How long is long enough?

- Modeling to predict genome assembly performance -

Hayan Lee@Schatz Lab

Feb 26, 2014 Quantitative Biology Seminar



Outline

- Background
 - Assembly history
 - Recent sequencing technology + Algorithm
- Motivation
 - Lander-Waterman statistics
 - Economical meaning (ROI)
- Our approach
 - Genome assembly challenges
 - Support Vector Regression (SVR)
 - Feature engineering
 - Model fitting
 - Prediction : Genome assembly performance
- Contribution

Cold Spring Harbor Laboratory

Genome Assembly

Nature Vol. 265 February 24 1977

articles

Nucleotide sequence of bacteriophage $\Phi\,X174\,\,DNA$

F. Sanger, G. M. Air', B. G. Barrell, N. L. Brown', A. R. Coulson, J. C. Fiddes, C. A. Hutchison IIF, P. M. Slocombe⁶ & M. Smith' MRC Laboratory of Molecular Biology, Hills Read, Cambridge CR2 2014, UK

A DNA sequence for the genome of bacteriophage ΦXI74 of approximately 5.175 multivalues has been determined sequence identifying and the sequence of the sequence identifying the production of the proteins of the name screening the for the production of the protein of the name screening the form organism, including initiation and termination states for the protein and RNAs. Two pairs of genes are colled by the protein and RNAs the optic reading frames.

This genome of bacteriophage ΦX174 is a single-strainfed circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques¹⁻¹ is A = A = C - D = J = F = A. Genes, F, G and H code for structural proteins of the virus capsid, and gene Usa defined be sequence work/ order for a small basic notein strand DNA of ΦN have been used sequence as the mRNA and, in ordinary, containing, will have physicases to a protocol and a containing will have physicases to a strain a sequence of the sequence of the sequence of the sequence sequence of the sequence physicase of the sequence of t

1977. Sanger *et al.* 1st Complete Organism X5375 bp



1995. Fleischmann *et al.* 1st Free Living Organism TIGR Assembler. 1.8Mbp



1998. C.elegans SC Ist Multicellular Organism BAC-by-BAC Phrap. 97Mbp



2000. Myers *et al.* 1st Large WGS Assembly. Celera Assembler. 116 Mbp



2001.Venter *et al.*, IHGSC Human Genome Celera Assembler/GigaAssembler. 2.9 Gbp



2010. Li *et al.* 1st Large SGS Assembly. SOAPdenovo 2.2 Gbp

Assembling a Genome

I. Shear & Sequence DNA



2. Construct assembly graph from overlapping reads

...AGCCTAG<mark>GGATGCGCGACACG</mark>T

GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph



4. Detangle graph with long reads, mates, and other links



Assembly Complexity







Long Read Sequencing Technology



PacBio SMRT Sequencing Road Map

October, 2013 **P5-C3** 9000 The new P5 polymerase and C3 8,500 bp chemistry combination (P5-C3) 8000 extends the industry-leading sequence read lengths to an 7000 average of approximately 8,500 6000 bases, with the longest reads exceeding 30,000 bases, doubling P4-C2 5000 throughput 4000 C2-C2 3000 ECR2 Early 2000 chemistries FCR 1000 1012 LPR 453 1734 0 2008 2009 2010 2011 2012 2013

Read Length (bp)

CSH Cold Spring Harbor Laboratory

S. cerevisiae W303

PacBio RS II sequencing at CSHL by Dick McCombie •Size selection using an 7 Kb elution window on a BluePippin[™] device from Sage Science



PacBio Assembly Algorithms

PacBioToCA

PBJelly

Gap Filling and Assembly Upgrade

English et al (2012) PLOS One. 7(11): e47768



Hybrid/PB-only Error Correction

Koren, Schatz, et al (2012) Nature Biotechnology. 30:693–700



PB-only Correction & Polishing

Chin et al (2013) Nature Methods. 10:563–569

< 5x

PacBio Coverage

> 50x

ogy

Many Genomes Are Sequenced... Many Questions Are Raised... But...

- How long should the read length be?
- What coverage should be used?

Given the read length and coverage,

- How long are the contigs?
- How many contigs?
- How many reads are in each contigs?
- How big are the gaps?

old Spring Harbor Laboratory

Previous Works

GENOMICS 2, 231-239 (1988)

Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis

ERIC S. LANDER* † AND MICHAEL S. WATERMAN‡

*Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142; †Harvard University, Cambridge, Massachusetts 02138; and ‡Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089

Received January 13, 1988; revised March 31, 1988

Results from physical mapping projects have recently been reported for the genomes of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, and similar projects are currently being planned for other organisms. In such projects, the physical map is assembled by first "fingerprinting" a large number of clones chosen at random from a recombinant library and then inferring overlaps between clones with sufficiently similar fingerprints. available region of up to several megabases and of studying its properties. In addition, the overlapping clones comprising the physical map would constitute the logical substrate for efforts to sequence an organism's genome.

Recently, three pioneering efforts have investigated the feasibility of assembling physical maps by means of "fingerprinting" randomly chosen clones. The fingerprints consisted of information about restriction

Lander-Waterman Statistics

G : Genome size L : read Length N : Number of reads C : Coverage = $\frac{NL}{G}$

Note : Poisson distribution is assumed! $P(x; \lambda) = \frac{\lambda^{x} e^{-\lambda}}{x!}$

P(No reads start in a given position) = $\frac{C^0 e^{-C}}{0!} = e^{-C}$ P(At least 1 read starts in a given position) = $1 - e^{-C}$

H Cold Spring Harbor Laboratory

Lander-Waterman Statistics

P(No reads start in a given position) = $\frac{C^0 e^{-C}}{0!} = e^{-C}$

P(At least 1 reads starts in a given position) = $1 - e^{-C}$

Expected # of bases in Gaps = $e^{-C}G$ Expected # of bases in Contigs = $(1 - e^{-C})G$ Expected # of contigs = $e^{-C}N$

Mean of contig size = $\frac{expected \# of bases in contigs}{expected \# of contigs} = \frac{(e^{C}-1)L}{C}$ (derivation) $\frac{(1-e^{-C})G}{e^{-C}N} = \frac{(e^{C}-1)G}{N} = \frac{(e^{C}-1)L}{C}$ Mean of contig size = $\frac{expeated \# of bases in contigs}{expected \# of contigs} = \frac{(e^{(1-\theta)C}-1)L}{C}$

SH Cold Spring Harbor Laboratory

HG19 Genome Assembly Performance by Lander-Waterman Statistics



Two key observations1. Contig over genome size2. Read Length vs. Coverage

Technology vs. Money

Empirical Data-driven Approach

- We selected 26 species across tree of life and exhaustively analyzed their assemblies using simulated reads for 4 different length (6 for HG19) and 4 different coverage per species
- For the extra long reads, we fixed the Celera Assembler(CA) to support reads up to 0.5Mbp



26 Species Across Tree of Life

Model	ID	Genome Size	# of	ploidy	Kingdom	Phylum	Class	Orde
Organism			chromosomes		(Domain)			
M.jannaschii	1	1,664,970	1	1	Archaea	Euryarchaeota	Methanococci	Meth
C.hydrogenoformans	2	2,401,520	1	1	Bacteria	Firmicutes	Clostridia	Clost
E.coli	3	4,639,675	1	1	Eubacteria	Proteobacteria	Gammaproteobacteria	Enter
Y.pestis	- 4	4,653,728	4	1	Eubacteria	Proteobacteria	Gammaproteobacteria	Enter
B.anthracis	5	5,227,293	1	1	Bacteria	Firmicutes	Bacilli	Bacil
A.mirum	6	8,248,144	1	1	Bacteria	Actinobacteria	Actinobacteria	Actin
yeast	7	12,157,105	16	1	Fungi	Ascomycota	Saccharomycetes	Sacc
Y.lipolytica	8	20,502,981	6	1	Fungi	Ascomycota	Saccharomycetes	Sacc
slime mold	9	34,338,145	6	1	Amoebozoa	Mycetozoa	Dictyostelia	Dicty
Red bread mold	10	41,037,538	7	1	Fungi	Ascomycota	Pezizomycotina	Sord
sea squirt	11	78,296,155	14	2	Animalia	Chordata	Ascidiacea	Enter
roundworm	12	100,272,276	6	2	Animalia	Nematoda	Chromadorea	Rhat
green alga	13	112,305,447	17	1	Plantae	Chlorophyta	Chlorophyceae	Chlar
arabidopsis	14	119,667,750	5+C+Mt	2	Plantae	Angiosperms	Magnoliopsida	Bras
fruitfly	15	130,450,100	3+XU+Mt	2	Animalla	Arthropoda	Insecta	Dipte
peach	16	227,252,106	8	2	Plantae	Angiosperms	Magnoliopsida	Rosa
rice	17	370,792,118	12	2	Plantae	Angiosperms	Monocots	Poale
poplar	18	417,640,243	19	2	Plantae	Angiosperms	Eudicots	Malp
tomato	19	781,666,411	12	2	Plantae	Magnoliophyta	Magnoliopsida	Solar
soybean	20	973,344,380	20	2	Plantae	Angiosperms	Eudicots	Faba
turkey	21	1,061,998,909	30+WZ+Mt	2	Animalla	Chordata	Aves	Gallif
zebra fish	22	1,412,464,843	25+MT	2	Animalla	Chordata	Actinopterygii	Cypr
lizard	23	1,799,126,364	6+abcdfgh	2	Animalla	Chordata	Reptilia	Squa
corn	24	2,066,432,718	10+Mt+Pt	2	Plantae	Angiosperms	Commelinids	Poale
mouse	25	2,654,895,218	19+XY	2	Animalla	Chordata	Mammalia	Rode
human	26	3,095,693,983	22+XY+Mt	2	Animalla	Chordata	Mammalia	Prim/

Stony Brook University

Dept. of Computer Science

HG19 Genome Assembly Performance by Our Simulation



H Cold Spring Harbor Laboratory

Why?

Lander-Waterman Statistics

• Assumptions!!!

Cold Spring Harbor Laboratory

• If genome is a random sequence, it will work

Our Approach

- Stop assuming that we cannot guarantee!!!
- We tried to assume as least as possible.
- Instead of building on top of assumptions, we let the model learn from the data
- Empirical data-driven approach

Cold Spring H

CSH

Repeats



a contractive Biology

CSH

Repeats in Rice



ititative Biology

Our Goal



Assembly Challenge

- Read Length
- Coverage
- Repeats
- Genome Size



Assembly Challenge (1) Read Length

- Read length is very important
- A matter of technology
- The longer is the better
- Quality was important but can be corrected
 - PacBio produces long reads, but low quality (~15% error rate)
 - Error correction pipeline are developed
 - Errors are corrected very accurately up to 99%

Cold Spring Harbor Laboratory

- Assembly Challenge (1) - Read Length

ZebraFish Assembly by Read Length



SH Cold Spring Harbor Laboratory

Assembly Challenge (2) Coverage

- A matter of money
- Using perfect reads, assembly performance increased for most genomes : Lower bound
- Using real reads, overall performance line will shift to the higher coverage
- The higher is the better (?)
- But still it suggests that there would be a threshold that can maximize your return on investment (ROI)

Assembly Challenge (2) Coverage



26

A. thaliana Ler-0

http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html



Genome size: Chromosome N50: Corrected coverage: 124.6 Mbp 23.0 Mbp 20x over 10kb A. thaliana Ler-0 sequenced at PacBio

•Sequenced using the previous P4 enzyme and C2 chemistry

•Size selection using an 8 Kb to 50 Kb elution window on a BluePippin[™] device from Sage Science

•Total coverage >119x

Sum of Contig Lengths:	149.5Mb
N50 Contig Length:	8.4 Mb
Number of Contigs:	1788

High quality assembly of chromosome arms Assembly Performance: 8.4Mbp/23Mbp = 36% MiSeq assembly: 63kbp/23Mbp = .2%

SH》Cold Spring Harbor Laboratory



Assembly Challenge (2) Coverage



Simons Center for Quantitative Biology

28

Assembly Challenge (3) **Repeats**

- Genome is not a random sequence
- Repeat hurts genome assembly performance
- Isolating the impact of repeats is not trivial
- Quantifying repeat characteristics is not trivial as well
 - The longest repeat size
 - # of repeats > read length



Assembly Challenge (3) **Repeats** Arabidopsis vs. Fruit fly

	Arabidopsis (120M) Longest repeat: 44kbp	Fruit fly (130M) Longest repeat: 30kbp
Mean Read Length	# of repeats > read length	<pre># of repeats > read length</pre>
3,650	210	5564
7,400	112	394
15,000	44	8
30,000	14	2

8 CSH

Assembly Challenge (3) **Repeats**



Dept. of Computer Science

Longest Repeat Size and Genome Size



Assembly Challenge (4) Genome Size

- Increase the assembly complexity
- Make a hard problem harder.



CS⊦

Assembly Challenge (4) Genome Size

S.cerevisiae Assembly by Coverage



5-

Dept. of Computer Science

S. cerevisiae W303

S288C Reference sequence

•12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler

•12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id







Assembly Challenge (4) Genome Size

Mouse Assembly by Coverage



H) Cold Spring Harbor Laboratory

Our Goal





Challenges for Prediction

- Sample size is small
- Quality is not guaranteed
- Predictive Power
- Overfitting

Support Vector Regression (SVR) Cross Validation



Support Vectors Non Support Vectors Regression Line Limits of 10 rule

Support Vector Regression (SVR)

• Epsilon insensitive loss function

$$L(y, f(x, w)) = \begin{cases} 0, & \text{if } |y - f(x, w)| \le \varepsilon \\ |y - f(x, w)| - \varepsilon, & \text{otherwise} \end{cases}$$

Support Vector Machine for Regression - Radial Basis Kernel



Benefits

- 1. Simplest fit
- **2.** Robust to outliers

(ex) Example of one-

regression function with

epsilon intensive band.

dimensional linear

Optimization(1)

• Minimize

$$\frac{1}{2} ||w||^2 + C \sum_{i=0}^n (\xi_i + \xi_i^*)$$

Subject to

$$y_i - (w_i x_{il}) - b \le \varepsilon + \xi_i$$

$$-y_i + (w_i x_{il}) + b \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$
 for $i = 1, 2, ..., l$



• Generalized Lagrange Multiplier (Karush–Kuhn–Tucker conditions)

CSH Cold Spring Harbor Laboratory

Optimization (2)

Primal

$$\min \frac{1}{2} ||w||^2 + C \sum_{i=0}^n (\xi_i + \xi_i^*) \qquad \text{s.t} \begin{cases} y_i - (w_i x_{il}) - b \le \varepsilon + \xi_i \\ -y_i + (w_i x_{il}) + b \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge for \ i = 1, 2, \dots, l \end{cases}$$

Karush–Kuhn–Tucker(KKT) conditions

Dual

$$\max \frac{1}{2} \sum_{i,j=0}^{l} (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) (\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=0}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=0}^{l} y_i (\alpha_i - \alpha_i^*)$$
s.t.
$$\begin{cases} \sum_{i=0}^{l} (\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases}$$
Kernel

CSH Cold Spring Harbor Laboratory

Optimization (3)

- Parameters
 - $\left\{\begin{array}{c} -C\\ -\varepsilon\end{array}\right\}$ SVM meta-parameters as user-defined inputs.
 - Kernel
 - Kernel type
 - Linear K(x_i, x_j)= $\langle x_i, x_j \rangle$
 - Non-Linear

usually based on applicationdomain knowledge and also should reflect distribution of input (x) values of the training data.

- Polynomial K(x_i, x_j)= $(x_i^T x_j + 1)^d$
- Radial Basis Function(RBF) K(x_i, x_j)=exp $\left(-\frac{1}{2\sigma^2}||x_i x_j||^2\right)$
- Kernel Parameters
 - Depends on kernel type

A Cold Spring Harbor Laboratory

SVR Model Fitting

- Using four features
 - Read Length
 - Coverage
 - Genome Size
 - # of Repeats > Read Length



Feature Engineering (1)

- Correlation Coefficient
 - Performance vs. genome size
 - R = -0.38
 - Performance vs. Read Length



Feature Engineering (2)

- Correlation Coefficient
 - Performance and *log* (genome size)
 - R = -0.49
 - Performance and *log* (read length)
 - R = 0.32
 - Performance and *log* (genome size)/ *log* (read length)
 - R = 0.6
 - Performance and log (coverage)
 - R = 0.58
 - Performance and log (# of repeats longer than read length)
 - R = -0.44

Cold Spring Harbor Laboratory

SVR Fit : Genome Assembly Using Genome Size and Read Length



How long is long enough?

SVR Fit: Genome Assembly Using Genome Size and Read Length



How long is long enough?

SVR Fit : Genome Assembly Using Genome Size and Read Length



How long is long enough?

Our Goal



How do we measure predictive power?
 How do we avoid overfitting?

Cross Validation

- K-fold Cross Validation
- A variation of Leave-One-Out Cross Validation (LOOCV)
- Leave one species out approach (LOSO) <- Our approach
 - A variation of Leave-One-Out Cross Validation (LOOCV)
 - Use 25 species as training data, test 1 species to measure predictive power
 - Avoid overfitting
- Model selection by predictive power





How long is long enough?

Prediction

- Average of residual is 10%
- We can predict the new genome assembly performance in 10% of error residual boundary
- Genome size, read length and coverage used explicitly
- Repeats are included implicitly



Web Service

Senome Assembly Performance Prediction - Mozilla Firefox							
<u>File Edit View History Bookmarks Tools H</u> elp							
Genome Assembly Performance Predicti +							
(
Most Visited 📴 etc 📴 CS 📴 erum							
Genome Assembly Performance Prediction							
This is the Genome Assembly Performance Prediction Service. If you have any queries please email Hayan Lee(<u>hlee@cshl.edu</u>).	Ξ						
Although assembly performance is a function of genome size, read length, coverage and repeats, in this prediction model, we only used 3 features; genome size, read length and coverage for the simplicity.							
Given genome size, we internally set read lengths and coverages for you. With 3 features, our model predicts the expected performance of assembly. Performance is defined as follows:							
Performance(%) = N50 of assembly / N50 of chromosomes							
Genome size : 1664000							
Submit							
Assembly Prediction of Genome Size 1664000							
By Coverage							
100	-						

How long is long enough?

Contribution

- Empirical data driven approach We selected 26 species across tree of life and exhaustively analyzed their
- For the extra long reads, we fixed the Celera Assembler(CA) to support reads up to 0.5Mbp
- We made a new model that predicts genome assembly performance
- Prediction in 10% of residue boundary.
- Our prediction is independent from any error model so that our model will provide upper bound and can be used for the general purpose
- The Marginal coverage forms around 20x-40x assuming no errors in reads
- Read Length is more important given that we have enough coverage

SH Cold Spring Harbor Laboratory

Contribution

• Recommendations

- < 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5 expect near perfect chromosome arms
- < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5 expect high quality assembly: contig N50 over 1Mbp
- > 1GB: hybrid/gap filling
 expect contig N50 to be 100kbp 1Mbp
- > 5GB: Email mschatz@cshl.edu

• Machine Learning & Big Data

Contact hlee@cshl.edu

Acknowledgements



CSHL Schatz Lab

Mike Schatz James Gurtowski Shoshana Marcus

BNL

Shinjae Yoo



McCombie Lab

Dick McCombie Eric Antoniou Elena Ghiban Melissa Kramer Panchajanya Deshpande Senem Mavruk Eskipehlivan Scott Ethe Sayers Sara Goodwin

A Cold Spring Harbor Laboratory

Thank You Q & A

