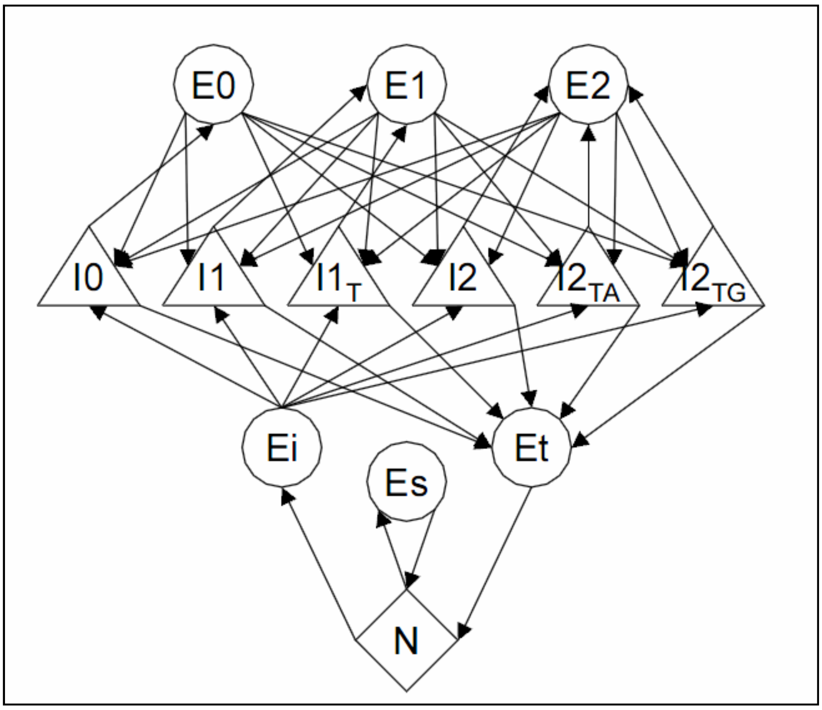


Figure 2.10 The HMM model of SNAP[7]

[7] The picture is copy from Gene finding in novel genomes Figure 1 SNAP HMM state diagram [23]

2.3 Metabolic Pathway Databases

2.3.1 Biocyc

BioCyc is a collection of 673 Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism[24].

2.3.2 Metacyc

MetaCyc is a database of nonredundant, experimentally elucidated metabolic pathways. MetaCyc contains more than 1,500 pathways from more than 1,900 different organisms, and is curated from the scientific experimental literature[25].

2.3.3 KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database of biological systems, consisting of genetic building blocks of genes and proteins, chemical building blocks of both endogenous and exogenous substances, molecular wiring diagrams of interaction and reaction networks, and hierarchies and relationships of various biological objects[26].

2.4 Metabolic Pathway Reconstruction Tools

2.4.1 Pathway tools

A popular metabolic reconstruction tool is Pathway Tool. It used a PathoLogic method computationally reconstructs organism-specific metabolic pathways and generates a new PGDB by matching the Enzyme Commission (EC) number and/or the name of the annotated gene product against enzymes in MetaCyc[27].

Chapter 3 Materials and methods

Our genome reconstruction platform includes three parts: sequence assembly, gene annotation and metabolic pathway. Sequence assembly is major part that support different sequencing platform. Gene annotation and metabolic pathway reconstruction are downstream analysis of assembled sequences.

3.1 Materials

Our pipeline integrates several biological data source and software. The description of the data source and software which integrate in our pipeline is in below.

Assembly tools

Phrap had been widely used in Human Genome Project, and the late version may support Roche 454 sequencing data. Velvet was one of first assemble short reads using de bruijn graph this algorithm can reduce effective of repeat region in short reads assembly. These two assemble tools are very popular on each support sequencing platform data. Velvet used to assemble short reads (Illumina reads), and phrap used to assemble long reads (Sanger or 454 reads) in our pipeline (Table 3.1).

Table 3.1 Assembly tools of materials

Category	Software	Reference
Assembling short reads	Velvet	[11]
Assembling long reads and mixture reads	phrap	[8]

Gene prediction tools

GENSCAN was one of first gene prediction based on GHMM, and it is a flag of gene prediction tools almost later gene prediction tools would compare performance and accuracy with GENSCAN. GlimmerHMM, AUGUSUT and SNAP are within different GHMM design, the performance of these tools are well as GENSCAN and these also suggested by EVidenceModeler.

EVidenceModeler

The EVidenceModeler (EVM) software combines ab intio gene predictions and protein and transcript alignments into weighted consensus gene structures. EVM provides a flexible and intuitive framework for combining diverse evidence types into a single automated gene structure annotation system[28].

GBrowse

The Generic Model Organism System Database Project (GMOD) seeks to develop reusable software components for model organism system databases. The Generic Genome Browser (GBrowse), a Web-based application for displaying genomic annotations and other features[29]. WormBase, FlyBase, and Human Genome Segmental Duplication Database build using GBrowse.

Table 3.2 lists software using in gene annotation of our pipeline.

Table 3.2 Gene prediction tools of materials

Category	**Software**	**Reference**
Ab initio gene prediction	GENSCAN	[17]
	GlimmerHMM	[19]
	AUGUSUT	[15]
	SNAP	[23]
Evidence-based gene prediction	BLAST	[20]
Combinational gene prediction	EVidenceModeler	[28]
Genome viewer	GBrowse	[29]

Gene annotation databases:

Swiss-Prot

The Swiss-Prot protein knowledgebase connects amino acid sequences with the current knowledge in the Life Sciences. Each protein entry provides an interdisciplinary overview of relevant information by bringing together experimental results, computed features and sometimes even contradictory conclusions[30].

Genbank

GenBank is a comprehensive database that contains publicly available nucleotide sequences for more than 260,000 named organisms, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects.

Table 3.3 list databases using in gene annotation of our pipeline.

Table 3.3 Gene annotation databases of materials

Category	**Software**	**Reference**
Protein database	Swiss-Prot	[30]
Gene database	Genbank	[31]

Metabolic pathway reconstruction tools

Pathway Tools has been generated 673 Pathway/Genome Databases in the Biocyc. It is the most widely used pathway reconstruction tool.

Table 3.4 lists software using in metabolic pathway reconstruction of our pipeline.

Table 3.4 Metabolic pathway reconstruction tools of materials

Category	**Software**	**Reference**
Metabolic pathway reconstruction	Pathway Tools	[27]

3.2 The processes of genome annotation

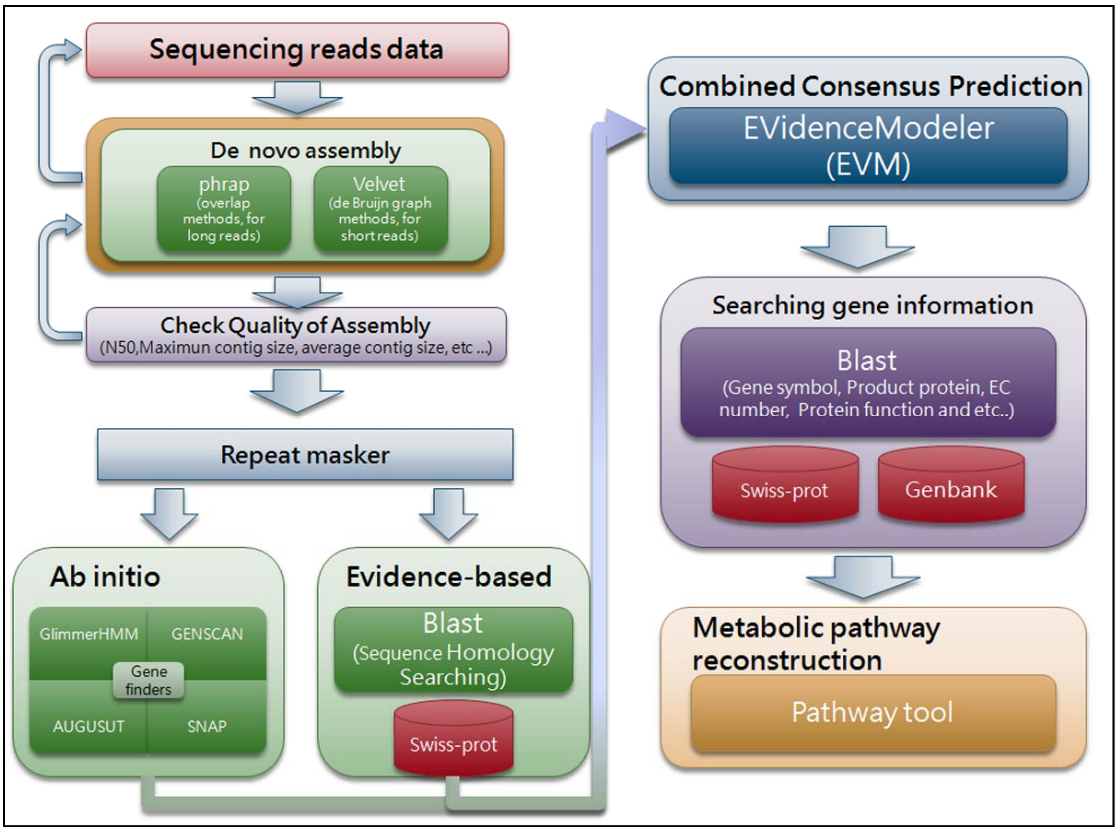


Figure 3.1 The schematic indicates the processes of annotating of a novel genome including sequence assembly, gene annotation, and metabolic pathway reconstruction.

Figure 3.1 presents the work flow of our computational pipeline. First step is sequence assembly, according to different sequencing data select assembly tools. This pipeline uses phrap[8] to assemble Sanger and 454 data and uses Velvet[11] to assemble Illumina data. After each assembly complete, it checks the quality of assembly with N50, maximum/average contig size and genome coverage. Before gene prediction, this pipeline uses Repeatmasker[32] to mask repeat region for avoid these region effect accuracy of gene prediction. This pipeline use four ab initio gene prediction tools: GlimmerHMM[19], AUGUSUT[15], GENSCAN[17] and SNAP[23] and an evidence-based gene prediction tool: BLAST[20]. BLAST search homology protein sequence against Swiss-prot[30]. After ab initio and evidence-based gene prediction, this pipeline uses EVidenceModeler[28] to combine consensus gene predictions, and uses BLAST to search Swiss-prot[30] and Genbank[31] for annotate gene information that includes gene symbol, product protein, EC (Enzyme Commission) number and protein function to these genes. According to the gene information, this pipeline use Pathway Tools[27] to reconstruct metabolic pathway for the organisms.

3.3 Methods

3.3.1 Sequence assembly

Our pipeline supports various sequencing raw data such as Sanger, 454, Illumina and mixture of those data. Assembling short reads was using Velvet[11]. Phrap[8] which is an overlap approach assemble tool assemble long sequences (Sanger or 454) in our pipeline, and phrap have been assemble whole genome shotgun sequence within Human genome project. In last version, phrap may support 454 reads and length of 454 reads was enough to assemble Sanger reads with less effect of repeat regions. Assembly of 454 and Illumina reads may be affected by repeat region. Thus, our pipeline used Velvet to assemble Illumina reads first and then combine the result of Velvet assembly with 454 reads. Our pipeline used phrap to assemble the mixture data. After each assembly complete, our pipeline checked the quality of assembly with N50, maximum/average contig size and genome coverage.

3.3.2 Gene annotation

In order to avoid transposons affect the accuracy of gene prediction, our pipeline used RepeaterMasker[32] to mask those region. RepeatMasker screens low complexity DNA sequence and LTRs(Long Terminal Repeats) in genome sequence and replace those region letters to N's. Our gene prediction combined ab initio and evidence-based approach. Our pipeline used four ab initio approach gene prediction tools (GlimmerHMM[19], AUGUSUT[15], GENSCAN[17] and SNAP[23]) and used BLAST[20] to search Swiss-prot[30] protein database for homologous proteins. EVidenceModeler[28] can combine ab initio and evidence-based gene predictions into weighted consensus gene structures. Our pipeline sets ab initio with weight 1 and evidence-based with weight 3. Consensus gene predictions generated by EVidenceModeler were identified gene information include gene symbol, product protein, EC number and protein function.

These gene information identified by using BLAST align gene sequence and protein sequence form Swiss-prot to searching similar protein and searching Genbank.

3.3.3 Metabolic pathway reconstruction

Metabolic pathway reconstruction was using Pathway Tools[27] with the information of gene annotation, and the work flow show on Figure 3.2. An initial PGDB (Pathway/Genome database) was created for each contig and each gene. The metabolic reactions identified by matching the EC number and the name of gene product against the MetaCyc[33], and the reactions known to be catalyzed are matched against all the pathways in MetaCyc. Pathway Tools imports the pathway and its associated reactions and substrates from MetaCyc into the initial PGDB. The initial PGDB have some pathway holes which are the enzymes missing each predicted pathways. Pathway holes occur when a protein has not been a specific function during annotation process, and reactions catalyzed by this protein will have a pathway hole in PGDB. The pathway hole-filler is implemented as part of the Pathway Tools[34]. The hole-filler uses isozyme sequences to search a genome for similar sequences. These isozyme sequences retrieve Swiss-Prot IDs directly from the ENZYME database[35] and retrieves PIR[36] IDs from the MetaCyc. Homology searching these isozyme sequences against the genome sequence using BLAST as candidates. Finally, hole-filler evaluates these candidate proteins to determine the probability that each candidate protein has the activity required by the missing reaction, and use these proteins to fill pathway holes.

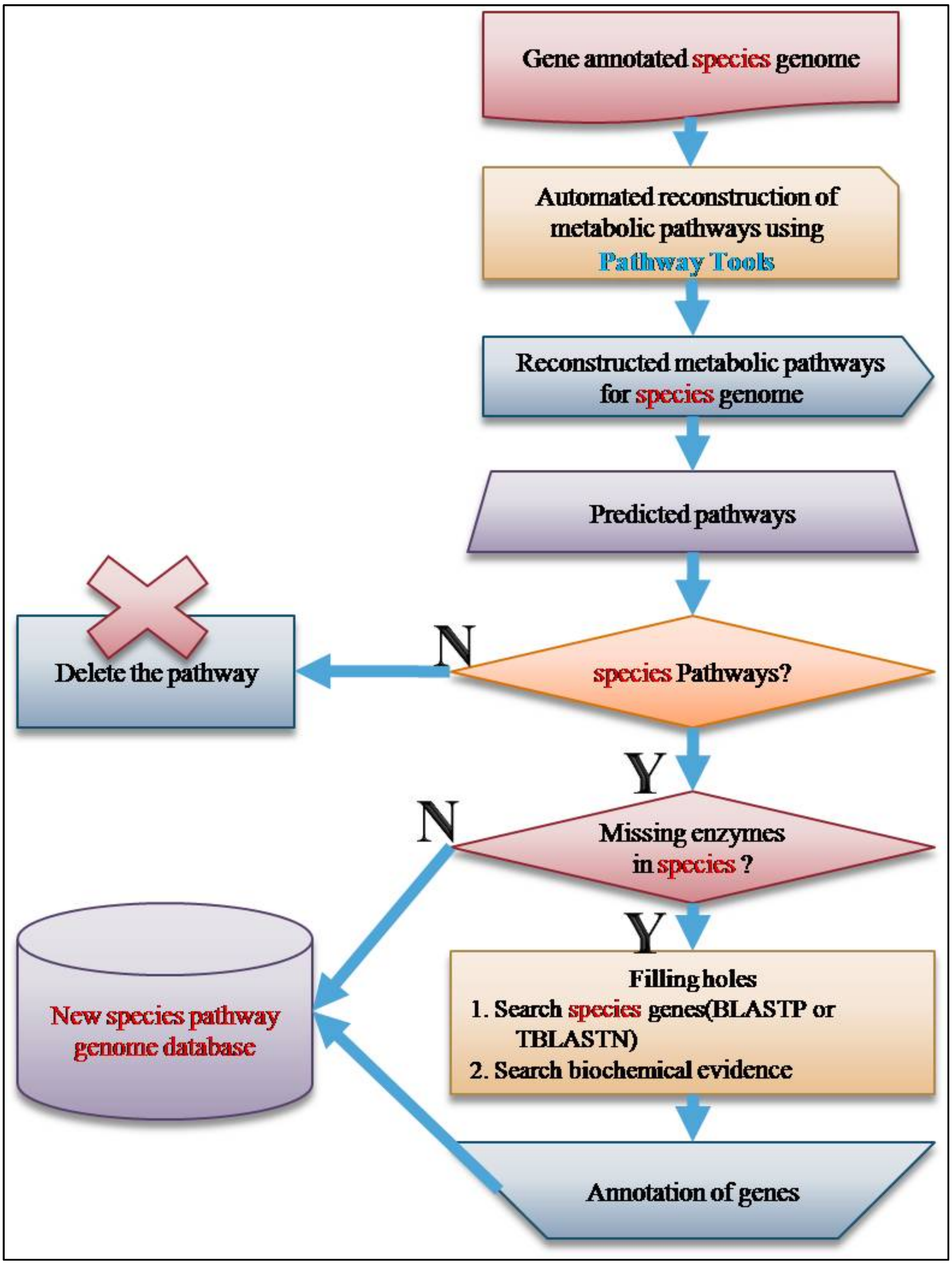


Figure 3.2 Metabolic pathway reconstruction work flow

Chapter 4 Results

We simulated and reconstructed Saccharomyces cerevisiae (yeast) genome. Saccharomyces cerevisiae included 16 chromosomes, 5861 protein coding genes and total length of all chromosomes is 12,244,764 bp.

4.1 Result of Sequence assembly

This simulation of Saccharomyces cerevisiae genome randomly spliced Saccharomyces cerevisiae into several type read data: 100bp~200bp, 200bp~400bp, 36bp single read (36bp_sr), 36bp paired-end with 100bp insert (36bp_pe), 100bp single read (100bp_sr), 100bp paired-end with 100bp insert (100bp_pe). The 100bp~200bp and 200bp~400bp were simulation 454 type data, and another reads data were simulation Illumina type data (Table 4.1). The 454 type data were with 10x coverage depth (12,244,764 bp*10) and 0.1% errors, and the Illumina type data were with 20x coverage depth (12,244,764 bp*20) and 1% error. Our pipeline assembled 454 type data using phrap[8] and Illumina type data using Velvet[11], and the assembly results shown on Table 4.2. Our pipeline evaluates the quality of assembly according to N50. N50 of 200bp~400bp data set is 153,047 and it is better than 100bp~200bp data set (N50 is 42,367) in 454 type data sets. In comparison of Illumina type data sets, paired-end data have a better quality than single read and 100bp length reads data is better than 36bp.

Table 4.1 Data sets for each sequencing platform

Simulation sequencing platform	**Data set**	**Coverage**
Roche 454	100bp~200bp	10x
	200bp~400bp	10x
Illumina	36bp single read	20x
	36bp paired-end with 100bp insert	20x
	100bp single read	20x
	100bp paired-end with 100bp insert	20x

Table 4.2 The comparison of data sets assembly

Reads data	**N50 (bp)**	**Number of contigs (bp)**	**Average length of contigs (bp)**	**Max length of contigs (bp)**	**Total length of contigs (bp)**
100bp~200bp	42,367	572	20,505.01	201,650	11,728,869
200bp~400bp	153,047	155	76,001.27	605,145	11,780,198
36bp_sr	1,816	9,700	1,160.63	10,044	11,258,206
36bp_pe	2,527	6,927	1,645.39	12,653	11,397,637
100bp_sr	8,597	3,011	3,753.70	41,128	11,302,409
100bp_pe	31,048	1,226	9,308.53	130,864	11,412,260

Because the 200bp~400bp and 100bp_pe had the best N50 in their simulation type, and select these two data set to mixture data simulation. Our pipeline assembled 200bp~400bp + 100bp_pe (454_Illumina) and 200bp~400bp + the result of velvet assembly 100bp_pe (454_velvet) using phrap, and the assembly results shown on Table 4.3.

In the mixture data assembly simulation, 454_velvet date set has a better N50 (322,842 bp) than 454_Illumina (222,376 bp), and it is also better than another simulation data set.

Illumina data assemble to longer contigs (result of velvet assembly) and then assembly with 454 data (454_velvet) will better than immediate assembly Illumina and 454 data.

Table 4.3 The comparison of two mixture data set

Reads data	**N50 (bp)**	**Number of contigs (bp)**	**Average length of contigs (bp)**	**Max length of contigs (bp)**	**Total length of contigs (bp)**
454_Illumina	222,376	417	28,813.92	653,700	12,015,405
454_velvet	322,842	121	101,649.92	1,101,185	12,299,641

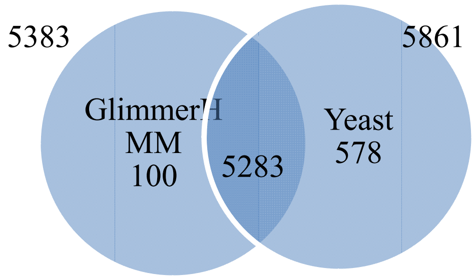
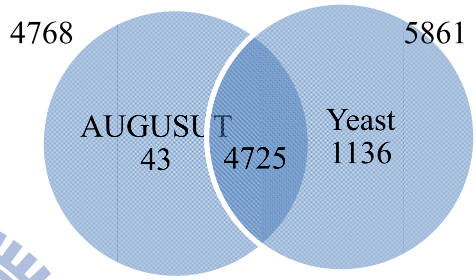
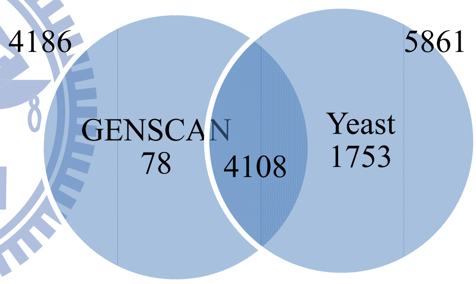
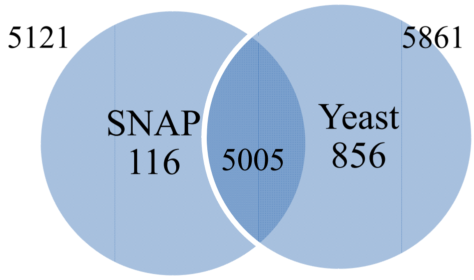
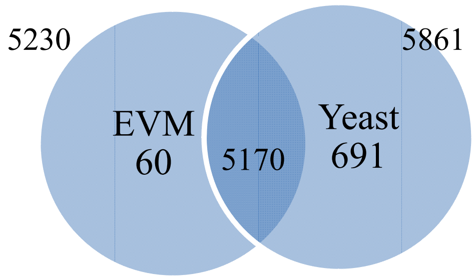
4.2 Result of Gene Annotation

Our pipeline used GlimmerHMM[19], AUGUSUT[15], GENSCAN[17] and SNAP[23] to ab initio gene prediction, BLAST to evidence-based gene prediction, and EVidenceModeler[28] to combine ab initio prediction result with weight 1 and evidence-based prediction result with weight 3. Prediction result of each prediction tools and comparison with Saccharomyces cerevisiae gene show on Table 4.4. The number of genes predicted by GlimmerHMM is 5383, the number of genes match to genes on Saccharomyces cerevisiae is 5283, the number of mismatch genes is 578 and the number of additional genes is 100. The number of genes predicted by AUGUSUT is 4768, the number of genes match to genes on Saccharomyces cerevisiae is 4725, the number of mismatch genes is 1136 and the number of additional genes is 43. The number of genes predicted by GENSCAN is 4186, the number of genes match to genes on Saccharomyces cerevisiae is 4108, the number of mismatch genes is 1753 and the number of additional genes is 78. The number of genes predicted by SNAP is 5121, the number of genes match to genes on Saccharomyces cerevisiae is 5005, the number of mismatch genes is 856 and the number of additional genes is 116. The number of genes predicted by

EvidenceModeler is 5230, the number of genes match to genes on Saccharomyces cerevisiae is 5170, the number of mismatch genes is 691 and the number of additional genes is 60.

GlimmerHMM match the most gene on Saccharomyces cerevisiae: 90% (5283/5861). GENSCAN match the least gene on Saccharomyces cerevisiae: 70% (4108/5861). EvidenceModeler match second gene on Saccharomyces cerevisiae: 88% (5170/5861).

Table 4.4 The comparison of gene predicted by each prediction tool with Saccharomyces cerevisiae gene

Gene prediction tools	**Number of gene**	**Match gene on yeas**	**Comparison of Predicted gene with evidence gene**
GlimmerHMM	5383	5283 (90%)	
AUGUSUT	4768	4725 (80%)	
GENSCAN	4186	4108 (70%)	
SNAP	5121	5005 (85%)	
EvidenceModeler (EVM)	5230	5170 (88%)	

EVidenceModeler predicted 5230 gene and searched homology proteins using BLAST against Swiss-prot. Our pipeline annotated gene symbol, product protein, EC number and Protein function to these gene form Swiss-prot and Genbank. These genes which predicted by EVidenceModeler include 5170 gene which matched to Saccharomyces cerevisiae gene and 691 gene which mismatch to Saccharomyces cerevisiae gene.

Our pipeline displayed gene location on contigs (Figure 4.1), information of annotation (Figure 4.2) and generated gene database using GBrowse[29].

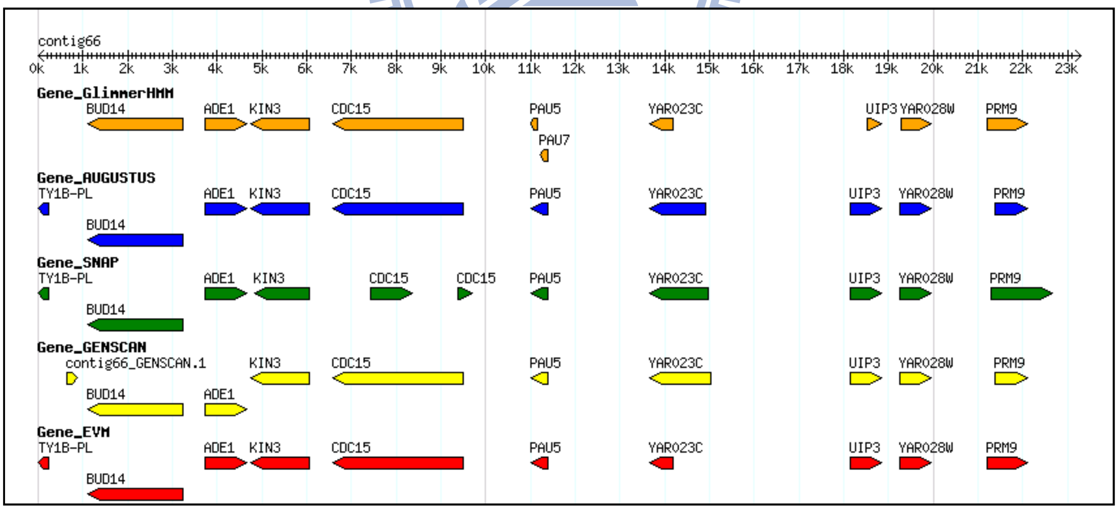


Figure 4.1 Gene locations on contig66

PAU5 Details

Name:	PAU5	
Type:	gene	
Description:	Seripauperin-5	
Source:	EVM	
Position:	contig66:11024..11392 (- strand)	
Length:	369	
Alias:	P43575	
	YFL020C	
Note:	Seripauperin-5	
gene:	PAU5	
load_id:	evm.TU.contig66.6	
Parts:	Type:	CDS
	Description:	
	Source:	EVM
	Position:	contig66:11024..11223 (- strand)
	Length:	200
	load_id:	evm.TU.contig66.6.e2
	parent_id:	evm.TU.contig66.6
	Type:	CDS
	Description:	
	Source:	EVM
	Position:	contig66:11278..11392 (- strand)
	Length:	115
	load_id:	evm.TU.contig66.6.e1
	parent_id:	evm.TU.contig66.6

Figure 4.2 PAU5 detail information

Table 4.5 list three type example genes: perfect match gene, good match gene that have few differences and bad match gene. Perfect match gene: SNC1 and its gene structure includes 102bp length CDS, 113bp length intron and 252bp length CDS. Good match gene: PAU5 and its gene structure includes 200bp length CDS, 54bp length intron and 115bp length CDS, but the real gene structure only include a 369bp length CDS without intron. Bad match gene: YPL278C and its gene structure includes 236bp length CDS, 459bp length intron and 103bp length CDS, but the real gene structure only include a 303bp length CDS.

Table 4.5 The three type example genes

Match type	Gene	Gene structure of our predicted genes	Gene structure of yeast genes	Total length of our predicted genes	Total length of yeast genes
Perfect	SNC1	CDS(102bp)→ intron(113bp)→ CDS(252bp)	CDS(102bp)→ intron(113bp)→ CDS(252bp)	467bp	467bp
Good	PAU5	CDS(200bp)→ intron(54bp)→ CDS(115bp)	CDS(369bp)	369bp	369bp
Bad	YPL278C	CDS(236bp)→ intron(459bp)→ CDS(103bp)	CDS(303bp)	798bp	303bp

4.3 Result of Metabolic Pathway Reconstruction

Our pipeline reconstructed metabolic pathway of Saccharomyces cerevisiae using Pathway Tools with information of gene annotation. The initial pathways included 200 metabolic pathways, but there were 268 pathway holes in the initial pathways. Our pipeline filled 268 pathway holes to 196, and the detail comparison of initial pathways (Initial) with the pathways which filled pathway holes (Hole-filled) showed on Table 4.6.

Table 4.6 The comparison of initial pathway with hole-filled pathway

Database statistics	**Initial**	**Hole-filled**
Metabolic pathways	200	200
Enzymatic reactions	1274	1294
Enzymes	1260	1269
Compounds	865	873
Number of Pathway Holes	268	196
Pathway Holes as a percentage of total reactions in pathways	37%	27%
Pathways with No Holes	97	121
Pathways with 1 Hole	47	34
Pathways with 2 Holes	18	14
Pathways with 3 Holes	12	15
Pathways with 4 Holes	7	4
Pathways with 5 Holes	5	2
Total Pathways with Holes	103	79

Table 4.7 showed comparison of the number of pathways in hole-filled pathway (Hole-filled) with the number of pathways in YeastCyc[37] pathway database.

Table 4.7 The comparison of hole-filled pathway with YeastCyc pathway database

Pathway Class	**The number of pathways in YeastCyc**	**The number of pathways in Hole-filled**
Biosynthesis	110	145
- Amines and Polyamines Biosynthesis	4	4
- Amino acids Biosynthesis	29	33
- Aminoacyl-tRNA Charging	0	2
- Aromatic Compounds Biosynthesis	1	2
- Carbohydrates Biosynthesis	7	9
- Cell structures Biosynthesis	0	0
- Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis	22	34
- Fatty Acids and Lipids Biosynthesis	13	23
- Hormones Biosynthesis	0	0
- Metabolic Regulators Biosynthesis	0	0
- Nucleosides and Nucleotides Biosynthesis	8	8
- Other Biosynthesis	0	2
- Secondary Metabolites Biosynthesis	1	9
- Siderophore Biosynthesis	0	0
Degradation/Utilization/Assimilation	40	67
- Alcohols Degradation	2	5
- Aldehyde Degradation	1	2
- Amines and Polyamines Degradation	1	3
- Amino Acids Degradation	18	23
- Aromatic Compounds Degradation	0	1
- C1 Compounds Utilization and Assimilation	1	2
- Carbohydrates Degradation	6	4

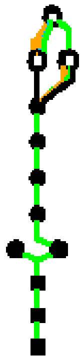
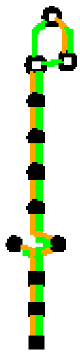
- Carboxylates Degradation	1	3
- Chlorinated Compounds Degradation	0	0
- Cofactors, Prosthetic Groups, Electron Carriers Degradation	1	0
- Degradation/Utilization/Assimilation - Other	0	2
- Fatty Acid and Lipids Degradation	3	9
- Hormones Degradation	0	0
- Inorganic Nutrients Metabolism	1	3
- Nucleosides and Nucleotides Degradation and Recycling	0	1
- Polymeric Compounds Degradation	0	1
- Secondary Metabolites Degradation	0	3
Generation of precursor metabolites and energy	11	21
Signal transduction pathways	0	0
Total	133	204

The pathways generated by our pipeline (Hole-filled) compares to YeastCyc with some important pathways. These pathways included: gluconeogenesis, glycerol degradation, glycolysis, pentose phosphate pathway, glyoxylate cycle, TCA cycle and fatty acid oxidation pathway.

Gluconeogenesis

The pathways generated by our pipeline (Hole-filled) included the most pathways in YeastCyc. The difference was our pathway had no pyruvate to oxaloacetic acid reaction catalyzed by 6.4.1.1. The more detail comparison showed on Table 4.8, Figure 4.3 and Figure 4.4

Table 4.8 The comparison gluconeogenesis between the pathways generated by our pipeline (Hole-filled) and YeastCyc

Hole-filled			**YeastCyc**		
Evidence Glyph	**Enzymes**	**Genes**	**Evidence Glyph**	**Enzymes**	**Genes**
	EC# 1.1.1.37	Malate dehydrogenase, mitochondrial: MDH1 Malate dehydrogenase, peroxisomal: MDH3 Malate dehydrogenase, cytoplasmic: MDH2		EC# 1.1.1.38	malic enzyme: MAE1
	EC# 4.1.1.31	None		EC# 6.4.1.1	pyruvate carboxylase: PYC1 pyruvate carboxylase: PYC2
	EC# 4.1.1.49	Phosphoenolpyruvate carboxykinase [ATP]: No Gene Name Phosphoenolpyruvate carboxykinase [ATP]: PCK1		EC# 1.1.1.37	peroxisome malate dehydrogenase: MDH3 mitochondrial malate dehydrogenase: MDH1 cytosolic malate dehydrogenase: MDH2
	EC# 1.1.1.40	NAD-dependent malic enzyme, mitochondrial: MAE1		EC# 4.1.1.49	phosphoenolpyruvate carboxylkinase: PCK1
	EC# 1.1.1.38	NAD-dependent malic enzyme, mitochondrial: MAE1		EC# 4.2.1.11	enolase: ENO2 enolase I: ENO1
	EC# 2.7.9.2	None		EC# 5.4.2.1	phosphoglycerate mutase: GPM1
	EC# 4.2.1.11	Enolase-related protein 3: ERR3 Enolase-related protein 1/2: ERR1 Enolase 2: ENO2 Enolase 1: ENO1		EC# 2.7.2.3	3-phosphoglycerate kinase: PGK1
	EC# 5.4.2.1	Phosphoglycerate mutase 1: GPM1 Phosphoglycerate mutase 2:		EC# 1.2.1.12	glyceraldehyde-3-phosphate dehydrogenase: TDH1 glyceraldehyde 3-phosphate

	GPM2 Probable phosphoglycerate mutase YOR283W: YOR283W Phosphoglycerate mutase 3: GPM3 Putative phosphoglycerate mutase DET1: DET1		dehydrogenase: TDH2 glyceraldehyde-3-phosphate dehydrogenase: TDH3
EC# 2.7.2.3	Phosphoglycerate kinase: PGK1	EC# 4.1.2.13	aldolase: FBA1
EC# 1.2.1.12	Glyceraldehyde-3-phosphate dehydrogenase 1: TDH1 Glyceraldehyde-3-phosphate dehydrogenase 2: TDH2 Glyceraldehyde-3-phosphate dehydrogenase 3: TDH3	EC# 3.1.3.11	fructose-1,6-bisphosphatase: FBP1
EC# 4.1.2.13	Fructose-bisphosphate aldolase: FBA1	EC# 5.3.1.9	glucose-6-phosphate isomerase: PGI1
EC# 3.1.3.11	Fructose-1,6-bisphosphatase: FBP1		
EC# 5.3.1.9	Glucose-6-phosphate isomerase: PGI1		

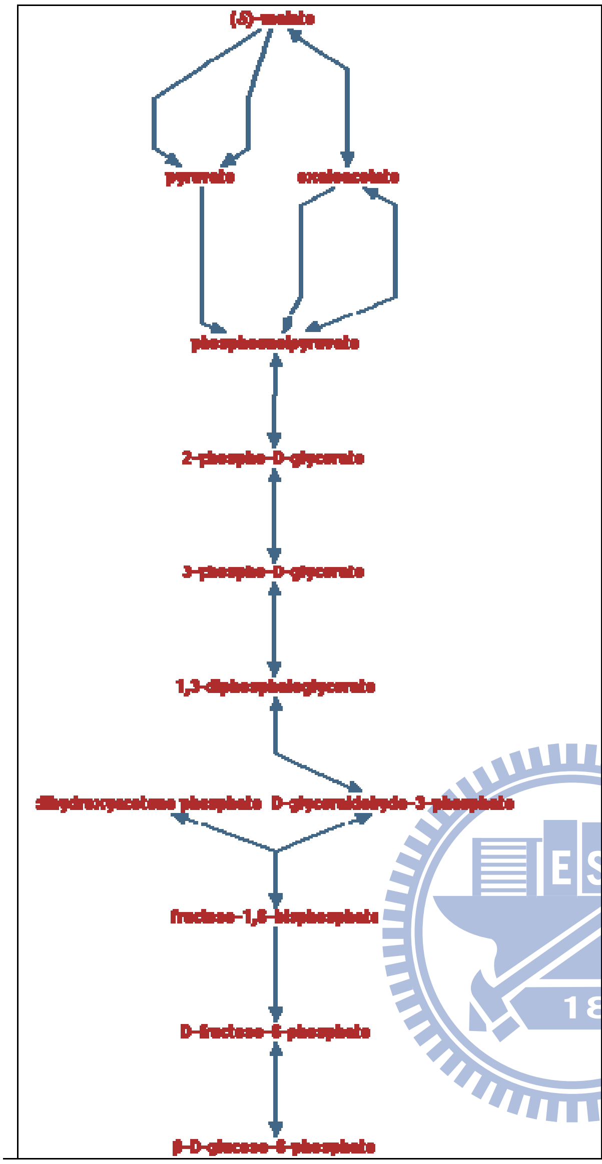


Figure 4.3 Gluconeogenesis in our pathways

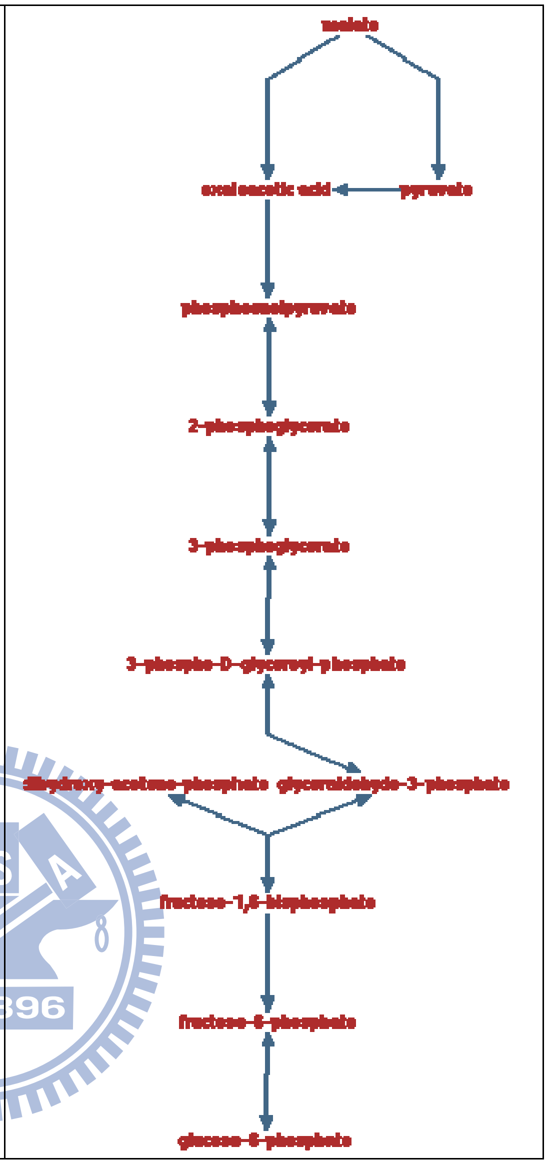


Figure 4.4 Gluconeogenesis in YeastCyc

Glycerol degradation

These pathways were same in the pathways generated by our pipeline (Hole-filled) and YeastCyc. EC 1.1.99.5 had been transferred to EC 1.1.5.3., they are same enzyme. The detail comparison information was on Table 4.9.

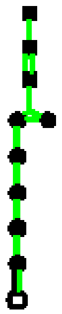
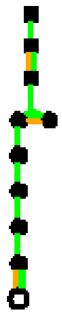
Table 4.9 The comparison glycerol degradation between the pathways generated by our pipeline (Hole-filled)y and YeastCyc

Hole-filled			YeastCyc		
Evidence Glyph	**Enzymes**	**Genes**	**Evidence Glyph**	**Enzymes**	**Genes**
	EC# 2.7.1.30	Glycerol kinase: GUT1		EC# 2.7.1.30	glycerol kinase: GUT1
	EC# 1.1.5.3	Glycerol-3-phosphate dehydrogenase, mitochondrial: GUT2		EC# 1.1.99.5	glycerol-3-phosphate dehydrogenase: GUT2

Glycolysis

These pathways were same in our pathways and YeastCyc. Some enzymes do not match because these enzymes catalyze reverse reactions. The detail comparison information was on Table 4.10.

Table 4.10 The comparison glycolysis between our pathway and YeastCyc

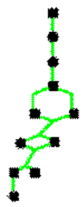
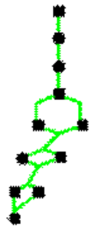
Hole-filled			YeastCyc		
Evidence Glyph	**Enzymes**	**Genes**	**Evidence Glyph**	**Enzymes**	**Genes**
	EC# 5.3.1.9	Glucose-6-phosphate isomerase: PGI1		EC# 5.3.1.9	glucose-6-phosphate isomerase: PGI1
	EC# 2.7.1.11	6-phosphofructokinase subunit beta: PFK2 6-phosphofructokinase subunit alpha: PFK1		EC# 2.7.1.11	phosphofructokinase: PFK2, PFK1
	EC# 3.1.3.11	Fructose-1,6-bisphosphatase: FBP1		EC# 4.1.2.13	aldolase: FBA1
	EC# 4.1.2.13	Fructose-bisphosphate aldolase: FBA1		EC# 5.3.1.1	triosephosphate isomerase: TPI1
	EC# 5.3.1.1	Triosephosphate isomerase: TPI1		EC# 1.2.1.12	glyceraldehyde-3-phosphate dehydrogenase: TDH1 glyceraldehyde 3-phosphate dehydrogenase: TDH2 glyceraldehyde-3-phosphate dehydrogenase: TDH3
	EC# 1.2.1.12	Glyceraldehyde-3-phosphate dehydrogenase 1: TDH1 Glyceraldehyde-3-phosphate dehydrogenase 2: TDH2 Glyceraldehyde-3-phosphate dehydrogenase 3: TDH3		EC# 2.7.2.3	3-phosphoglycerate kinase: PGK1
	EC# 2.7.2.3	Phosphoglycerate kinase: PGK1		EC# 5.4.2.1	phosphoglycerate mutase: GPM1
	EC# 5.4.2.1	Phosphoglycerate mutase 1: GPM1 Phosphoglycerate mutase 2: GPM2 Probable phosphoglycerate mutase YOR283W: YOR283W Phosphoglycerate mutase 3:		EC# 4.2.1.11	enolase: ENO2 enolase I: ENO1

	GPM3 Putative phosphoglycerate mutase DET1: DET1		
EC# 4.2.1.11	Enolase-related protein 3: ERR3 Enolase-related protein 1/2: ERR1 Enolase 2: ENO2 Enolase 1: ENO1	EC# 2.7.1.40	pyruvate kinase: PYK2 pyruvate kinase: CDC19
EC# 2.7.1.40	Pyruvate kinase 1: PYK1 Pyruvate kinase 2: PYK2		
EC# 2.7.9.2	None		

Pentose phosphate pathway

These pathways were same in the pathways generated by our pipeline (Hole-filled) and YeastCyc. The detail comparison information was on Table 4.11.

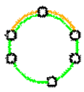
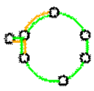
Table 4.11 The comparison pentose phosphate pathway between the pathways generated by our pipeline (Hole-filled) and YeastCyc

Hole-filled			**YeastCyc**		
Evidence Glyph	**Enzymes**	**Genes**	**Evidence Glyph**	**Enzymes**	**Genes**
	EC# 1.1.1.49	Glucose-6-phosphate 1-dehydrogenase: ZWF1		EC# 1.1.1.49	glucose-6-phosphate dehydrogenase: ZWF1
	EC# 3.1.1.31	6-phosphogluconolactonase 3: SOL3 6-phosphogluconolactonase 4: SOL4		EC# 3.1.1.31	6-phosphogluconolactonase: SOL4 6-phosphogluconolactonase: SOL3
	EC# 1.1.1.44	6-phosphogluconate dehydrogenase, decarboxylating 1: GND1 6-phosphogluconate dehydrogenase, decarboxylating 2: GND2		EC# 1.1.1.44	6-phosphogluconate dehydrogenase, decarboxylating: GND1 6-phosphogluconate dehydrogenase: GND2
	EC# 5.3.1.6	Ribose-5-phosphate isomerase: RKI1 Ribose-5-phosphate isomerase: RKI1		EC# 5.3.1.6	ribose-5-phosphate ketol-isomerase: RKI1
	EC# 5.1.3.1	Ribulose-phosphate 3-epimerase: RPE1		EC# 5.1.3.1	D-ribulose-5-Phosphate 3-epimerase: RPE1
	EC# 2.2.1.1	Transketolase 1: TKL1 Transketolase 2: TKL2		EC# 2.2.1.1	transketolase: TKL1 transketolase: TKL2
	EC# 2.2.1.2	Transaldolase NQM1: NQM1 Transaldolase NQM1: NQM1 Transaldolase: TAL1		EC# 2.2.1.2	transaldolase: TAL1
	EC# 2.2.1.1	Transketolase 1: TKL1 Transketolase 2: TKL2		2TRANSKETO-RXN	transketolase: TKL1

Glyoxylate cycle

These pathways were same in the pathways generated by our pipeline (Hole-filled) and YeastCyc. The additional node in YeastCyc evidence glyph includes in our pathway and it just not show on the graph. The detail comparison information was on Table 4.12.

Table 4.12 The comparison glyoxylate cycle between the pathways generated by our pipeline (Hole-filled) and YeastCyc

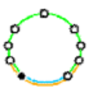
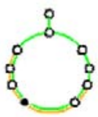
Hole-filled			**YeastCyc**		
Evidence Glyph	**Enzymes**	**Genes**	**Evidence Glyph**	**Enzymes**	**Genes**
	EC# 2.3.3.9	Malate synthase 2, glyoxysomal: MSL2 Malate synthase 1, glyoxysomal: MLS1		EC# 1.1.1.37	peroxisome malate dehydrogenase: MDH3 mitochondrial malate dehydrogenase: MDH1 cytosolic malate dehydrogenase: MDH2
	EC# 1.1.1.37	Malate dehydrogenase, mitochondrial: MDH1 Malate dehydrogenase, peroxisomal: MDH3 Malate dehydrogenase, cytoplasmic: MDH2		EC# 2.3.3.1	citrate synthase: CIT3 citrate synthase: CIT1 citrate synthase: CIT2
	EC# 2.3.3.1	Citrate synthase, mitochondrial: CIT1 Citrate synthase, peroxisomal: CIT2 Citrate synthase 3: CIT3		EC# 4.2.1.3	aconitase: ACO1 aconitate hydratase: ACO2
	EC# 4.2.1.3	Probable aconitate hydratase 2: ACO2 Aconitate hydratase, mitochondrial: ACO1 Aconitate hydratase,		EC# 4.2.1.3	aconitase: ACO1

		mitochondrial: ACO1			
	EC# 4.2.1.3	Probable aconitate hydratase 2: ACO2 Aconitate hydratase, mitochondrial: ACO1 Aconitate hydratase, mitochondrial: ACO1		EC# 4.1.3.1	isocitrate lyase: ICL1
	EC# 4.1.3.1	Isocitrate lyase: ICL1		EC# 2.3.3.9	malate synthase: MLS1 malate synthase 2: DAL7

TCA cycle

These pathways were same in the pathways generated by our pipeline (Hole-filled) and YeastCyc. The additional enzyme is EC 6.4.1.1 that link pyruvate and TCA cycle. The detail comparison information was on Table 4.13.

Table 4.13 The comparison TCA cycle between the pathways generated by our pipeline (Hole-filled) and YeastCyc

Hole-filled			**YeastCyc**		
Evidence Glyph	**Enzymes**	**Genes**	**Evidence Glyph**	**Enzymes**	**Genes**
	EC# 4.2.1.2	Fumarate hydratase, mitochondrial: FUM1		EC# 6.4.1.1	pyruvate carboxylase: PYC1 pyruvate carboxylase: PYC2
	EC# 1.1.1.37	Malate dehydrogenase, mitochondrial: MDH1 Malate dehydrogenase, peroxisomal: MDH3 Malate dehydrogenase, cytoplasmic: MDH2		EC# 2.3.3.1	citrate synthase: CIT3 citrate synthase: CIT1 citrate synthase: CIT2
	EC# 2.3.3.1	Citrate synthase, mitochondrial: CIT1		EC# 4.2.1.3	aconitase: ACO1 aconitate hydratase: ACO2

	Citrate synthase, peroxisomal: CIT2 Citrate synthase 3: CIT3		
EC# 4.2.1.3	Probable aconitate hydratase 2: ACO2 Aconitate hydratase, mitochondrial: ACO1 Aconitate hydratase, mitochondrial: ACO1	EC# 4.2.1.3	aconitase: ACO1
EC# 4.2.1.3	Probable aconitate hydratase 2: ACO2 Aconitate hydratase, mitochondrial: ACO1 Aconitate hydratase, mitochondrial: ACO1	EC# 1.1.1.41	NAD-dependent isocitrate dehydrogenase: IDH2, IDH1
EC# 1.1.1.41	Isocitrate dehydrogenase [NAD] subunit 2, mitochondrial: IDH2 Isocitrate dehydrogenase [NAD] subunit 1, mitochondrial: IDH1	α-ketoglutarate oxidative decarboxylation	2-ketoglutarate dehydrogenase complex: KGD2, KGD1, LPD1
α-ketoglutarate oxidative decarboxylation	2-oxoglutarate dehydrogenase, mitochondrial: KGD1	EC# 6.2.1.5	succinyl-CoA ligase: LSC2, LSC1
EC# 6.2.1.5	Succinyl-CoA ligase [ADP-forming] subunit alpha, mitochondrial: LSC1 Succinyl-CoA ligase [ADP-forming] subunit beta, mitochondrial: LSC2	EC# 1.3.5.1	minor succinate dehydrogenase (ubiquinone): SDH1b, SDH2, SDH3, SDH4 succinate dehydrogenase (ubiquinone): SDH1, SDH2, SDH3, SDH4
EC# 1.3.5.1	Succinate dehydrogenase [ubiquinone] iron-sulfur subunit, mitochondrial: SDH2 Succinate dehydrogenase	EC# 4.2.1.2	fumarate hydralase: FUM1

	[ubiquinone] flavoprotein subunit, mitochondrial: SDH1 Succinate dehydrogenase [ubiquinone] flavoprotein subunit 2, mitochondrial: YJL045W		
		EC# 1.1.1.37	peroxisome malate dehydrogenase: MDH3 mitochondrial malate dehydrogenase: MDH1 cytosolic malate dehydrogenase: MDH2

Fatty acid oxidation

There is no the reaction catalyzed by EC 5.3.3.8 and the pathway does not link to a cycle (Figure 4.5 and 4.6) in the pathways generated by our pipeline (Hole-filled). The detail comparison information was on Table 4.14.

Table 4.14 The comparison fatty acid oxidation between the pathways generated by our pipeline (Hole-filled) and YeastCyc

Hole-filled			YeastCyc		
Evidence Glyph	**Enzymes**	**Genes**	**Evidence Glyph**	**Enzymes**	**Genes**
	EC# 6.2.1.3	Long-chain-fatty-acid--CoA ligase 3: FAA3 Long-chain-fatty-acid--CoA ligase 1: FAA1 Long-chain-fatty-acid--CoA ligase 2: FAA2 Long-chain-fatty-acid--CoA ligase 4: FAA4		EC# 5.3.3.8	delta(3,5)-delta(2,4)-dienoyl-CoA isomerase: DCI1 d3,d2-Enoyl-CoA Isomerase: ECI1
	EC# 1.3.3.6	Acyl-coenzyme A oxidase: POX1		EC# 6.2.1.3	long chain fatty acyl:CoA synthetase: FAA1 long chain fatty acyl:CoA synthetase: FAA4 acyl-CoA synthase: FAA3 acyl-CoA synthetase: FAA2 fatty acid transporter: FAT1
	EC# 4.2.1.17	None		EC# 1.3.3.6	fatty-acyl coenzyme A oxidase: POX1

EC# 1.1.1.35	None	EC# 4.2.1.17	3-hydroxyacyl-CoA dehydrogenase: FOX2
EC# 2.3.1.16	3-ketoacyl-CoA thiolase, peroxisomal: FOX3	EC# 1.1.1.35	3-hydroxyacyl-CoA dehydrogenase: FOX2
		EC# 2.3.1.16	3-oxoacyl CoA thiolase: POT1

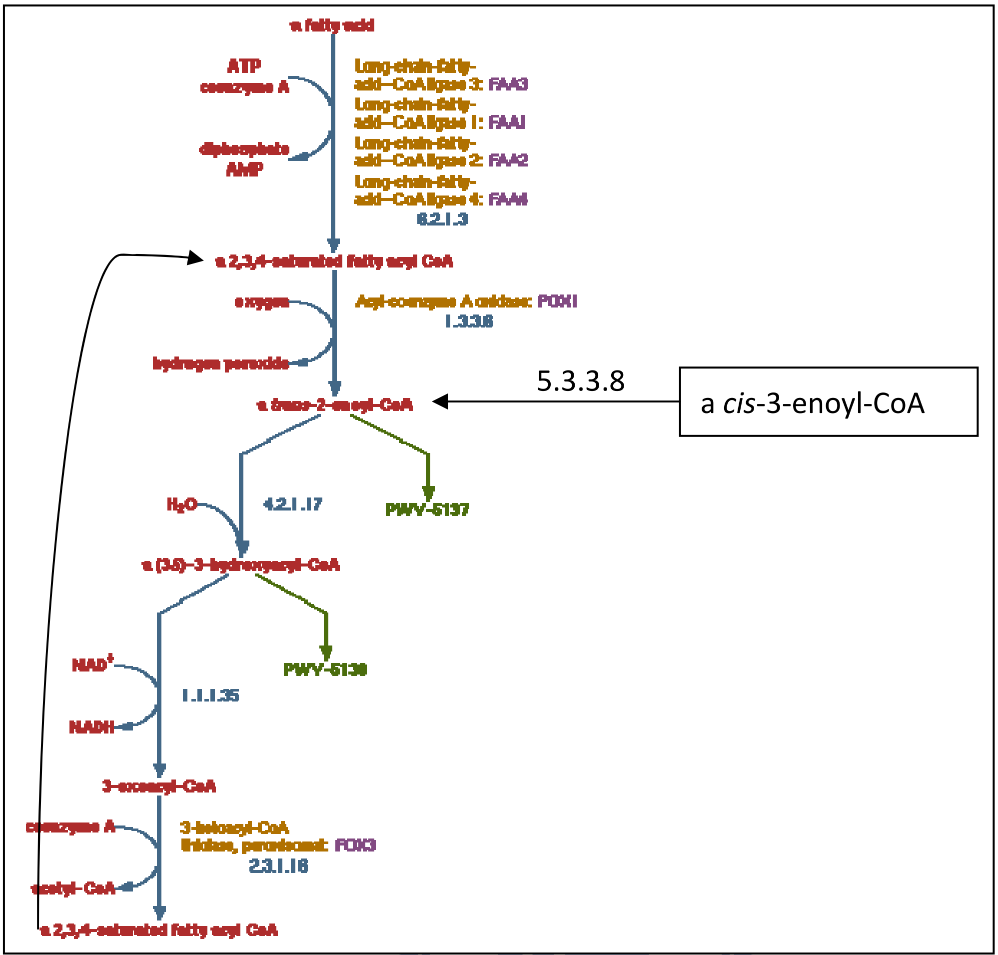


Figure 4.5 Fatty acid oxidation in our pathways

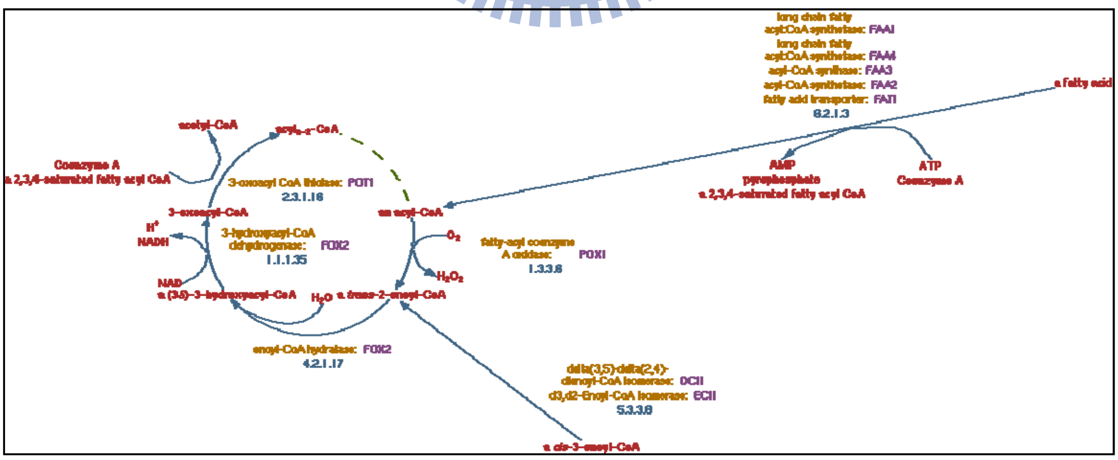


Figure 4.6 Fatty acid oxidation in YeastCyc

Chapter 5 Discussion

We implement an integrated pipeline for sequence assembly, gene annotation and metabolic pathway reconstruction. Our pipeline had been tested by three types of dataset including the sequencing results of Roche 454, Illumina and hybrid dataset that combine the contigs generated by velvet assemble Illumina and Roche 454. Our results show that the assembly result of the hybrid dataset is better than solely dataset of Roche 454 or Illumina based on the value of N50. Our pipeline can successfully assemble the reads from the next generation sequencing techniques. Following the sequence assembly process is the gene annotations. Our pipeline integrates four ab initio gene prediction tools and one evidence-based gene prediction tool. The ab initio tools include GlimmerHMM, AUGUSUT, GENSCAN, and SNAP. The evidence-based prediction tool is BLAST. Then, our pipeline use EvidenceModeler to combine the results from all of the gene prediction tools. The gene annotation simulation of yeast genome shows 88% of genes in yeast are well annotated. Our hybrid gene prediction approach can annotate more genes than the number of genes predicted by BLAST (86%) or by single ab inito gene prediction tools, which the average recall ratio is 81%. In short, the EvidenceModeler can significant increase the gene recall rate. Our platform can also reconstruct the metabolic pathways of the predicted genes, which belong to the tier 3 databases in Biocyc. We compare some of the housekeeping pathways between our annotation and annotated in YeastCyc which is a manual curated metabolic pathway database in Yeast. We found that the pathways are almost identical. Overall, our platform can assembly and annotate the genome sequenced by the next generation sequencing techniques in short time and provide the data of genomic sequence, genes and metabolic pathways.

We found that around 88% of yeast genes were predicted by EvidenceModeler which is less than the number of genes predicted by GlimmerHMM (90%). The reason of the drawback is that the genes predicted by EvidenceModeler is the gene prediction combination from various tools. Due to some of the genes can only be annotated by GilmmerHMM, the EvidenceModeler cannot agree the prediction results from single prediction tool. In other words, the gene predicted by EvidenceModeler must be predicted by most of the prediction tools.

The manually curated metabolic pathways are usually different to the pathways predicted by computational approaches. For example, the number of metabolic pathways which are computationally annotated in CattleCyc is 243. After the manual curation, the number of pathways shared with the computational approached is 113[38]. We have similar problem of our metabolic pathway reconstruction. YeastCyc include 133 pathways and our platform predicted 204 pathways. Around 55 pathways are shared between YeastCyc annotations and our pathway annotations. Hence the computational annotated metabolic pathways must be curated manually.

Chapter 6 Future work

The sequence assembly quality might be improved if we can design a filter to control the quality of the reads. The filter removes the reads according to the region of low complexity region and the error rate of read.

Currently, the parameters for sequence assembly are manually adjusted. The best parameters are different for various organisms and experiments. Hence, finding the best parameters for each new experiment manually is labor intensive. An automatic process to figure out the best parameter for each annotation might be reduced the time of finding parameters.

During the evaluation of our gene prediction, we found the accuracy of the gene prediction is not higher enough to recover as more genes as possible. In order to improve drawbacks, we may include more evidence-based gene prediction data such as cDNA and EST sequence.

A well integrated graphical user interface could improve the usability of our annotation platforms because the gene annotation results and the metabolic pathways are shown in distinct web sites. It is hard for user to find the annotation linkage between different annotations. If we can provide a user interface such as UCSC genome browser, the user can access the annotation more convenient.

References

1. Miller, J.R., S. Koren, and G. Sutton, *Assembly algorithms for next-generation sequencing data.* Genomics, 2010. **95**(6): p. 315-27.
2. Imelfort, M. and D. Edwards, *De novo sequencing of plant genomes using second-generation technologies.* Brief Bioinform, 2009. **10**(6): p. 609-18.
3. Voelkerding, K.V., S.A. Dames, and J.D. Durtschi, *Next-generation sequencing: from basic research to diagnostics.* Clin Chem, 2009. **55**(4): p. 641-58.
4. Holt, R.A. and S.J. Jones, *The new paradigm of flow cell sequencing.* Genome Res, 2008. **18**(6): p. 839-46.
5. Huang, X. and A. Madan, *CAP3: A DNA sequence assembly program.* Genome Res, 1999. **9**(9): p. 868-77.
6. Huang, X., et al., *PCAP: a whole-genome assembly program.* Genome Res, 2003. **13**(9): p. 2164-70.
7. Batzoglou, S., et al., *ARACHNE: a whole-genome shotgun assembler.* Genome Res, 2002. **12**(1): p. 177-89.
8. Green, P. *Phrap documentation.* 1996; Available from: http://www.phrap.org/phredphrap/phrap.html.
9. Myers, E.W., et al., *A whole-genome assembly of Drosophila.* Science, 2000. **287**(5461): p. 2196-204.
10. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
11. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.
12. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* Genome Res, 2009. **19**(6): p. 1117-23.
13. Butler, J., et al., *ALLPATHS: de novo assembly of whole-genome shotgun microreads.* Genome Res, 2008. **18**(5): p. 810-20.
14. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Res, 2010. **20**(2): p. 265-72.
15. Stanke, M. and S. Waack, *Gene prediction with a hidden Markov model and a new intron submodel.* Bioinformatics, 2003. **19 Suppl 2**: p. ii215-25.
16. Salamov, A.A. and V.V. Solovyev, *Ab initio gene finding in Drosophila genomic DNA.* Genome Res, 2000. **10**(4): p. 516-22.
17. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.
18. Lukashin, A.V. and M. Borodovsky, *GeneMark.hmm: new solutions for gene*

finding. Nucleic Acids Res, 1998. **26**(4): p. 1107-15.
19. Majoros, W.H., M. Pertea, and S.L. Salzberg, *TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.* Bioinformatics, 2004. **20**(16): p. 2878-9.
20. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
21. Romero, P., et al., *Computational prediction of human metabolic pathways from the complete human genome.* Genome Biol, 2005. **6**(1): p. R2.
22. Warren, R.L., et al., *Assembling millions of short DNA sequences using SSAKE.* Bioinformatics, 2007. **23**(4): p. 500-1.
23. Korf, I., *Gene finding in novel genomes.* BMC Bioinformatics, 2004. **5**: p. 59.
24. Karp, P.D., et al., *Expansion of the BioCyc collection of pathway/genome databases to 160 genomes.* Nucleic Acids Res, 2005. **33**(19): p. 6083-9.
25. Caspi, R., et al., *The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2008. **36**(Database issue): p. D623-31.
26. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
27. Karp, P.D., S. Paley, and P. Romero, *The Pathway Tools software.* Bioinformatics, 2002. **18 Suppl 1**: p. S225-32.
28. Haas, B.J., et al., *Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.* Genome Biol, 2008. **9**(1): p. R7.
29. Stein, L.D., et al., *The generic genome browser: a building block for a model organism system database.* Genome Res, 2002. **12**(10): p. 1599-610.
30. Bairoch, A. and B. Boeckmann, *The SWISS-PROT protein sequence data bank.* Nucleic Acids Res, 1991. **19 Suppl**: p. 2247-9.
31. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2010. **38**(Database issue): p. D46-51.
32. A.F.A. Smit, R.H.P.G. *RepeatMasker*. Available from: http://repeatmasker.org.
33. Karp, P.D., et al., *The MetaCyc Database.* Nucleic Acids Res, 2002. **30**(1): p. 59-61.
34. Green, M.L. and P.D. Karp, *A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.* BMC Bioinformatics, 2004. **5**: p. 76.
35. Bairoch, A., *The ENZYME database in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 304-5.
36. Wu, C.H., et al., *The Protein Information Resource.* Nucleic Acids Res, 2003.

31(1): p. 345-7.
37. *YeastCyc Database*. Available from: http://pathway.yeastgenome.org/.
38. Seo, S. and H.A. Lewin, *Reconstruction of metabolic pathways for the cattle genome.* BMC Syst Biol, 2009. **3**: p. 33.