The Science of Information: From Communication to DNA Sequencing

David Tse
Dept. of EECS
U.C. Berkeley

Telecom Paris
August 31, 2012

Research supported by NSF Center for Science of Information.

Joint work with A. Motahari, G. Bresler, M. Bresler.

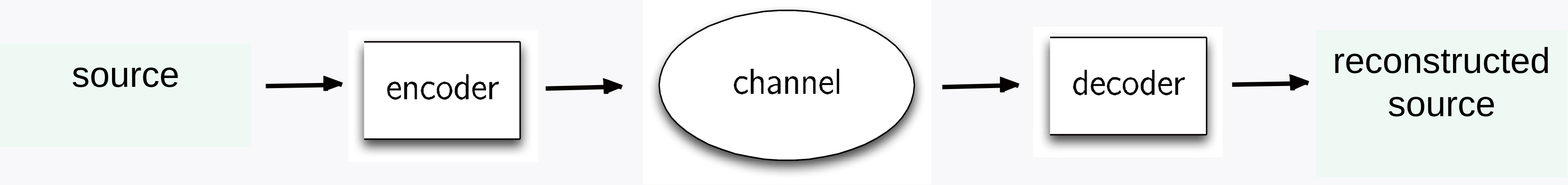
Acknowledgements: S. Batzolgou, L. Pachter, Y. Song

Communication: the beginning

- Prehistoric: smoke signals, drums.
- 1837: telegraph
- 1876: telephone
- 1897: radio
- 1927: television

Communication design tied to the specific source and specific physical medium.

Grand Unification



Model all sources and channels statistically.

channel capacity C bits/ sec

source entropy rate H bits/ source sym

Shannon 48

Theorem:

$$\boxed{\begin{array}{l} \text{max. rate of reliable communication} \\ = \dfrac{C}{H} \text{ source sym / sec.} \end{array}}$$

A unified way of looking at all communication problems in terms of information flow.

60 Years Later

- All communication systems are designed based on the principles of information theory.
- A benchmark for comparing different schemes and different channels.
- Suggests totally new ways of communication.

Secrets of Success

- Information , then computation.

  It took 60 years, but we got there.

- Simple models, then complex.

  The discrete memoryless channel
  ………… is like the Holy Roman Empire.

- Infinity, and then back.

  Allow us to think in terms of typical behavior.

“Asymptotic limit is the first term in the Taylor series expansion at infinity.

And theory is the first term in the Taylor series of practice.”

Tom Cover, 1990

Looking Forward

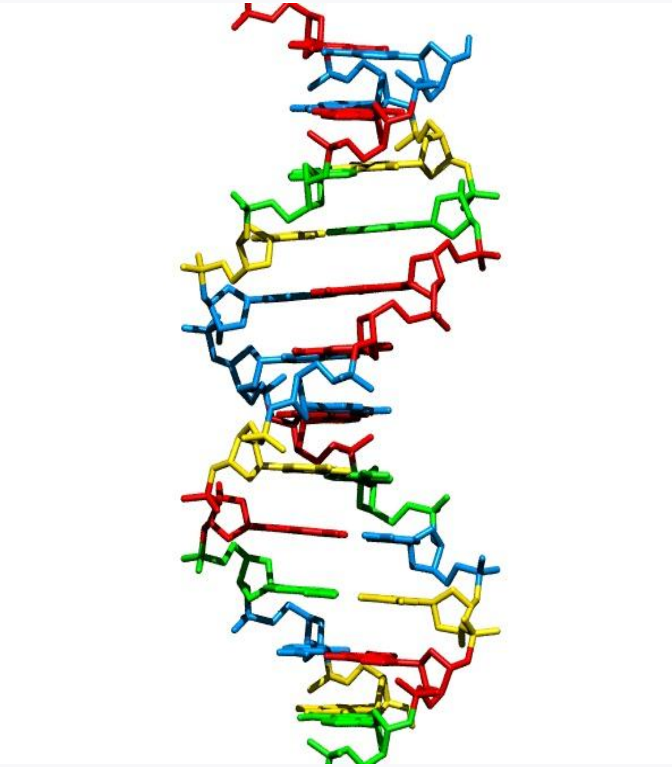
Can the success of this way of thinking be broadened to other fields?

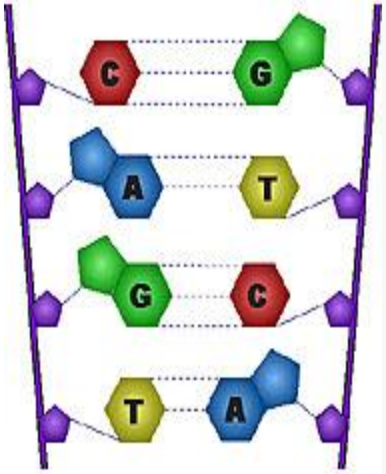
Information Theory of DNA Sequencing

DNA sequencing

DNA: the blueprint of life

Problem: to obtain the sequence of nucleotides.

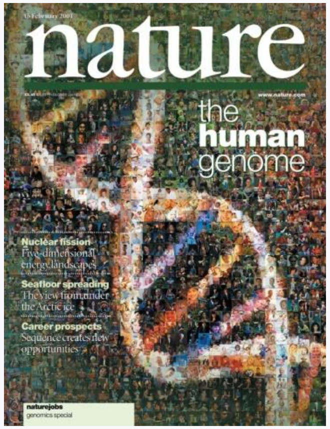


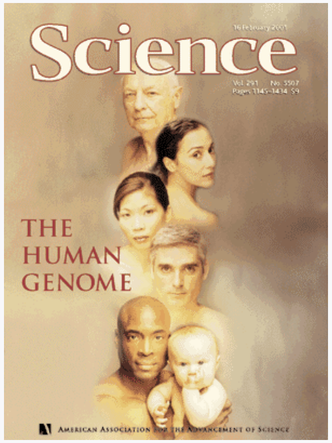


…ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT…

courtesy: Batzoglou

Impetus: Human Genome Project


nature
the
human
genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special



1990: Start

2001: Draft

2003: Finished

3 billion nucleotides

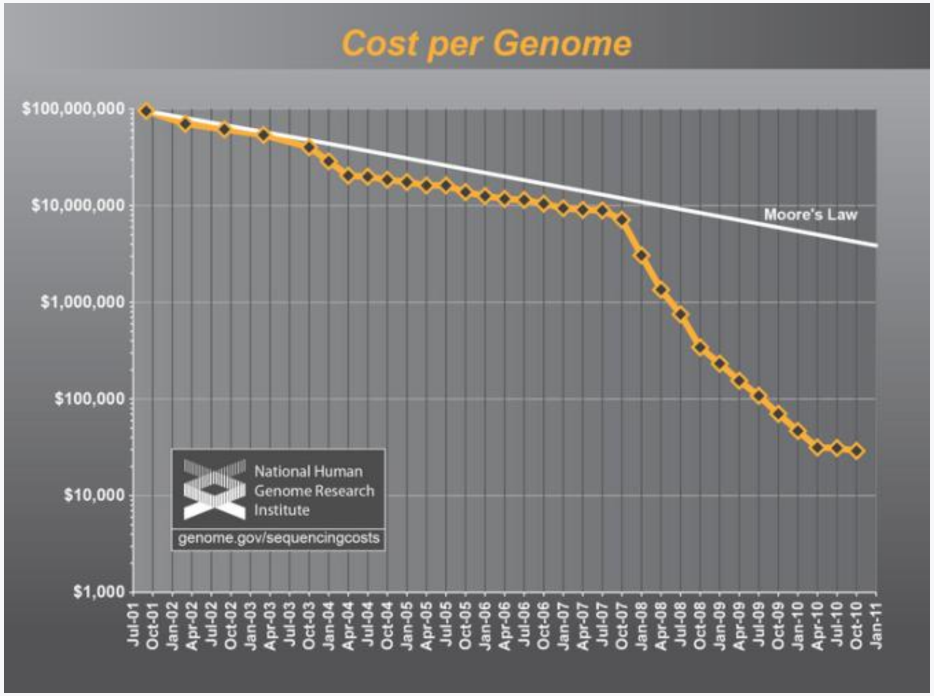
3 billion $$$$

courtesy: Batzoglou

Sequencing gets cheaper and faster

Cost of one human genome

HGP:	$ 3 billion
2004:	$30,000,000
2008:	$100,000
2010:	$10,000
2011:	**$4,000**
2012-13:	$1,000
???:	$300

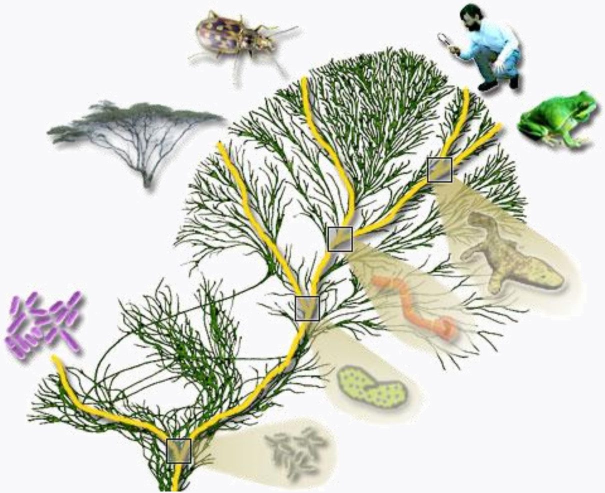


courtesy: Batzoglou

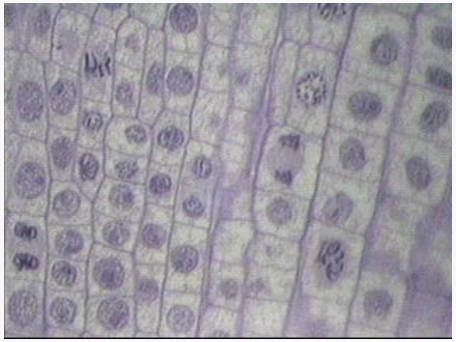
Time to sequence one genome: years → days

Massive parallelization.

But many genomes to sequence



100 million species
(e.g. phylogeny)

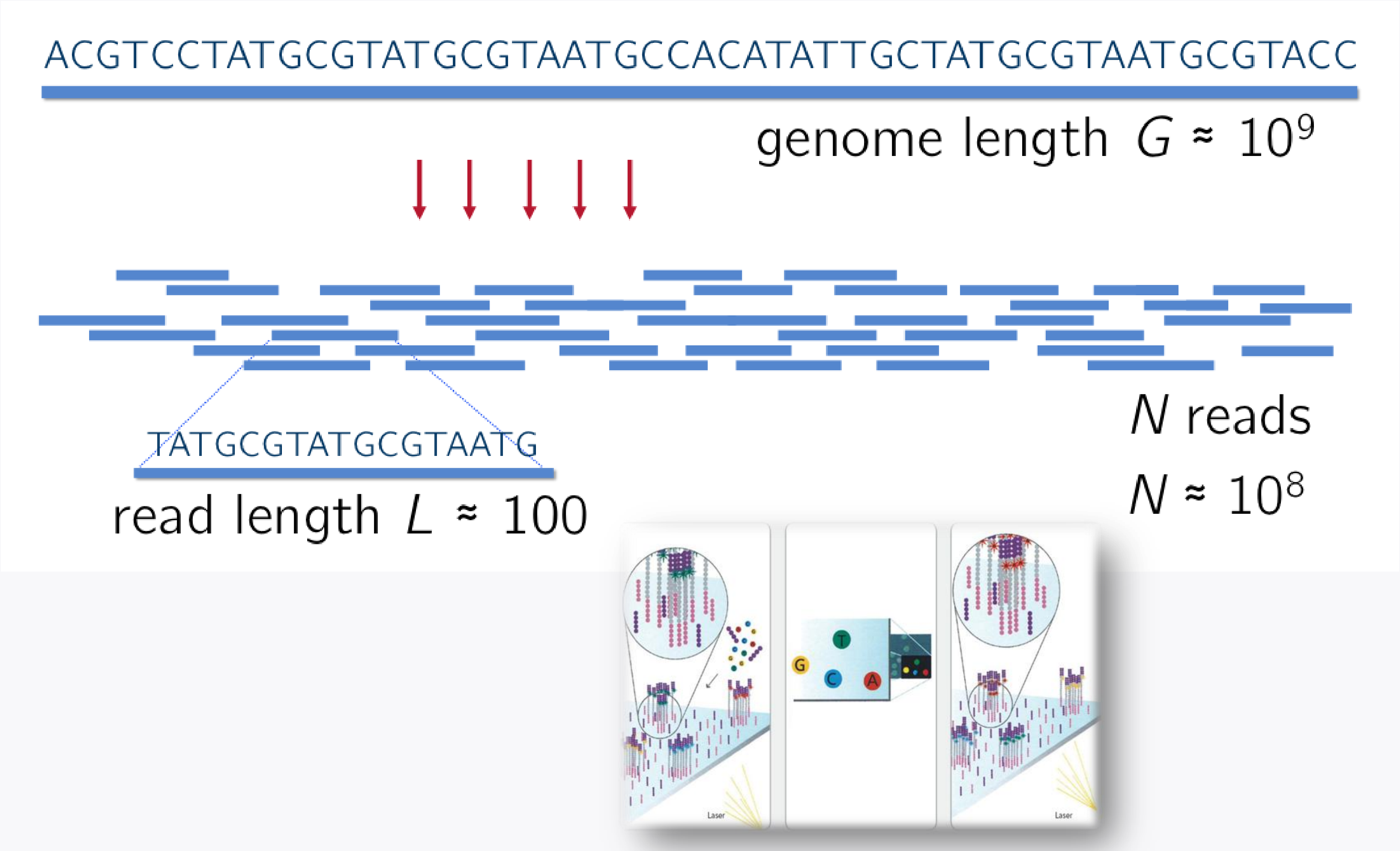


10^{13} cells in a human
(e.g. somatic mutations
such as HIV, cancer)

7 billion individuals
(SNP, personal genomics)

courtesy: Batzoglou

Whole Genome Shotgun Sequencing


ACGTCCTATGCGTATGCGTAATGCCACATATTGCTATGCGTAATGCGTACC
genome length $G \approx 10^9$
TATGCGTATGCGTAATG
read length $L \approx 100$
N reads
$N \approx 10^8$
G
T
C
A
Laser
Laser

Reads are assembled to reconstruct the original DNA sequence.

A Gigantic Jigsaw Puzzle

I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

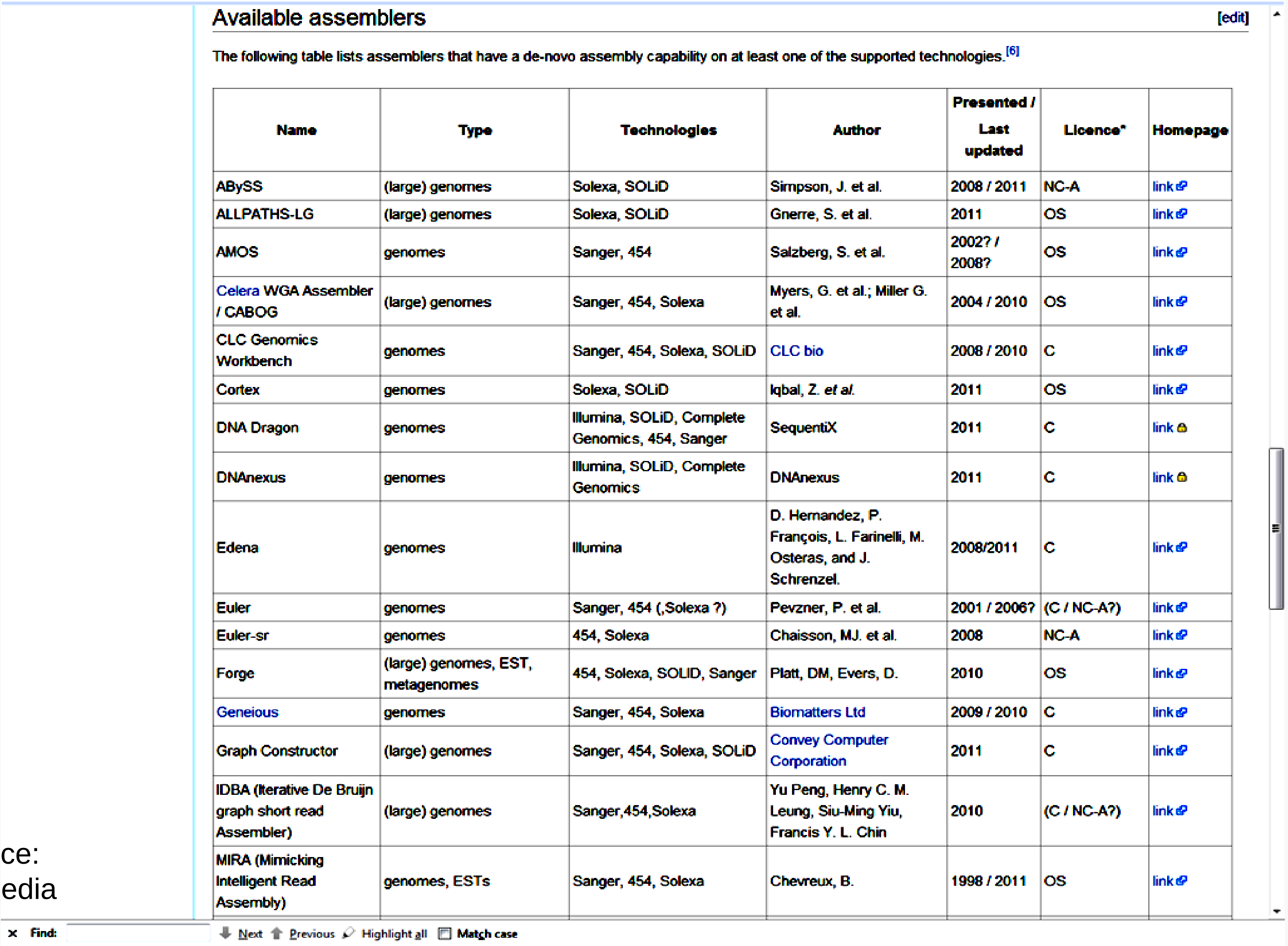
Many Sequencing Technologies

- HGP era: single technology (Sanger)

- Current: multiple “next generation” technologies (eg. Illumina, SoLiD, Pac Bio, Ion Torrent, etc.)

- Each technology has different read lengths, noise profiles, etc

Many assembly algorithms



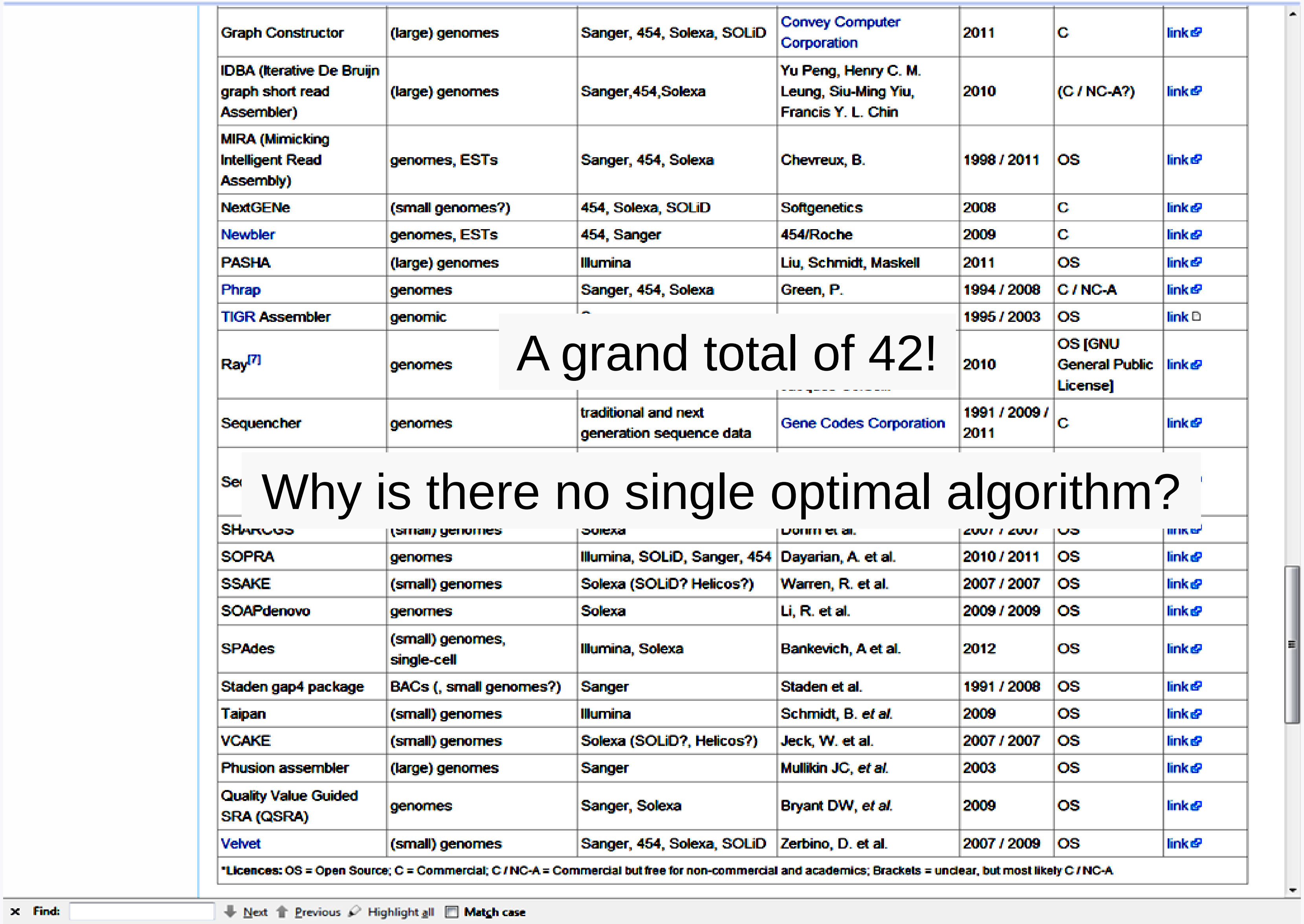
Available assemblers

The following table lists assemblers that have a de-novo assembly capability on at least one of the supported technologies.[6]

Name	Type	Technologies	Author	Presented / Last updated	Licence*	Homepage
ABySS	(large) genomes	Solexa, SOLiD	Simpson, J. et al.	2008 / 2011	NC-A	link
ALLPATHS-LG	(large) genomes	Solexa, SOLiD	Gnerre, S. et al.	2011	OS	link
AMOS	genomes	Sanger, 454	Salzberg, S. et al.	2002? / 2008?	OS	link
Celera WGA Assembler / CABOG	(large) genomes	Sanger, 454, Solexa	Myers, G. et al.; Miller G. et al.	2004 / 2010	OS	link
CLC Genomics Workbench	genomes	Sanger, 454, Solexa, SOLiD	CLC bio	2008 / 2010	C	link
Cortex	genomes	Solexa, SOLiD	Iqbal, Z. *et al.*	2011	OS	link
DNA Dragon	genomes	Illumina, SOLiD, Complete Genomics, 454, Sanger	SequentiX	2011	C	link 🔒
DNAnexus	genomes	Illumina, SOLiD, Complete Genomics	DNAnexus	2011	C	link 🔒
Edena	genomes	Illumina	D. Hernandez, P. François, L. Farinelli, M. Osteras, and J. Schrenzel.	2008/2011	C	link
Euler	genomes	Sanger, 454 (,Solexa ?)	Pevzner, P. et al.	2001 / 2006?	(C / NC-A?)	link
Euler-sr	genomes	454, Solexa	Chaisson, MJ. et al.	2008	NC-A	link
Forge	(large) genomes, EST, metagenomes	454, Solexa, SOLID, Sanger	Platt, DM, Evers, D.	2010	OS	link
Geneious	genomes	Sanger, 454, Solexa	Biomatters Ltd	2009 / 2010	C	link
Graph Constructor	(large) genomes	Sanger, 454, Solexa, SOLiD	Convey Computer Corporation	2011	C	link
IDBA (Iterative De Bruijn graph short read Assembler)	(large) genomes	Sanger,454,Solexa	Yu Peng, Henry C. M. Leung, Siu-Ming Yiu, Francis Y. L. Chin	2010	(C / NC-A?)	link
MIRA (Mimicking Intelligent Read Assembly)	genomes, ESTs	Sanger, 454, Solexa	Chevreux, B.	1998 / 2011	OS	link

Source: Wikipedia

And many more…….



Graph Constructor	(large) genomes	Sanger, 454, Solexa, SOLiD	Convey Computer Corporation	2011	C	link
IDBA (Iterative De Bruijn graph short read Assembler)	(large) genomes	Sanger,454,Solexa	Yu Peng, Henry C. M. Leung, Siu-Ming Yiu, Francis Y. L. Chin	2010	(C / NC-A?)	link
MIRA (Mimicking Intelligent Read Assembly)	genomes, ESTs	Sanger, 454, Solexa	Chevreux, B.	1998 / 2011	OS	link
NextGENe	(small genomes?)	454, Solexa, SOLiD	Softgenetics	2008	C	link
Newbler	genomes, ESTs	454, Sanger	454/Roche	2009	C	link
PASHA	(large) genomes	Illumina	Liu, Schmidt, Maskell	2011	OS	link
Phrap	genomes	Sanger, 454, Solexa	Green, P.	1994 / 2008	C / NC-A	link
TIGR Assembler	genomic	[illegible]	[illegible]	1995 / 2003	OS	link
Ray[7]	genomes	[illegible]	[illegible]	2010	OS [GNU General Public License]	link
Sequencher	genomes	traditional and next generation sequence data	Gene Codes Corporation	1991 / 2009 / 2011	C	link
Se[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
SHARCGS	(small) genomes	Solexa	Dohm et al.	2007 / 2007	OS	link
SOPRA	genomes	Illumina, SOLiD, Sanger, 454	Dayarian, A. et al.	2010 / 2011	OS	link
SSAKE	(small) genomes	Solexa (SOLiD? Helicos?)	Warren, R. et al.	2007 / 2007	OS	link
SOAPdenovo	genomes	Solexa	Li, R. et al.	2009 / 2009	OS	link
SPAdes	(small) genomes, single-cell	Illumina, Solexa	Bankevich, A et al.	2012	OS	link
Staden gap4 package	BACs (, small genomes?)	Sanger	Staden et al.	1991 / 2008	OS	link
Taipan	(small) genomes	Illumina	Schmidt, B. *et al.*	2009	OS	link
VCAKE	(small) genomes	Solexa (SOLiD?, Helicos?)	Jeck, W. et al.	2007 / 2007	OS	link
Phusion assembler	(large) genomes	Sanger	Mullikin JC, *et al.*	2003	OS	link
Quality Value Guided SRA (QSRA)	genomes	Sanger, Solexa	Bryant DW, *et al.*	2009	OS	link
Velvet	(small) genomes	Sanger, 454, Solexa, SOLiD	Zerbino, D. et al.	2007 / 2009	OS	link

*Licences: OS = Open Source; C = Commercial; C / NC-A = Commercial but free for non-commercial and academics; Brackets = unclear, but most likely C / NC-A

A grand total of 42!

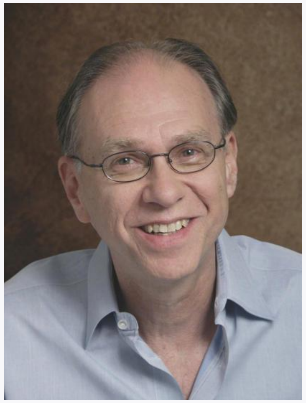
Why is there no single optimal algorithm?

A basic question

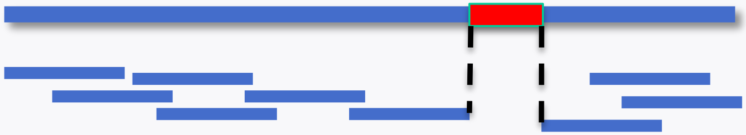
- What is the minimum number of reads required for reliable reconstruction?
- How much intrinsic information does each read provide about the DNA sequence?
- A benchmark for comparing different algorithms and different technologies.
- An open question!

Coverage Analysis

- Pioneered by Lander-Waterman



- What is the minimum number of reads to ensure there is no gap between the reads with a desired prob.?

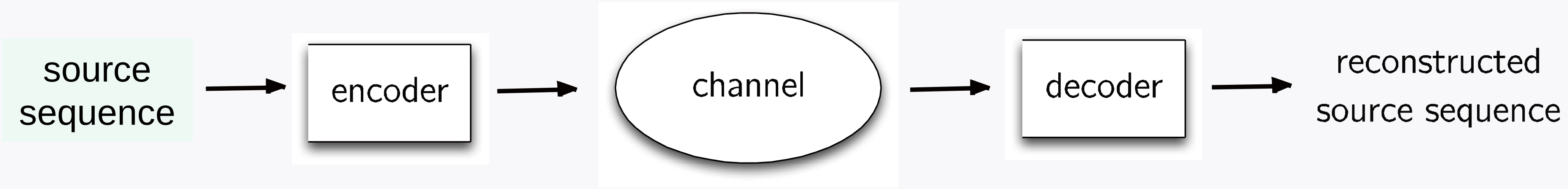


- Only provides a lower bound on the minimum number of reads to reconstruct.

- Clearly not tight.

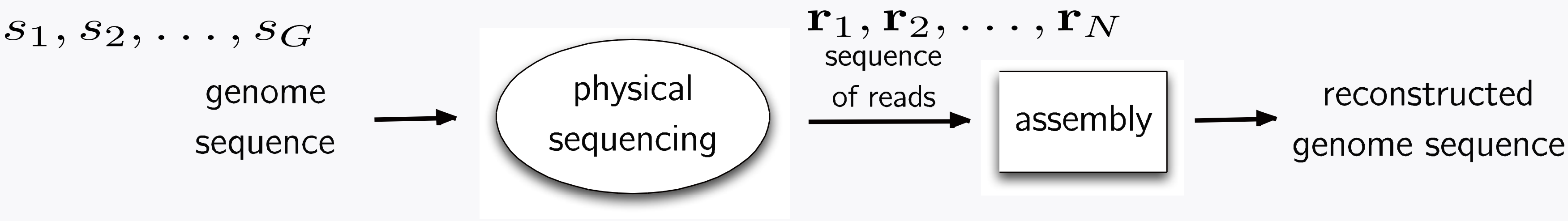
Communication and Sequencing: An Analogy

Communication:



$$\text{max. communication rate } R = \frac{C_{\text{channel}}}{H_{\text{source}}} \text{ source sym / sec.}$$

Sequencing:



$$\text{sequencing rate } R = \frac{G}{N} \text{ DNA sym / read}$$

Question: what is the max. sequencing rate such that reliable reconstruction is asymptotically possible?

A Basic Model

- DNA sequence: i.i.d. with marginal distribution

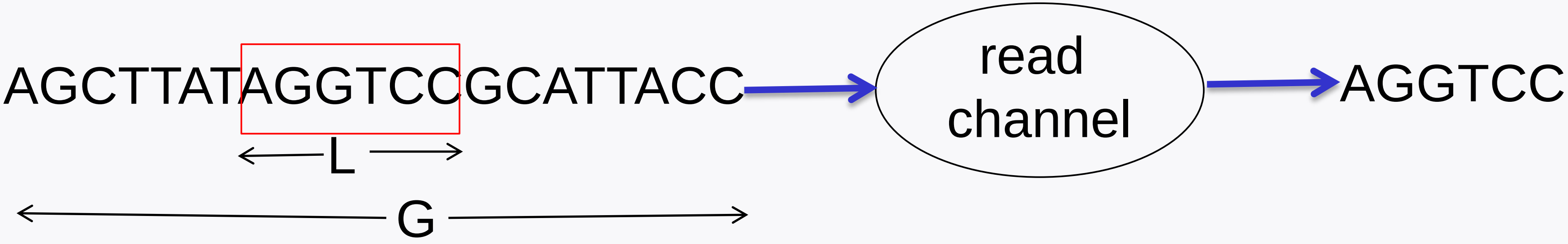
$$\mathbf{p}=(p_1, p_{\mathbf{2}}, p_3, p_4).$$

- Starting positions of reads: i.i.d. uniform on the DNA sequence.

- Read process: noiseless.

Will build on this to look at statistics from genomic data.

The read channel

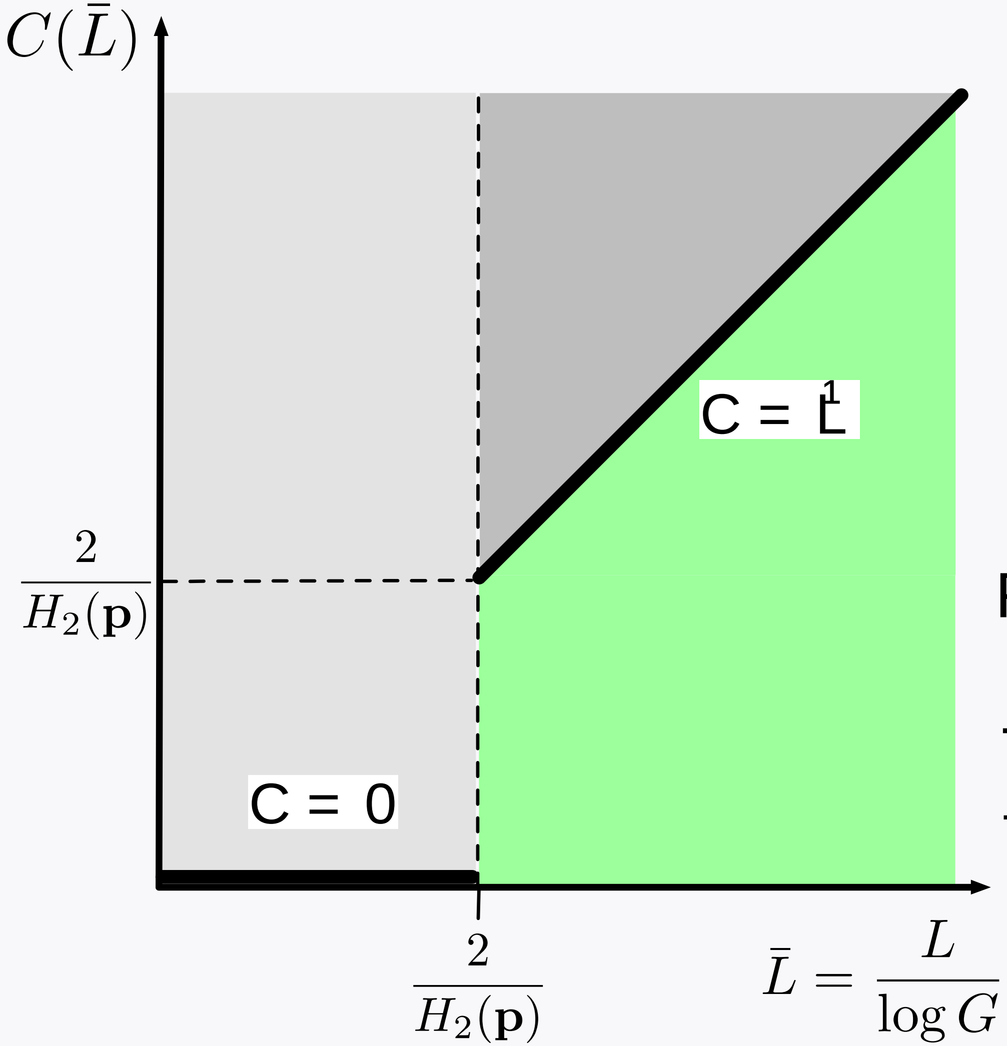


- Capacity depends on
  - read length: L $L \uparrow \Rightarrow C \uparrow$
  - DNA length: G $G \uparrow \Rightarrow C \downarrow$

- Normalized read length: $\bar{L} := \frac{L}{\log G}$

- Eg. L = 100, G = 3 £ 10^9 : $\bar{L} = 4{:}6$

Result: Sequencing Capacity



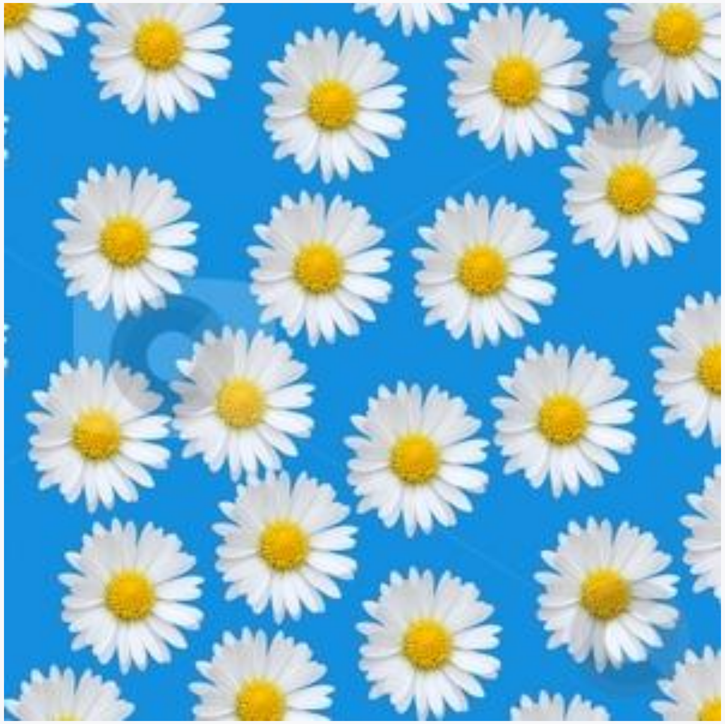
$$H_2(\mathbf{p}) = -\log \sum_{i=1}^{4} p_i^2$$

Renyi entropy of order 2

The higher the entropy, the easier the problem!

Complexity is in the eye of the beholder

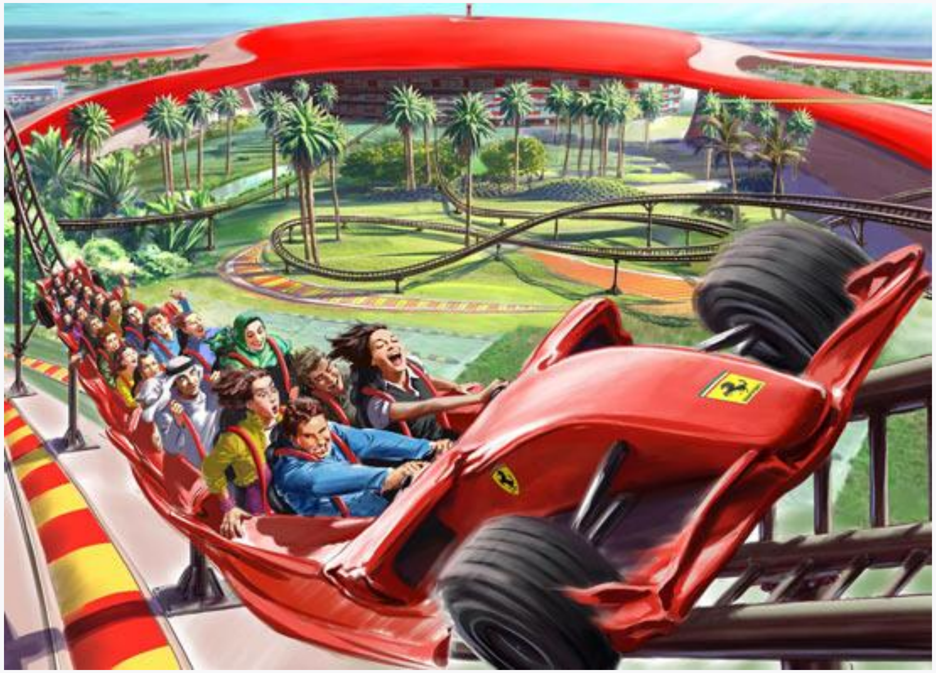
Low entropy



easier to compress

harder jigsaw puzzle

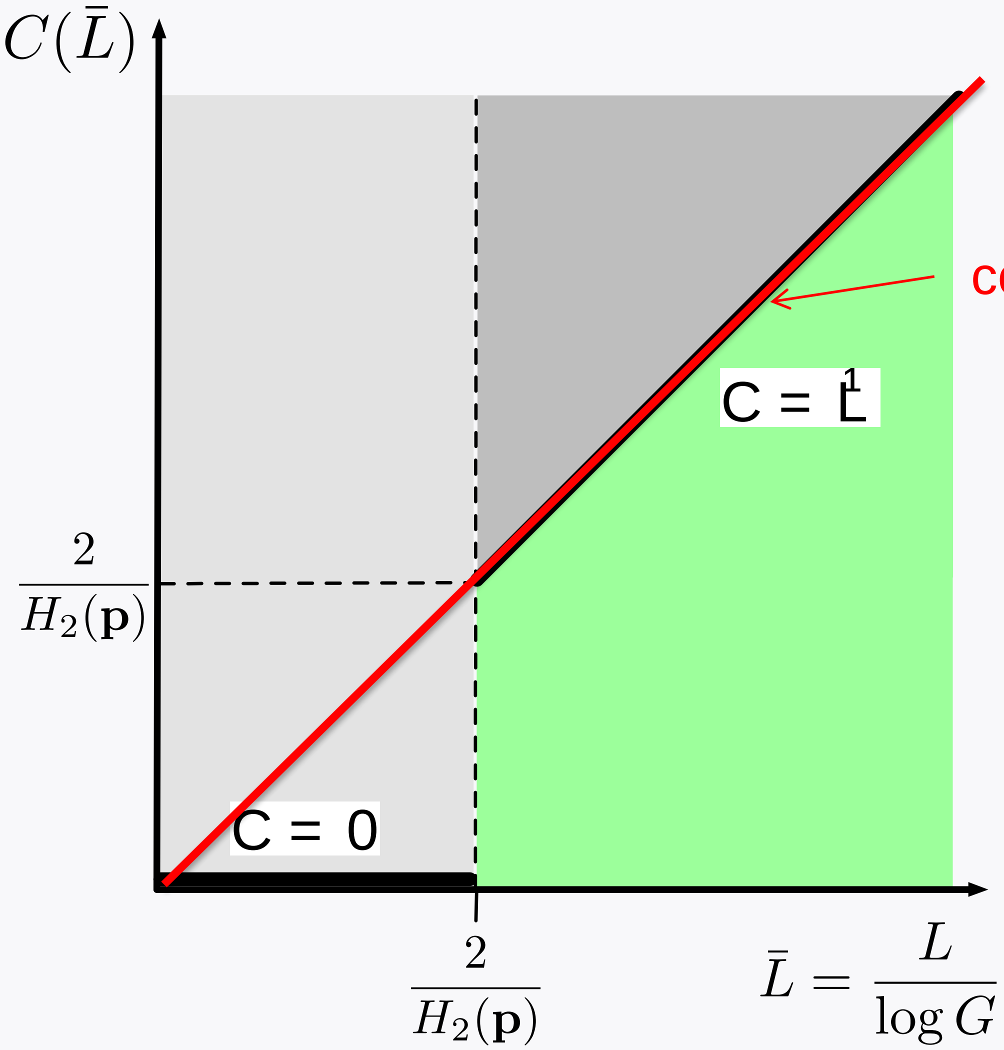
High entropy



harder to compress

easier jigsaw puzzle

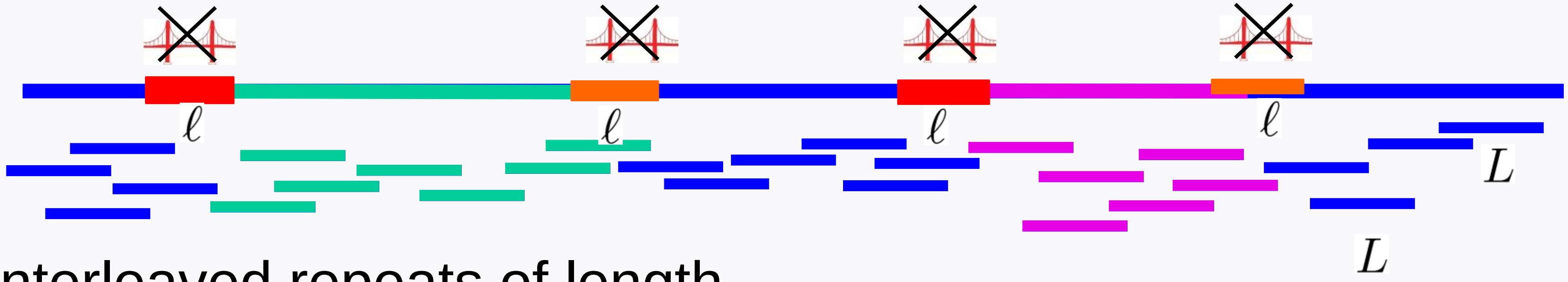
Capacity Result Explained



coverage constraint

N > G=L log G

A necessary condition for reconstruction



interleaved repeats of length

ℓ

None of the copies is straddled by a read (unbridged).

Reconstruction is impossible!

Special cases:

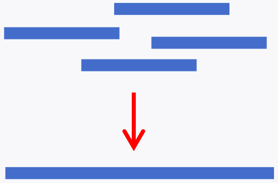
$\ell = L-1:$ No interleaved repeats of length $L-1$ (Ukkonen)

$\ell = 1:$ roughly equivalent to coverage

Under i.i.d. model, greedy is optimal for mixed reads.

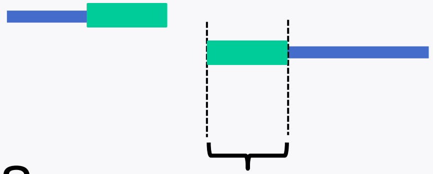
A sufficient condition: greedy algorithm

Input: the set of N reads of length L



contig

1. Set the initial set of contigs as the reads
2. Find two contigs with largest overlap and merge them into a new contig



overlap

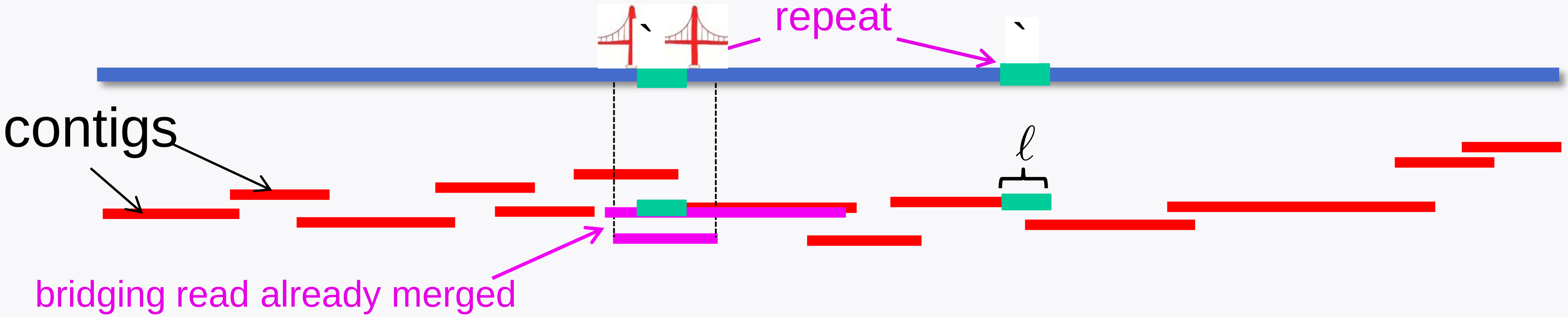
3. Repeat step 2 until only one contig remains

Algorithm progresses in stages:

at stage $\ell = L-1, L-2, \ldots, 1$

merge reads at overlap ℓ

Greedy algorithm: stage ℓ



merge mistake => there must be a ℓ-repeat

and

each copy is not bridged by a read.

A sufficient condition for reconstruction:

There is no unbridged ℓ- repeat for any ℓ.

Summary

Necessary condition for reconstruction:

No unbridged interleaved ℓ-repeats for any ℓ .

Sufficient condition (via greedy algorithm)

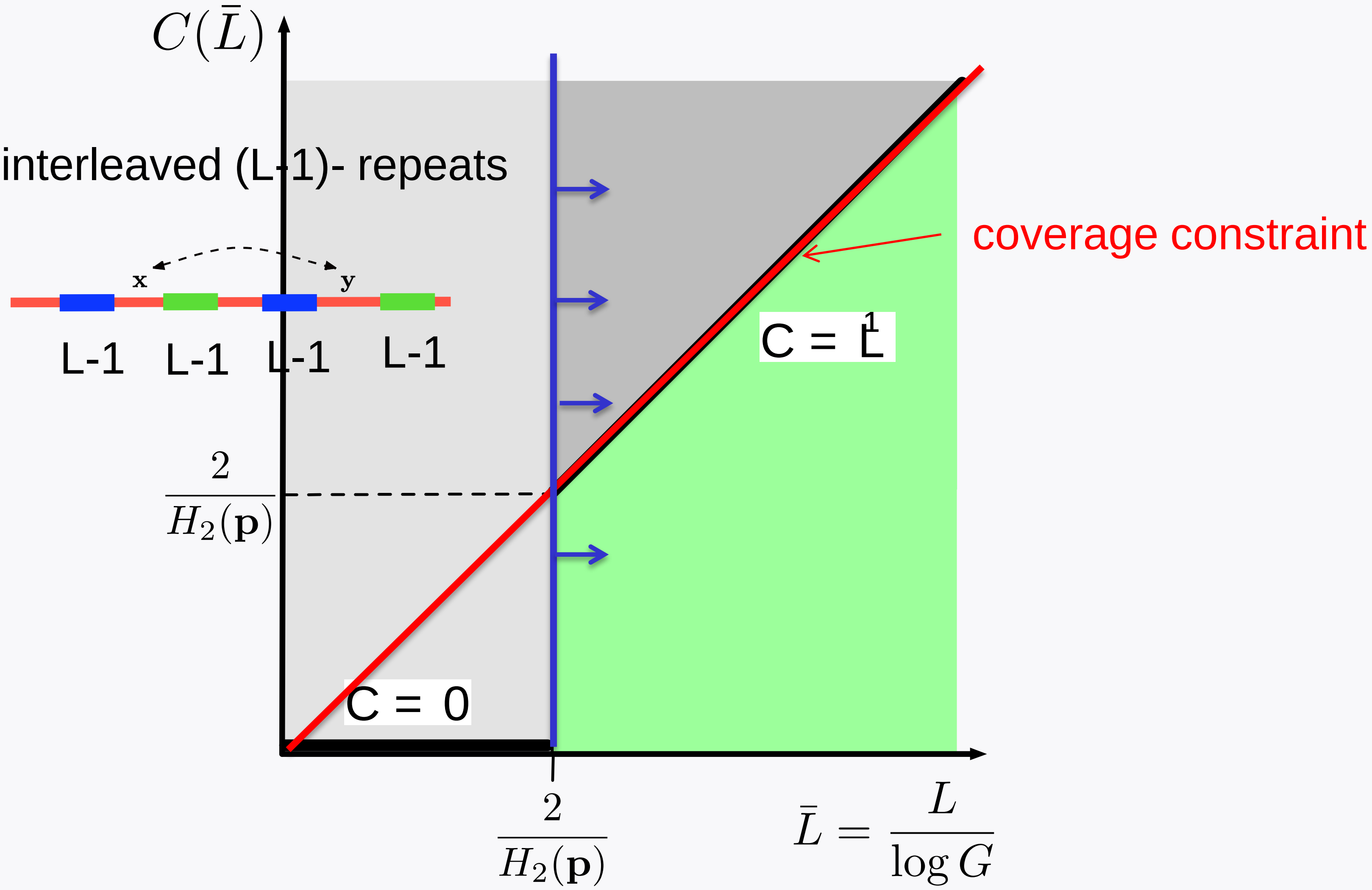
No unbridged ℓ-repeats for any ℓ .

For the i.i.d. DNA model:

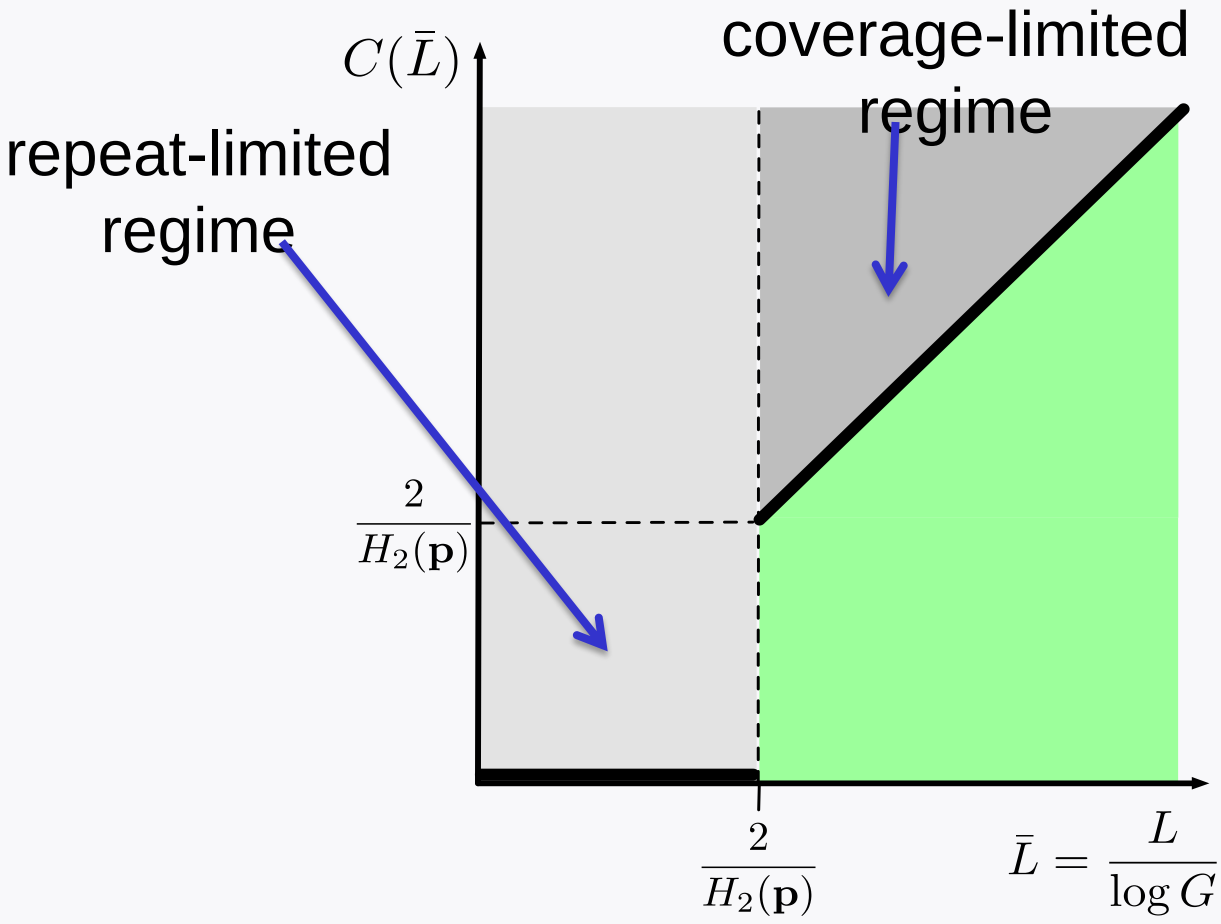
1) If there are no unbridged interleaved repeats, then w.h.p. there are no unbridged repeats.

2) The probability is dominated when either $\ell = L - 1 \text{ or } \ell = 1$

Capacity Result Explained



Summary: Two Regimes



Question:

Is this clean state of affairs tied to the i.i.d. DNA model?

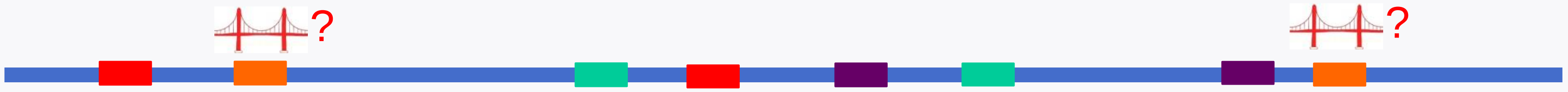
I.I.D. DNA vs real DNA

- Mammalian DNA has many long repeats.

- How will the greedy algorithm perform for general DNA statistics?

- Will there be a clean decomposition into two regimes?

Greedy algorithm: general DNA statistics



- Reconstruction if there are no unbridged repeats.

- Performance depends on the DNA statistics through the number of ` - repeats:

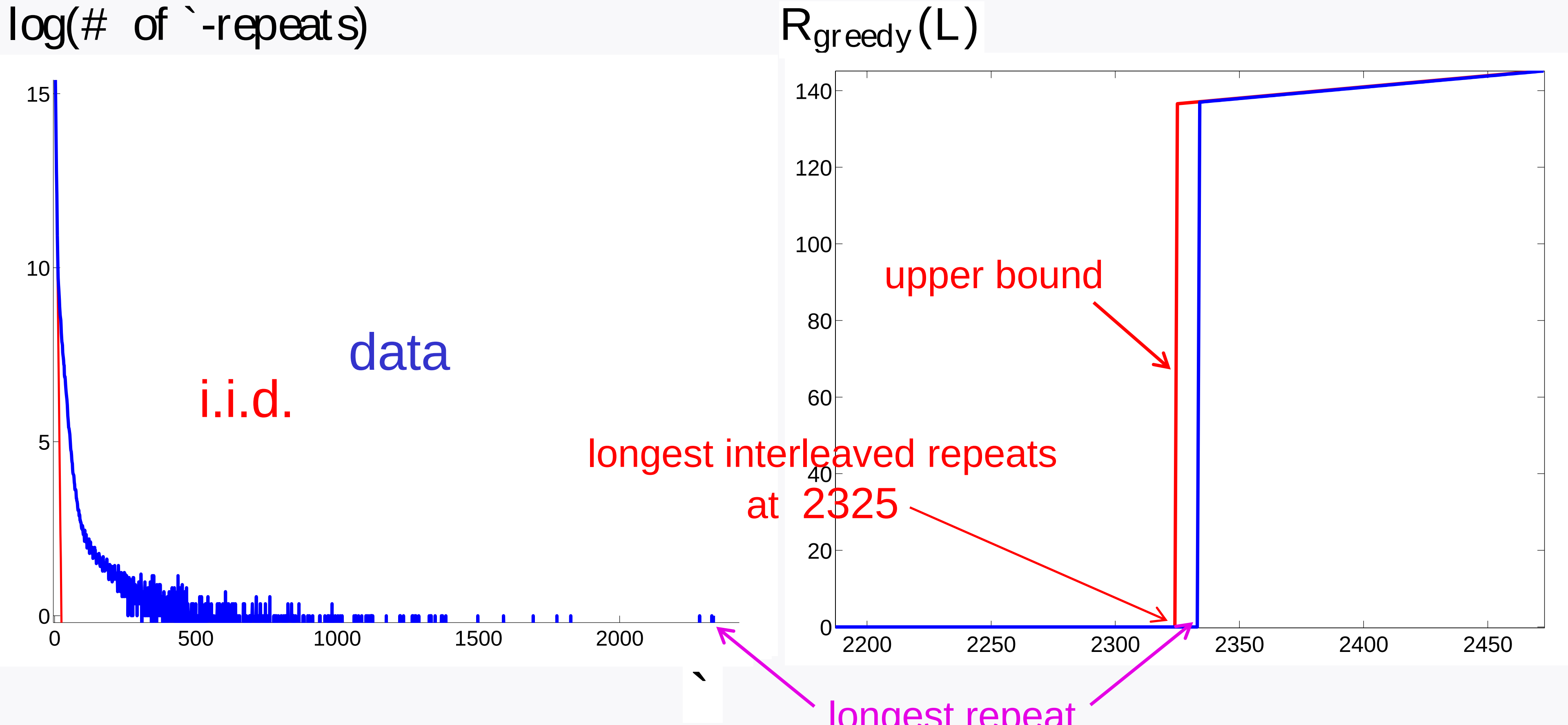
$$R_{\text{greedy}}(L) = \min_{1\leq\ell\leq L-1} \frac{2(L-\ell-1)}{\log(\#\text{ of }\ell\text{-repeats})}$$

- Necessary condition translates similarly to an upper bound on capacity:

$$C(L) \leq \min_{1\leq\ell\leq L-1} \frac{4(L-\ell-1)}{\log(\#\text{ of interleaved }\ell\text{-repeats})}$$

I.I.D. DNA vs real DNA

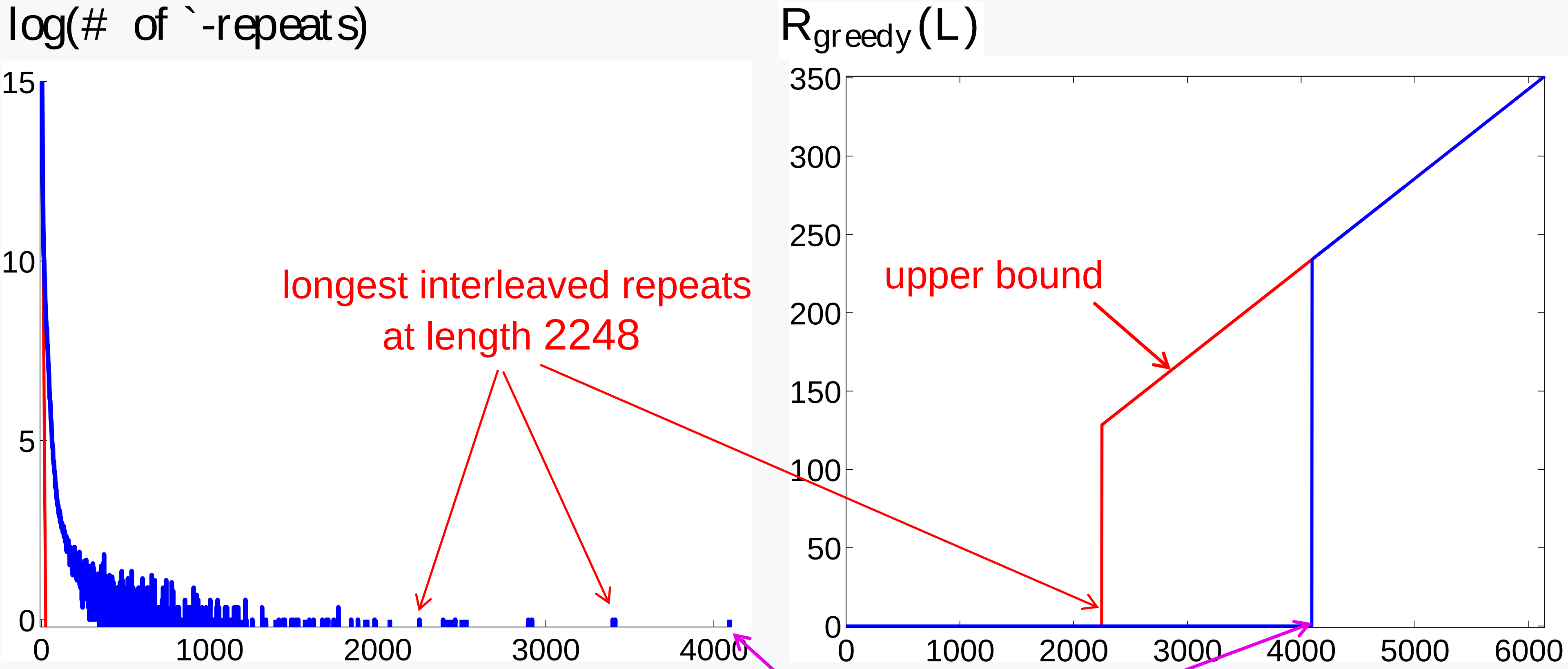
$$R_{\text{greedy}}(L) = \min_{1 \le \ell \le L-1} \frac{2(L-\ell-1)}{\log(\#\text{ of }\ell\text{-repeats})}$$



GRCh37 Chr 22 (G = 35M)

Chromosome 19

There is another more sophisticated algorithm that would close the gap and in fact near optimal on all 22 chromosomes.



longest interleaved repeats at length 2248

upper bound

longest repeat at $\ell = 4092$

GRCh37 Chr 19 (G = 55M)

Ongoing work

- Noisy reads
- Reference-based assembly.
- Partial reconstruction.

………..

Conclusion

- Information theory has made a huge impact on communication.

- Its success stems from focusing on something fundamental: information.

- This philosophy may be useful for other important engineering problems.

- DNA sequencing is a good example.