SeqZip – a tool for reconstruction of genome sequences using Solexa/Illumina machine data

Victor Solovyev[1], Denis Vorobyev[2], Peter Kosarev[2], Igor Seledtsov[2]

[1]Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX,UK;
[2]Softberry Inc., 116 Radio Circle, Suite 400, Mount Kisco, NY 10549, USA

ABSTRACT

SeqZip *tool has been developed recently by Softberry research team (www.softberry.com).*
By processing millions of short reads generated by Solexa sequencing machine it provides effective solutions to the following three tasks:

1) ab initio reconstruction of genomic sequence;
2) reconstruction of sequence using a reference genome from the same or close organism;
3) mutation profiling and SNP discovery in a given set of genes.

The SeqZip tool is using L-plets hashing technique enable fast data processing and taking into account the reads quality information.

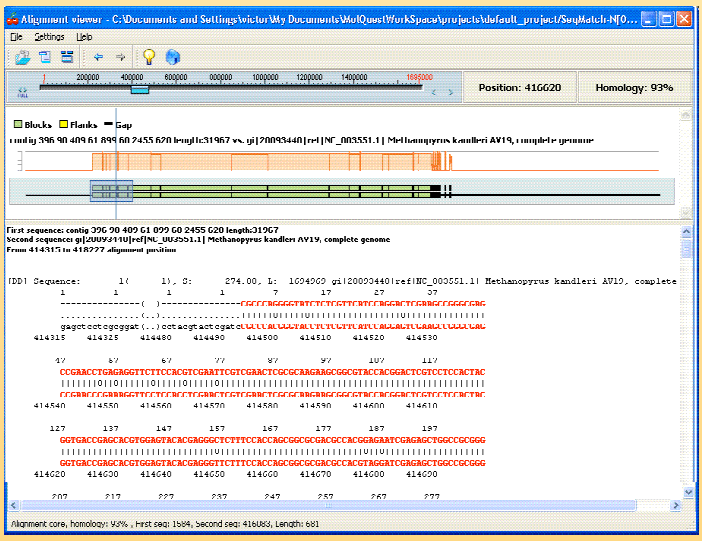
I. Ab initio sequence assembling:

Our ab initio sequence reconstruction was tested on assembling a few phage and bacterial genomes and demonstrated superior clustering power comparing with the earlier published approach (Bioinformatics, 2007, 23(4):500-501): simulated error-free 25mers of bacteriophage PhiX174 and coronavirus SARS TOR2 were assembled perfectly, while we produced approximately twice longer contigs for bacteria Haemophilus influenzae for the same level of genome coverage.

II. Sequence assembling using reference genome:

To reconstruct sequences using a reference genome we applied SeqZip to assembling seven bacterial genomic sequences related to the known genome of Methanopyrus kandleri AV19 from 5-7 millions reads (for each genome) produced by the sequencing laboratory of Harvard Partners HealthCare Center for Genetics and Genomics. The AV19 genome itself was assembled perfectly. Also from the set of AV19 oligonucleotides there is one additional contig was produced that happened to be completely identical to the whole genome sequence of phi-X174 phage. The other related genome TAG11 was reconstructed (using AV19 as reference genome) resulting in a few hundred contigs each.

In the following figure we can see *(with Softberry sequence comparison viewer)* the alignments of one contig of TAG11 genome reconstructed from short Solexa reads with part of AV19 genome.



Statistics of TAG11 assembling using AV19 reference genome:

Length >	# contigs	Genome coverage
75	325	98.37
700	221	97.24
1000	210	96.68

A few new software modules of SeqZip tool are currently under development for resolving some difficulties to treat repeated sequences that will result in producing longer contigs. SeqZip tolls will be incorparated to Softberry WEB server as well as in a new version of Softberry Molquest 2 (www.molquest.com) bioinformatics package for Windows.

Annotation of assembled genomes:

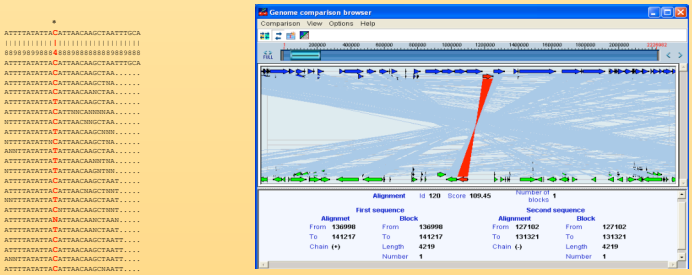
Using sequences of aligned parts of both genomes we run **Fgenesb gene annotation pipeline** on both sequences. Two fragments of the annotation is presented below (The first is for TAG11 contig and the second is for AV19 sequence. We can see that pipeline predicted almost the same genes in both genomes (while they have small differences in their length).

```
Prediction of potential genes in microbial  genomes
 Time:    Tue Nov 13 12:41:03 2007
 Seq name: contig  of TAG11    length:31967
 Length of sequence - 31967 bp
 Number of predicted genes - 41, with homology - 34
 Number of transcription units - 16, operones - 9 average op.length -  3.8
     N      Tu/Op   Conserved  S              Start          End     Score
                        pairs(N/Pv)
     1     1 Op  1      .       -     CDS          3 -       1353     391  ## COG0144 tRNA and rRNA cytosine-C5-methylases
     2     1 Op  2      .       -     CDS       1310 -       1534     112  ##
                               +     Prom      1499 -       1558     2.6
     3     2 Tu  1    1/0.667   +     CDS       1622 -       2263     471  ## COG4034 Uncharacterized protein conserved in
     4     3 Tu  1      .       +     CDS       2397 -       3242     516  ## COG0157 Nicotinate-nucleotide pyrophosphoryl
     5     4 Op  1      .       -     CDS       3264 -       4277     567  ##
     6     4 Op  2      .       -     CDS       4217 -       4711     247  ## COG0028 Thiamine pyrophosphate-requiring enz
     7     4 Op  3      .       -     CDS       4728 -       5096     234  ##
     8     5 Tu  1      .       +     CDS       5218 -       5730     220  ## COG1813 Predicted transcription factor, homo
     9     6 Op  1      .       -     CDS       5734 -       6363     283  ##
    10     6 Op  2    1/0.667   -     CDS       6293 -       7843     740  ## COG0849 Actin-like ATPase involved in cell d
    11     6 Op  3    2/0.000   -     CDS       7923 -       8234     146  ## COG1694 Predicted pyrophosphatase
    12     6 Op  4      .       -     CDS       8231 -       8779     212  ## COG0500 SAM-dependent methyltransferases
                               +     Prom      9049 -       9108     2.8
    13     7 Op  1    2/0.000   +     CDS       9262 -       9609     267  ## COG4921 Uncharacterized protein conserved in
    14     7 Op  2      .       +     CDS       9614 -      10849     686  ## COG2262 GTPase
```

```
Prediction of potential genes in microbial  genomes
 Time:    Tue Nov 13 12:36:21 2007
 Seq name: gi|20093440|ref|NC_003551.1| Methanopyrus kandleri AV19, complete genome 414494 447500 length 1694969
 Length of sequence - 33007 bp
 Number of predicted genes - 44, with homology - 34
 Number of transcription units - 17, operones - 10 average op.length -  3.7
     N      Tu/Op   Conserved  S              Start          End     Score
                        pairs(N/Pv)
     1     1 Op  1      .       -     CDS          3 -       1248     606  ## COG0144 tRNA and rRNA cytosine-C5-methylases
     2     1 Op  2      .       -     CDS       1289 -       1540     146  ##
     3     2 Tu  1    1/0.667   +     CDS       1643 -       2269     431  ## COG4034 Uncharacterized protein conserved in
                               +     Prom      2282 -       2341     2.1
     4     3 Tu  1      .       +     CDS       2420 -       3268     493  ## COG0157 Nicotinate-nucleotide pyrophosphoryl
     5     4 Op  1      .       -     CDS       3286 -       4299     640  ##
     6     4 Op  2      .       -     CDS       4239 -       4676     272  ## COG0028 Thiamine pyrophosphate-requiring enz
     7     4 Op  3      .       -     CDS       4750 -       5118     238  ##
     8     5 Tu  1      .       +     CDS       5240 -       5755     212  ## COG1813 Predicted transcription factor, homo
     9     6 Op  1      .       -     CDS       5759 -       6391     347  ##
    10     6 Op  2    1/0.667   -     CDS       6321 -       7871     640  ## COG0849 Actin-like ATPase involved in cell d
    11     6 Op  3    2/0.000   -     CDS       7951 -       8262     149  ## COG1694 Predicted pyrophosphatase
    12     6 Op  4      .       -     CDS       8259 -       8807     234  ## COG0500 SAM-dependent methyltransferases
                               +     Prom      9077 -       9136     2.8
    13     7 Op  1    2/0.000   +     CDS       9290 -       9637     273  ## COG4921 Uncharacterized protein conserved in
    14     7 Op  2      .       +     CDS       9642 -      10877     690  ## COG2262 GTPases
```

III. SNP discovery

Solexa sequencing provide possibility of analyzing many individuals DNA during one run. As one example of using Solexa data for SNP finding we present a fragment of Homo sapiens eps8 binding protein gene. The known SNP **C-> T** substitution is marked here by *. It can be observed that approximately 40% of 268 mapped to this region oligonucleotides (Solexa reads) support occurrence of this SNP. This example demonstrates application of Solexa sequencing for SNP discovery and it can be used to find specific SNP in some selected populations (people with some type of disease, for example).



Example of Comparing of assembled bacterial genomes in Genome comparison Browser

Example of main interface of Molquest 2 including Pipeline construction tools