DNA Sequence Algorithms & FISH

Conor Camplisson

Genome 540

March 7th, 2023

Genome Sciences
UNIVERSITY OF WASHINGTON

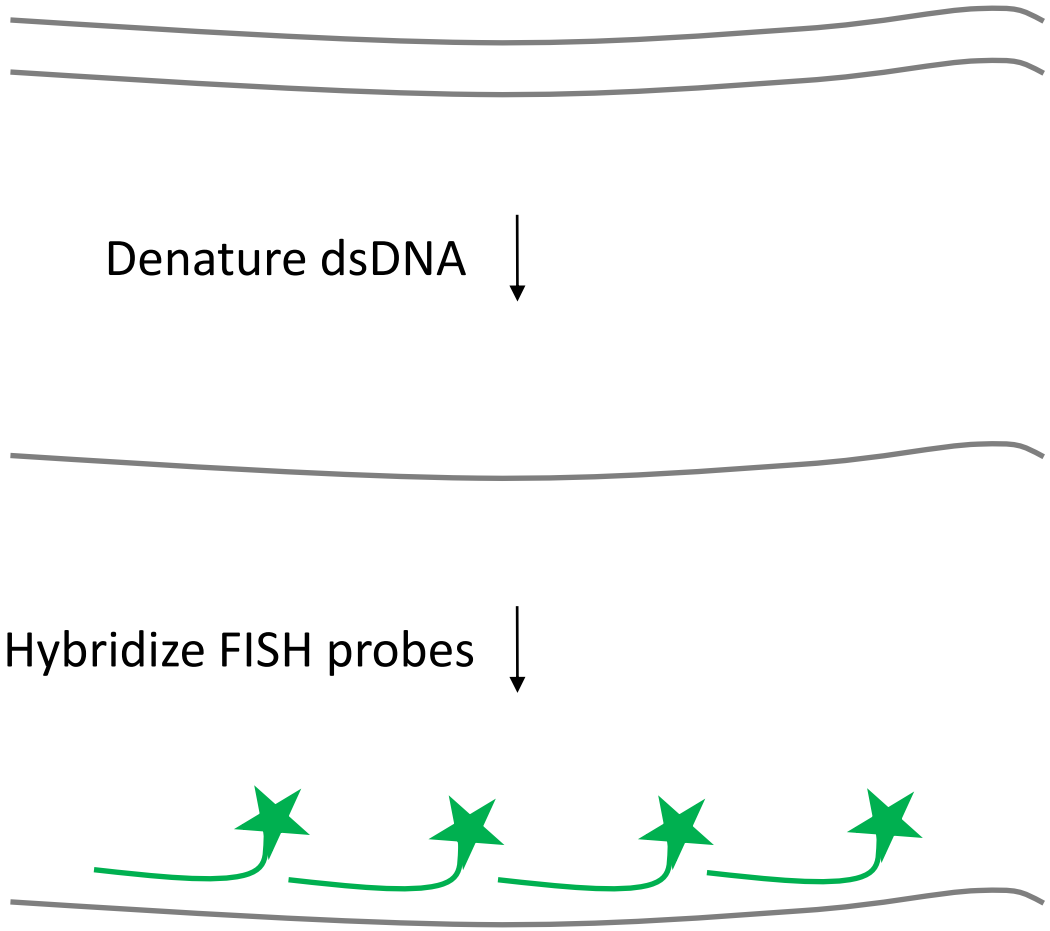
Outline

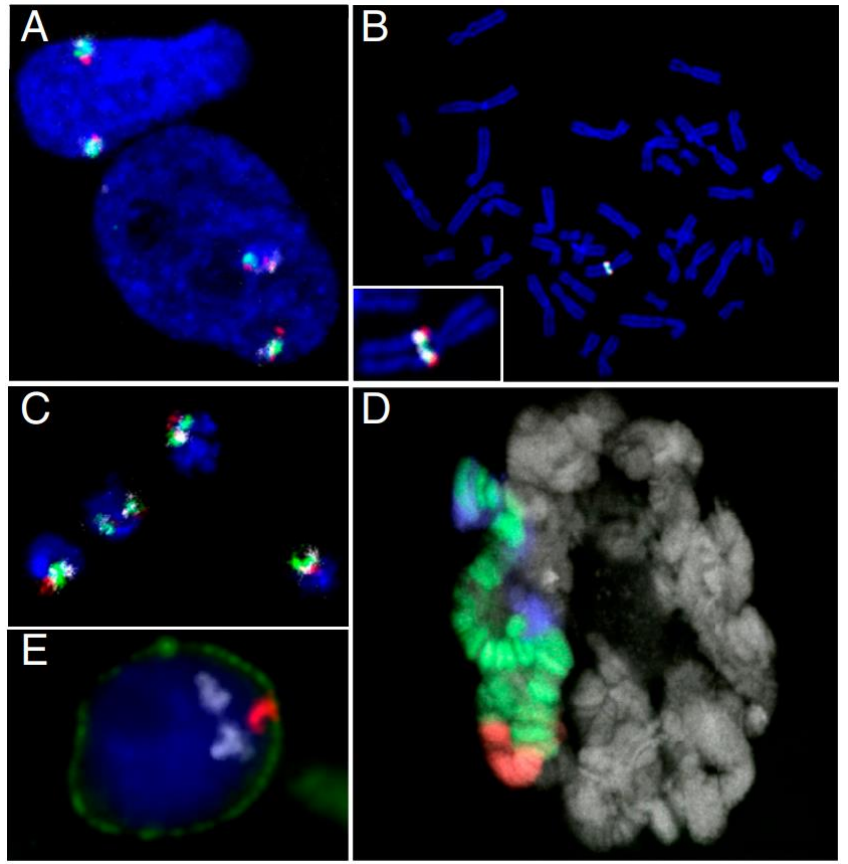
- Fluorescence in situ hybridization (FISH)
- Information theory & FISH
  - Compressed sensing
  - Linear block theory, error-correction
- de Bruijn sequences
- Orthogonal DNA sequence set design
  - k-mer symmetry minimization
  - Hamming distance approaches
- New algorithm idea…

Outline

- Fluorescence in situ hybridization (FISH)
- Information theory & FISH
  - Compressed sensing
  - Linear block theory, error-correction
- de Bruijn sequences
- Orthogonal DNA sequence set design
  - k-mer symmetry minimization
  - Hamming distance approaches
- New algorithm idea…

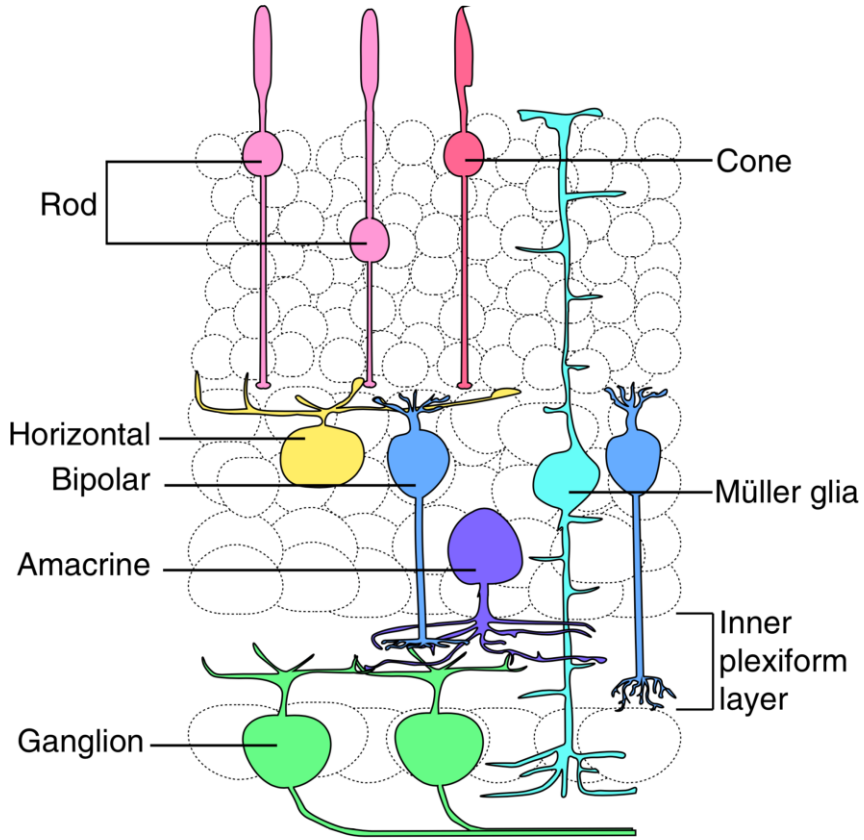
FISH: Fluorescence *in situ* Hybridization

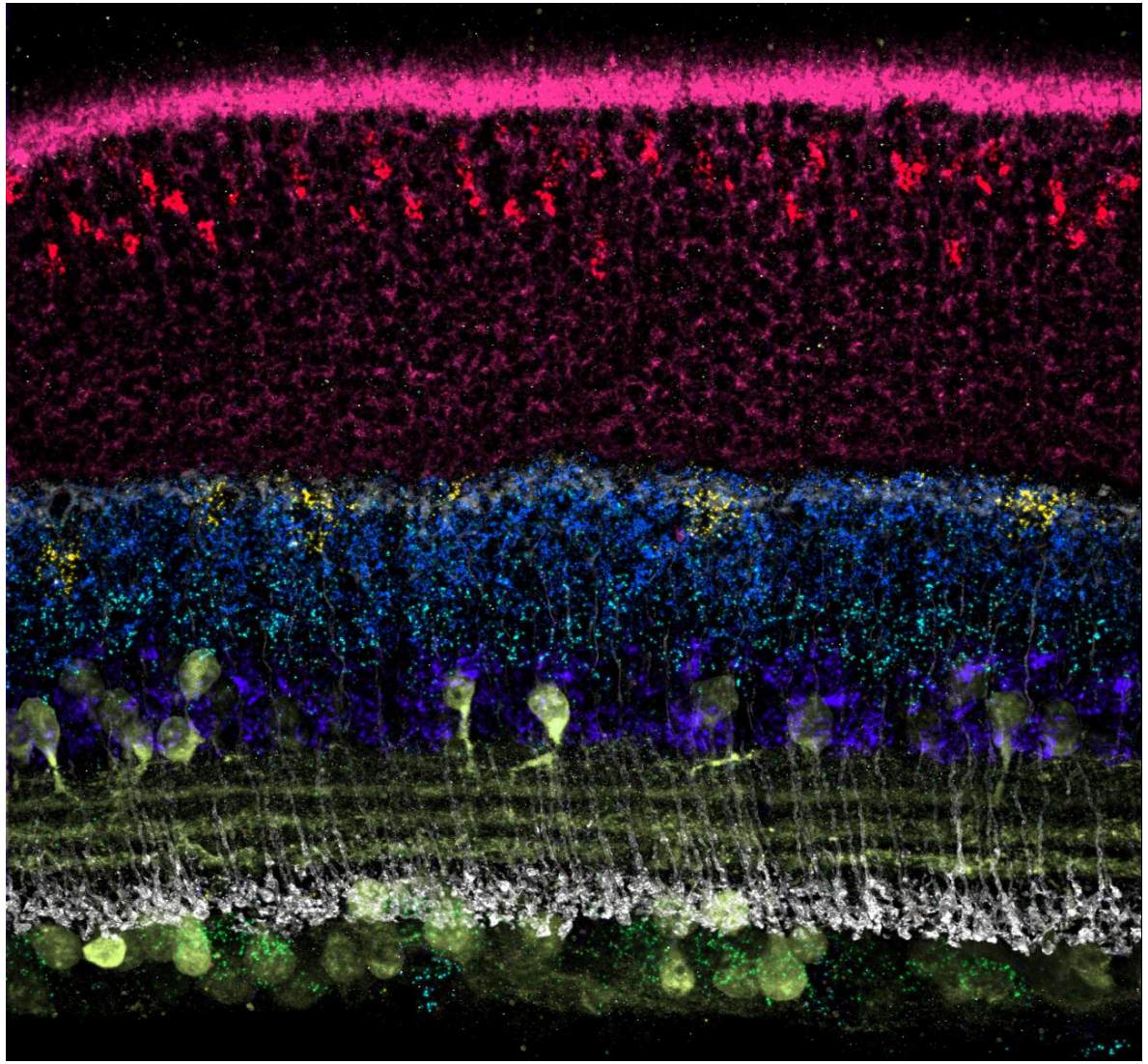




Beliveau, B.J. Et al. *PNAS* **109 (52)**, 21301-6 (2012)

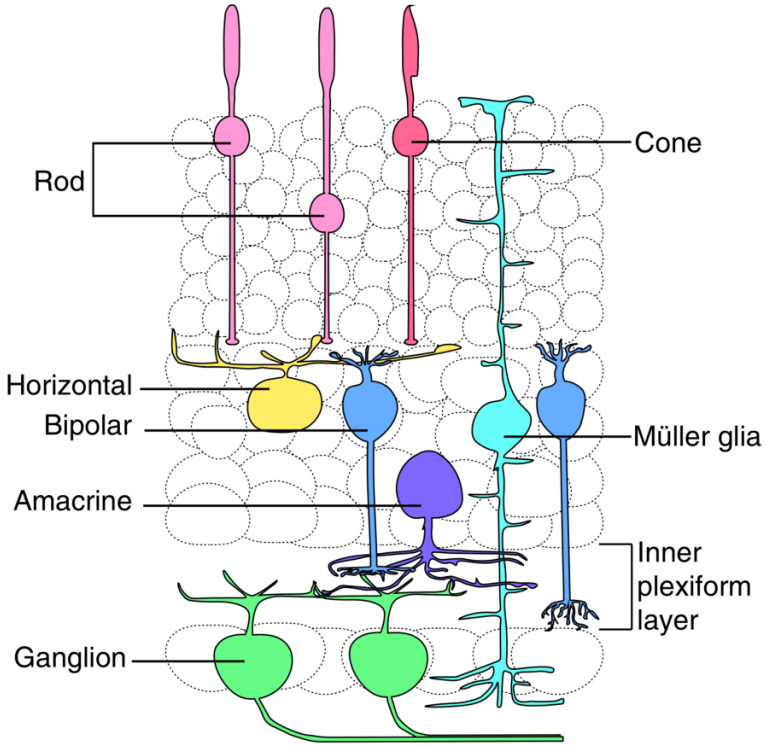
Targeting RNA & Proteins



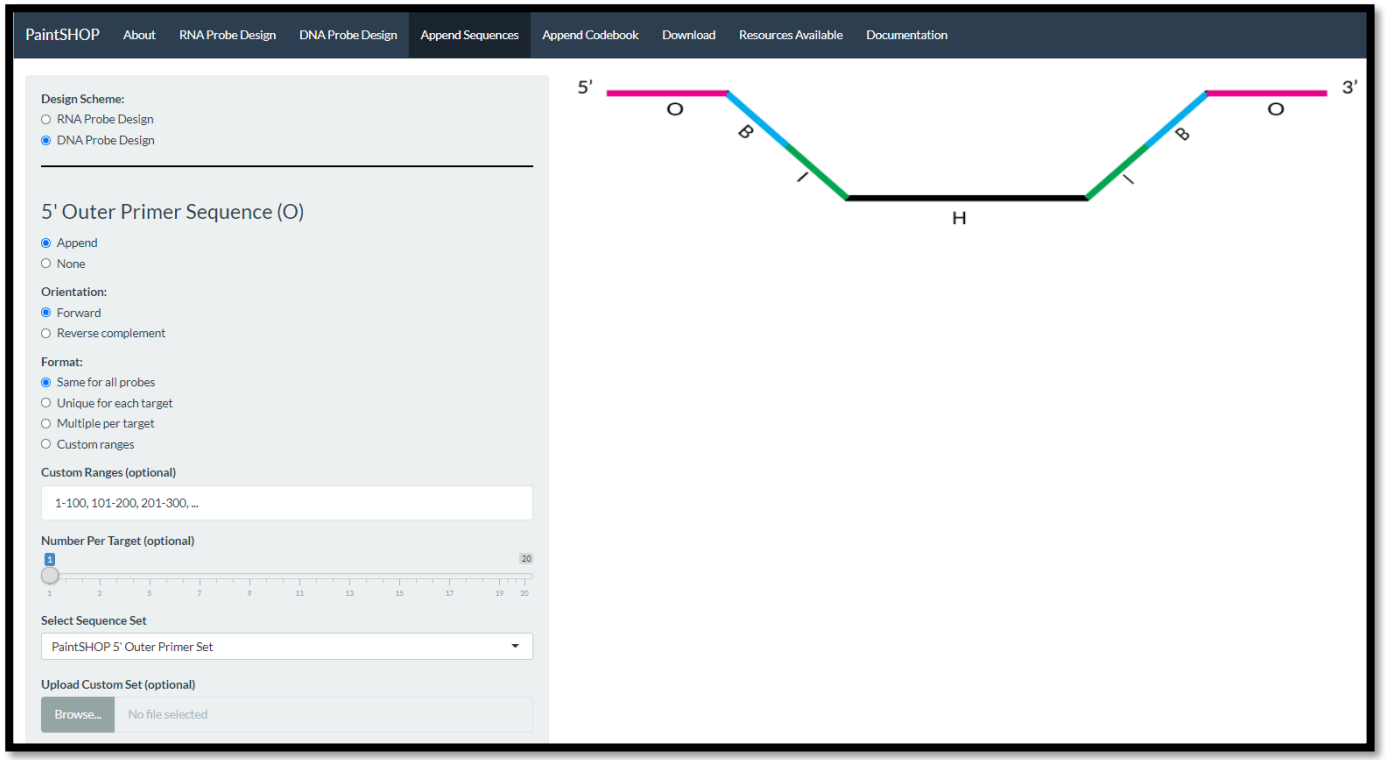


Kishi, J.Y., Lapan, S.W., Beliveau, B.J. et al. *Nat Methods* **16**, 533–544 (2019)

PaintSHOP simplifies complex probe design

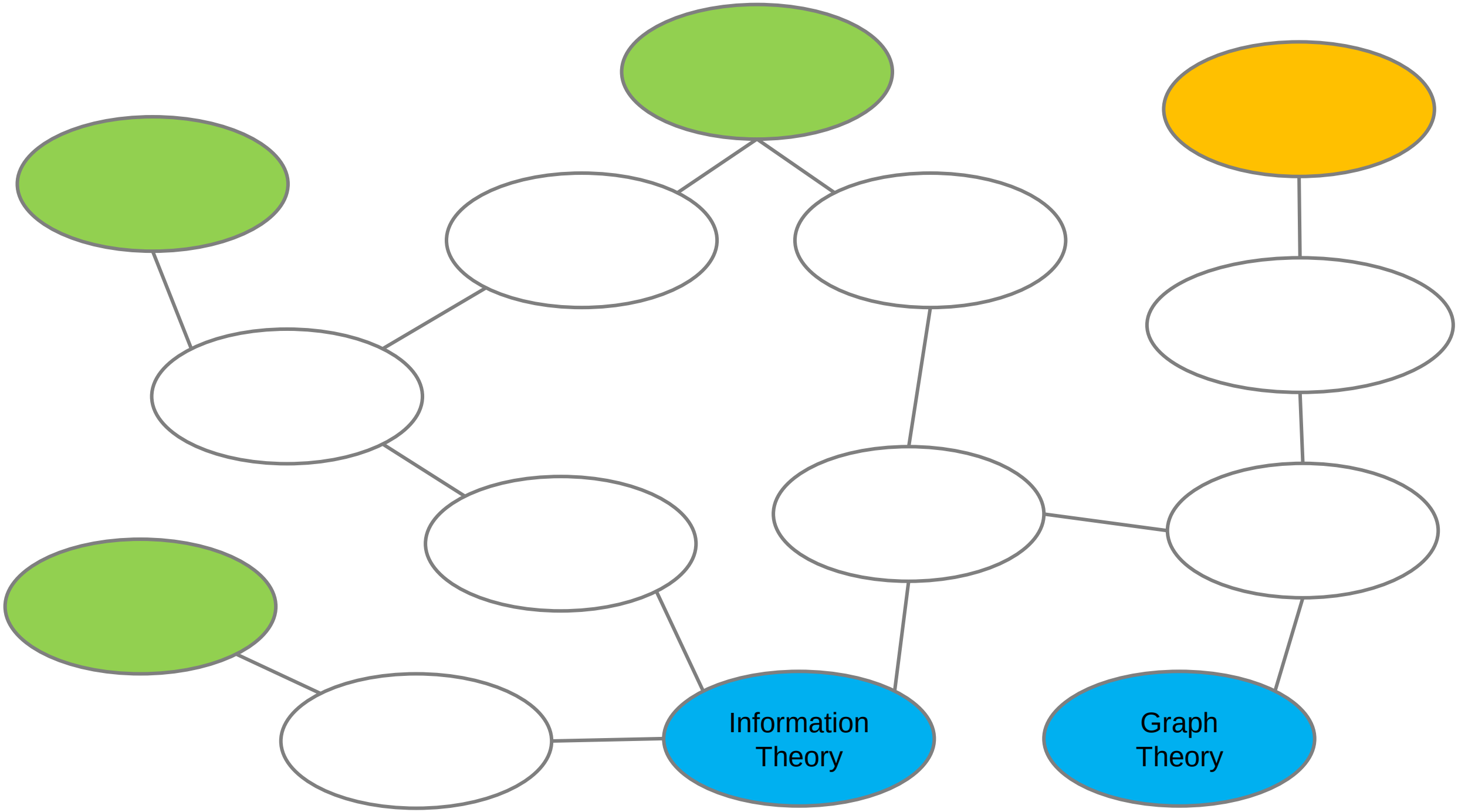


web server: paintshop.io



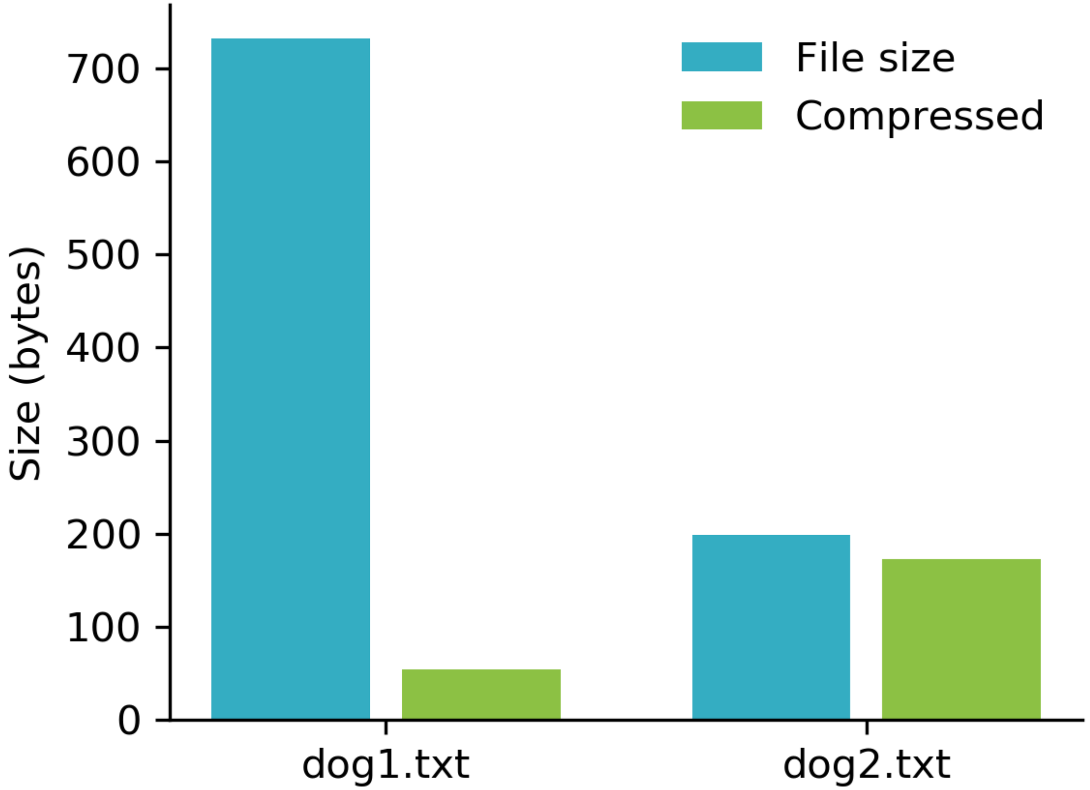
Features for multiplexing, barcoding probes

Kishi, J.Y., Lapan, S.W., Beliveau, B.J. et al. *Nat Methods* **16**, 533–544 (2019)


Information
Theory
Graph
Theory

Information Content

file	size	compressed
dog1.txt	732	54
dog2.txt	199	173



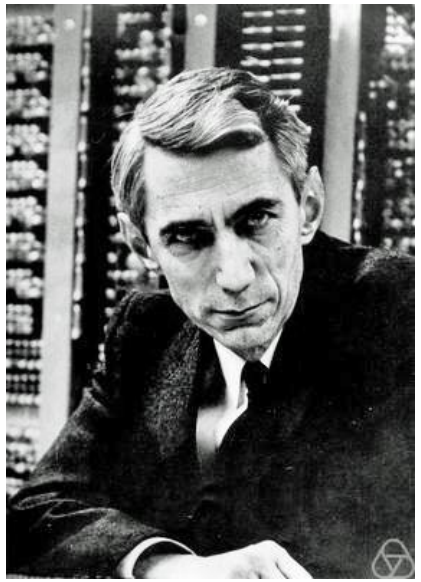

```
In [61]: print_contents(file_1)

DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG
DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG DOG

In [62]: print_contents(file_2)

Lassie is a fictional character created by Eric Knight.

She is a female Rough Collie dog, and is featured in a short story
that was later expanded to a full-length novel called Lassie Come-Home.
```

Information Theory

Quantifying information

Heads ✗

Tails ✓



Claude Shannon

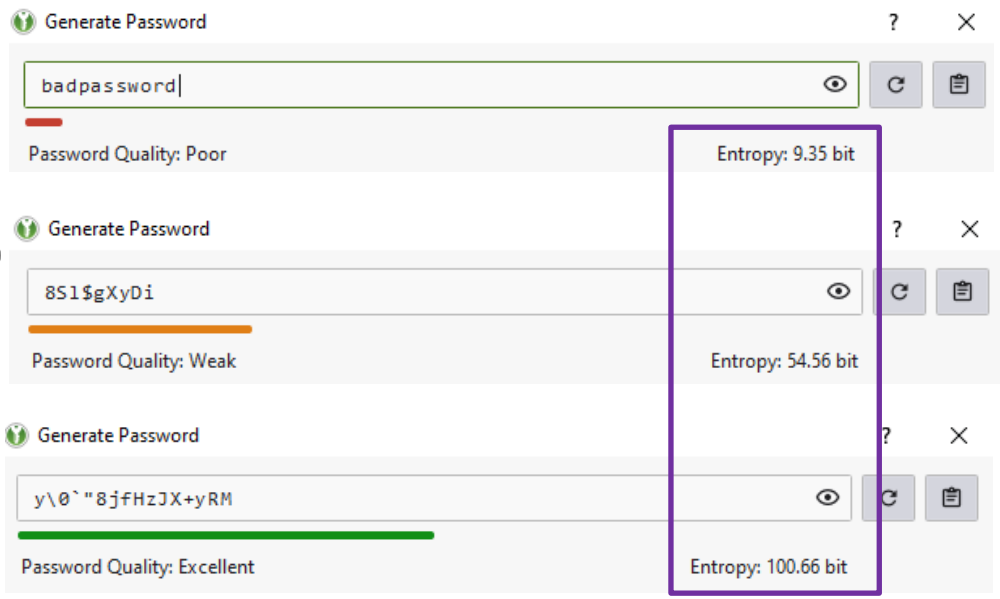
Information entropy

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

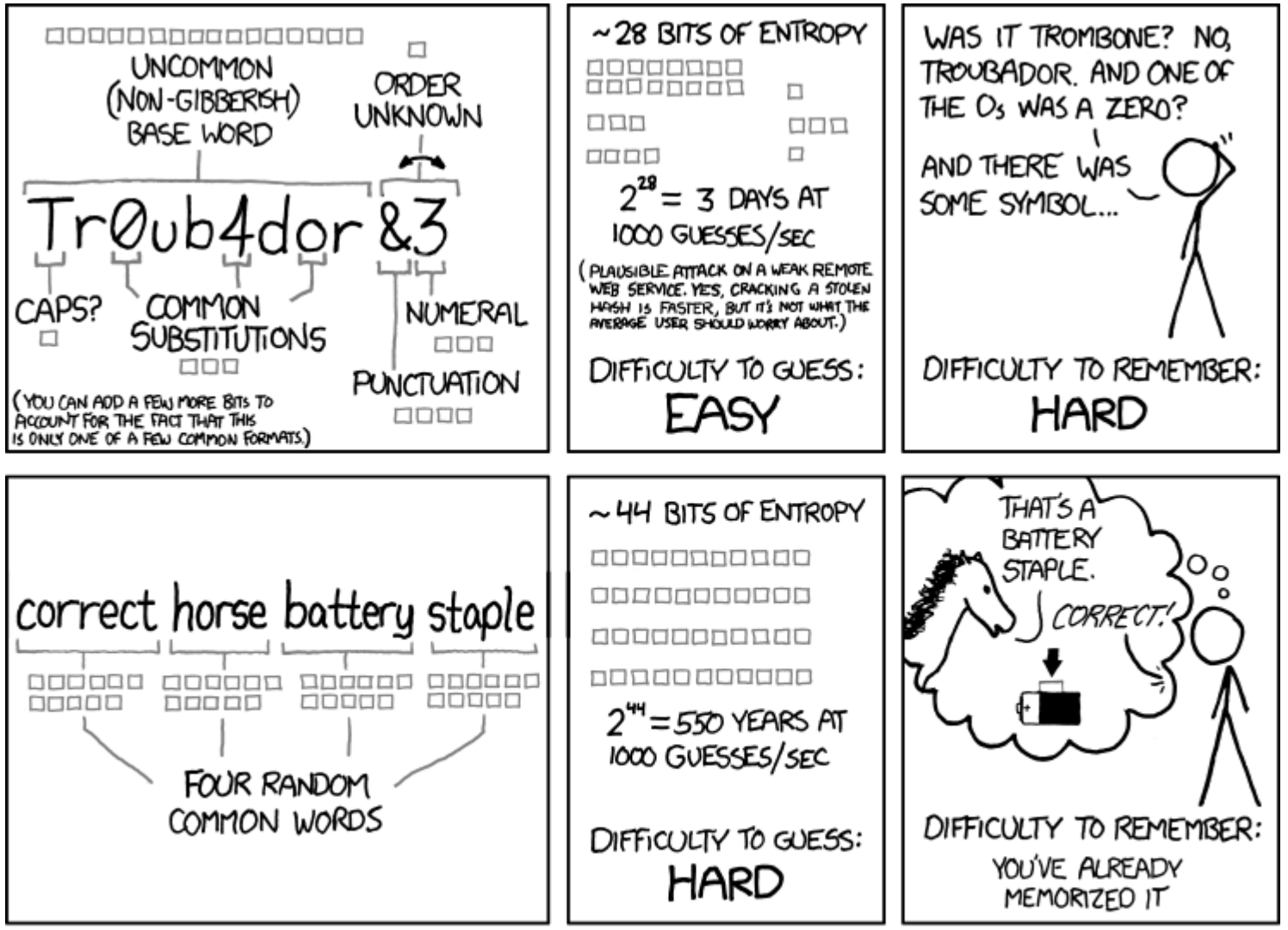
H(toss) = -(p(heads) * log2 (p(heads) + p(tails) * log2(p(tails))

H(toss) = -(0.5*log2(0.5) + 0.5*log2(0.5)) = **1.0 bit**

"binary digit" → "bit"



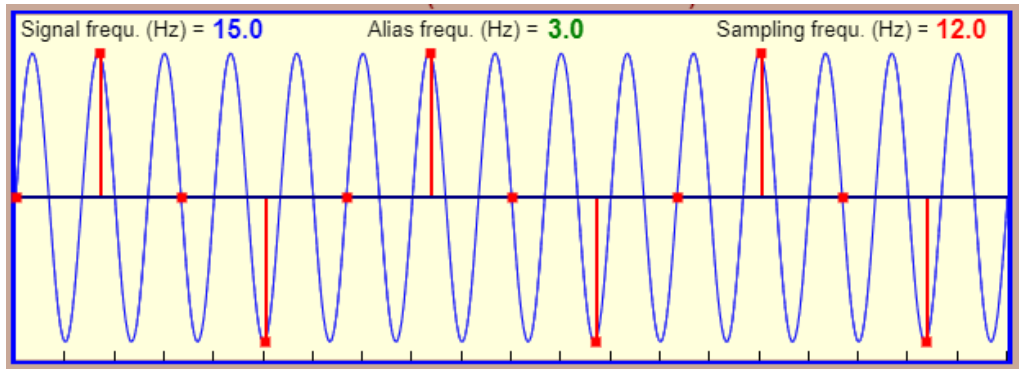
There's always a relevant xkcd

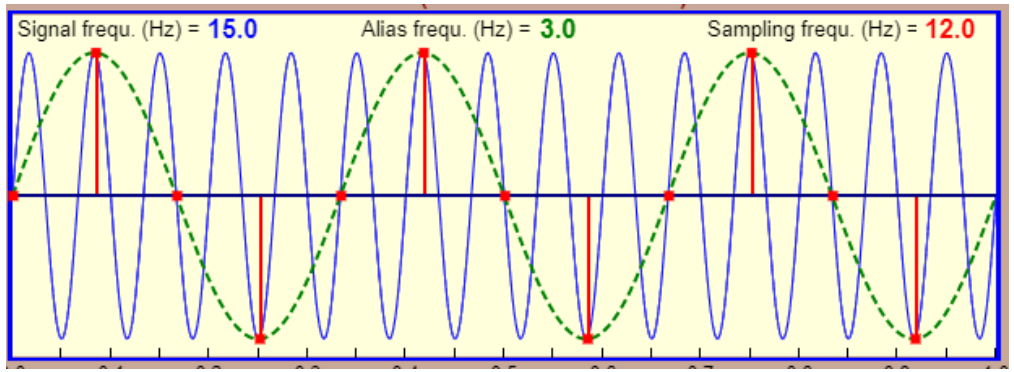


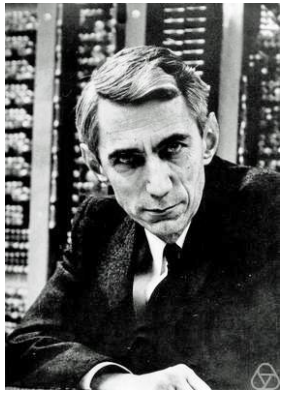
https://imgs.xkcd.com/comics/password_strength.png

Shannon-Nyquist Sampling Theorem

Sub-Nyqist sampling:



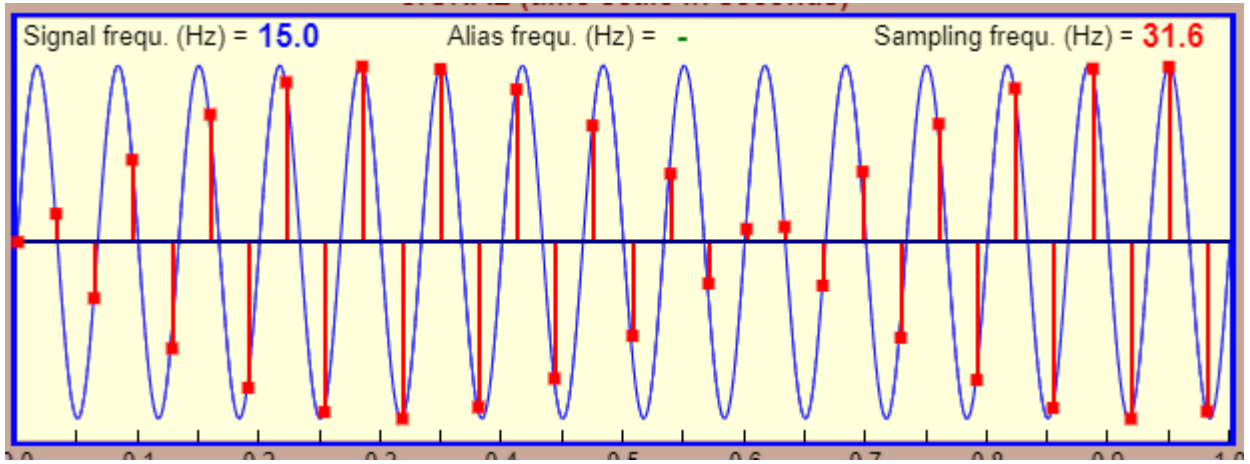




Claude Shannon

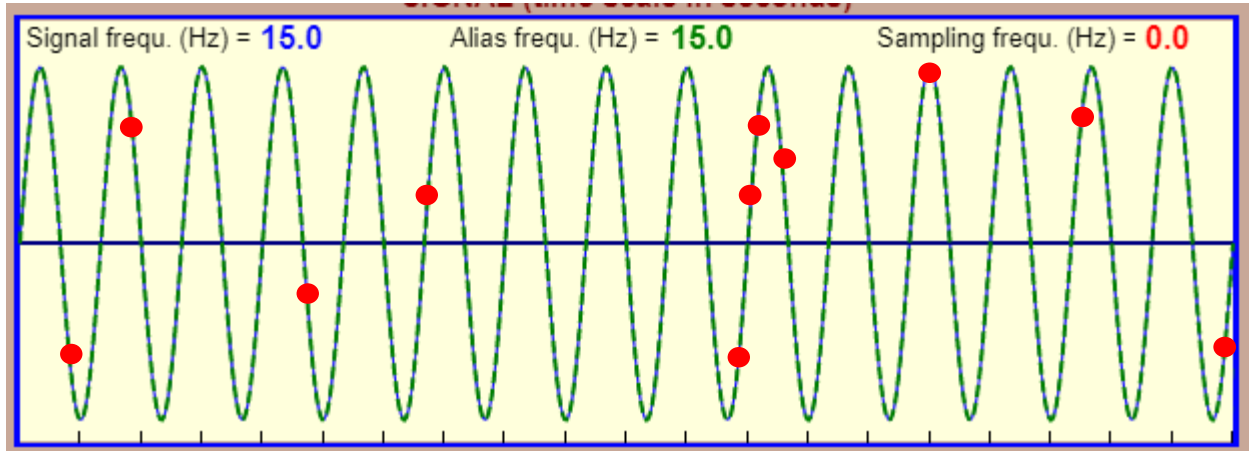
Harry Nyquist

Sampling above Nyquist rate:

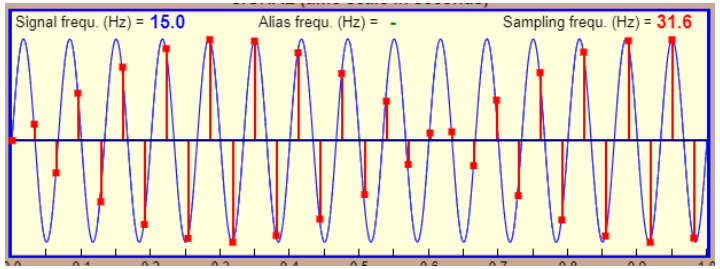


Compressed Sensing

Randomly downsampled



^ Far fewer samples than Nyquist rate:



Compressed Sensing

nature biotechnology

LETTERS

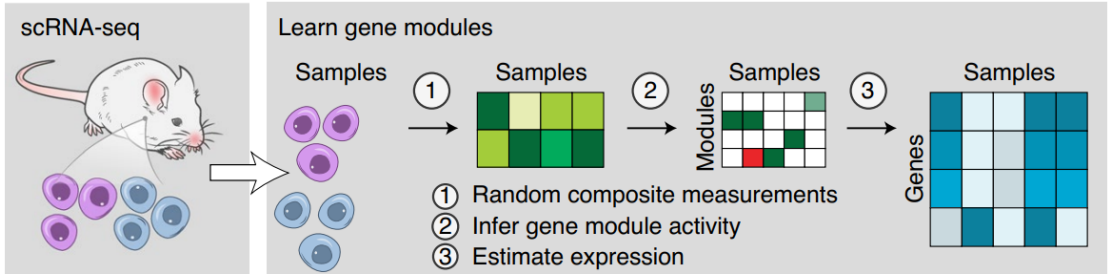
https://doi.org/10.1038/s41587-021-00883-x

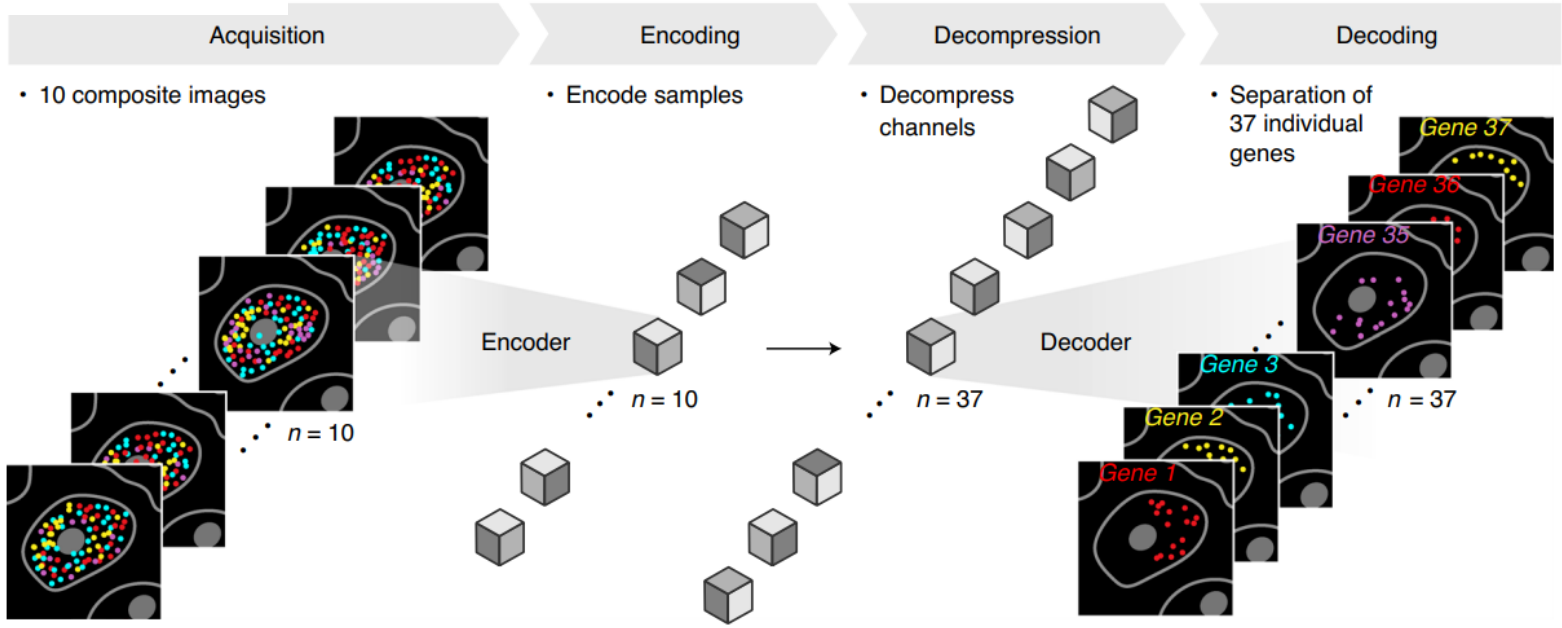
Check for updates

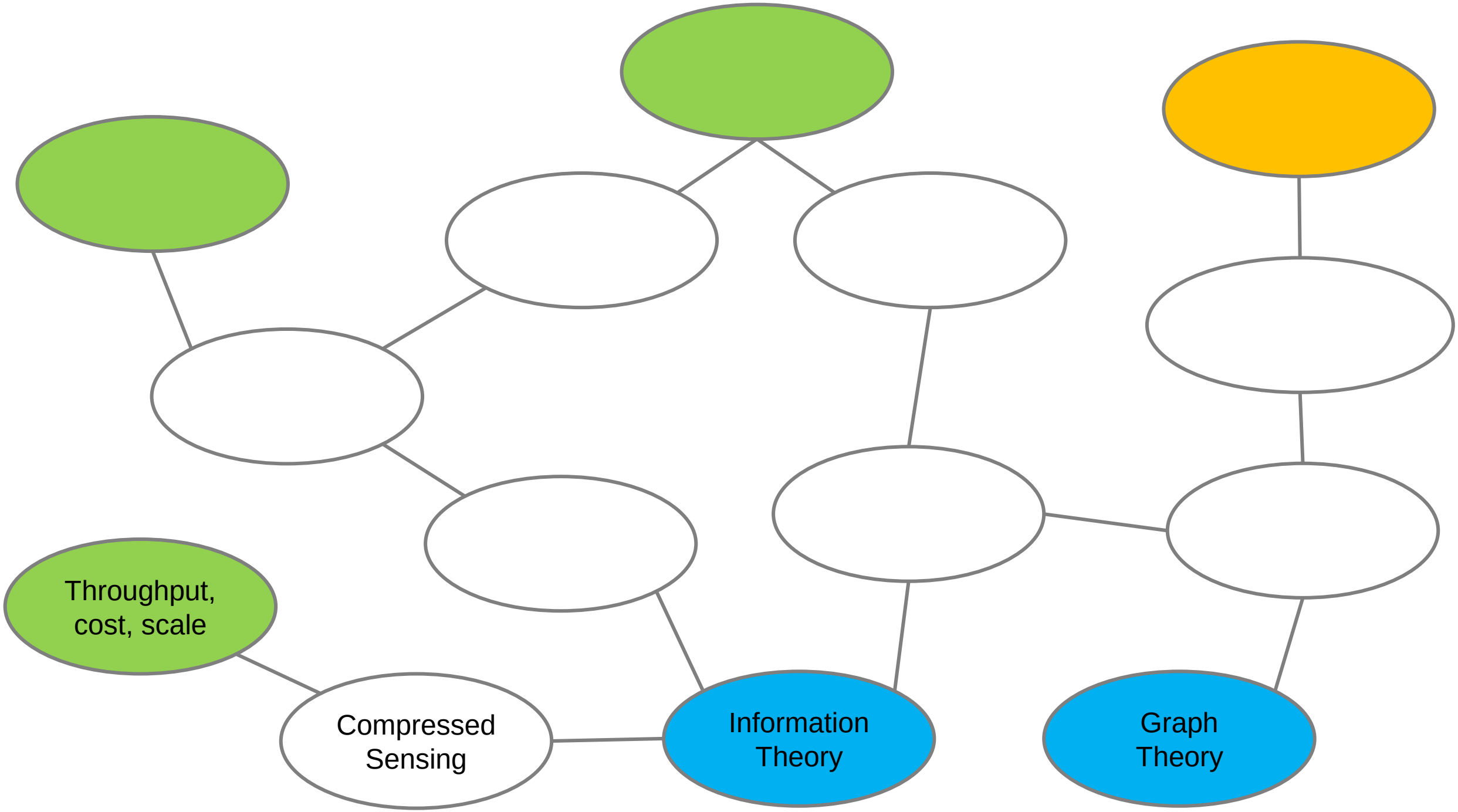
Compressed sensing for highly efficient imaging transcriptomics

Brian Cleary[1]✉, Brooke Simonton[1], Jon Bezney[1], Evan Murray[1], Shahul Alam[1], Anubhav Sinha[1,2], Ehsan Habibi[1], Jamie Marshall[1], Eric S. Lander[1,3,4]✉, Fei Chen[1,5]✉ and Aviv Regev[3,6,7,8]✉

Received: 16 August 2019; Accepted: 11 March 2021;
Published online: 15 April 2021






Throughput,
cost, scale
Compressed
Sensing
Information
Theory
Graph
Theory

Information Theory

sending information over a noisy channel

Information Theory

sending information over a noisy channel

Harry Nyquist

Certain Topics in Telegraph Transmission Theory

H. NYQUIST, MEMBER, A. I. E. E.

Classic Paper

Bell Laboratories

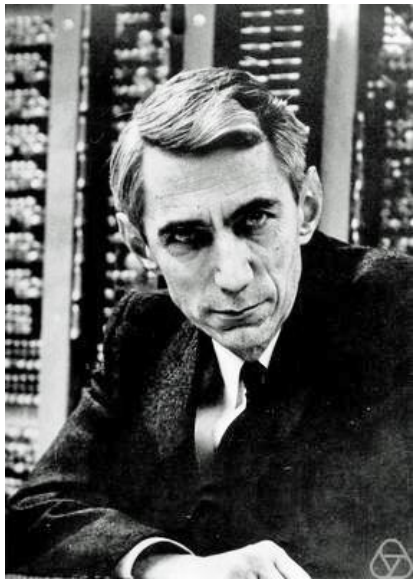
- Radio astronomy
- Transistor
- LASER
- Photovoltaic cell
- Charge-coupled device (CCD)
- UNIX, C, C++, AWK, others
- 9 Nobel Prizes
- *Information Theory*

Communication in the Presence of Noise*

CLAUDE E. SHANNON†, MEMBER, IRE

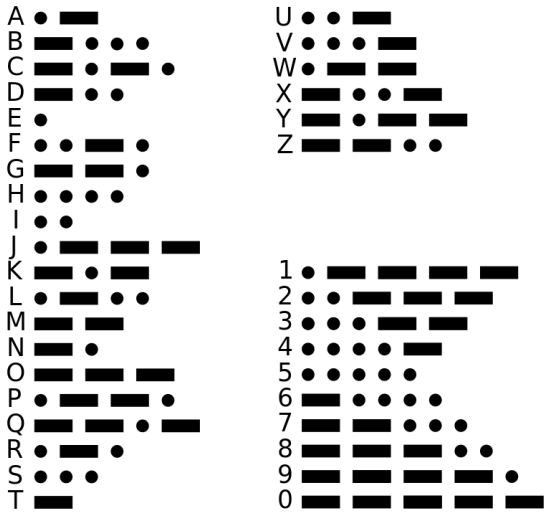
I. INTRODUCTION

ENERAL COMMUNICATIONS system is
hown schematically in Fig. 1. It consists essen-
ially of five elements.
information source. The source selects one mes-
n a set of possible messages to be transmitted to
ving terminal. The message may be of various
r example, a sequence of letters or numbers, as
aphy or teletype, or a continuous function of
, as in radio or telephony.
transmitter. This operates on the message in
y and produces a signal suitable for transmis-
he receiving point over the channel. In teleph-

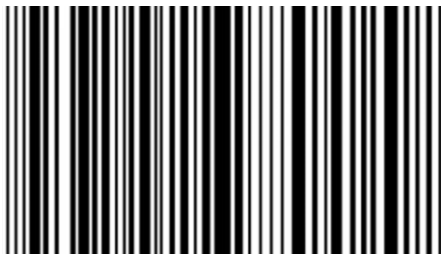


Claude Shannon

Codes

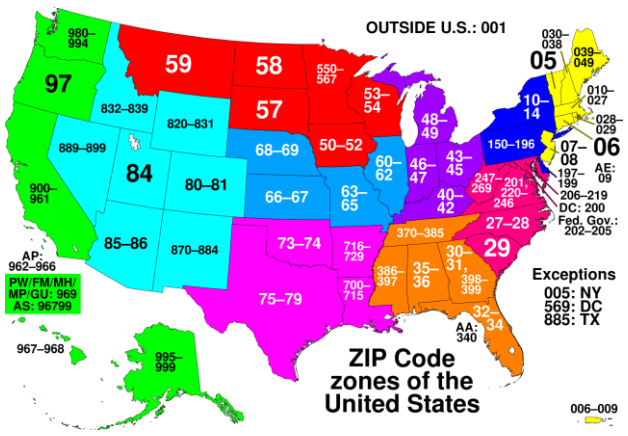


Morse code

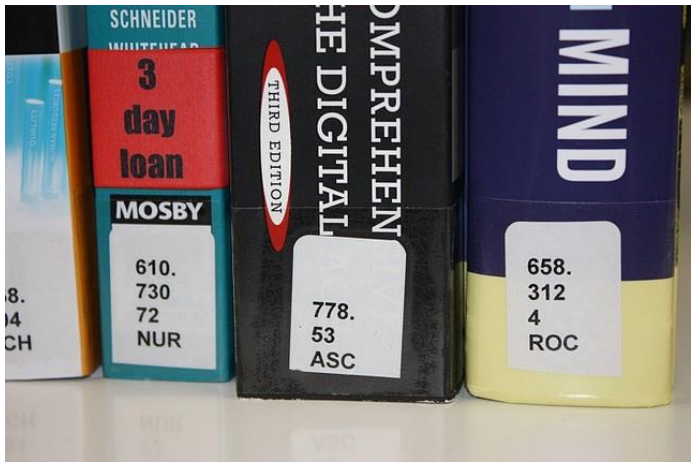


Barcode

QR code



Postal codes



Dewey decimal code

LOL IDK

OMG ROFL

SMS code

UUU, UUC: Phe; UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr; UAA, UAG: Stop	UGU, UGC: Cys; UGA: Stop; UGG: Trp
CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His; CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg
AUU, AUC, AUA: Ile; AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn; AAA, AAG: Lys	AGU, AGC: Ser; AGA, AGG: Arg
GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp; GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly

Genetic code

Genomically Recoded Organisms Expand Biological Functions

Marc J. Lajoie,[1,2] Alexis J. Rovner,[3,4] Daniel B. Goodman,[1,5] Hans-Rudolf Aerni,[4,6] Adrian D. Haimovich,[3,4] Gleb Kuznetsov,[1] Jaron A. Mercer,[7] Harris H. Wang,[8] Peter A. Carr,[9] Joshua A. Mosberg,[1,2] Nadin Rohland,[1] Peter G. Schultz,[10] Joseph M. Jacobson,[11,12] Jesse Rinehart,[4,6] George M. Church,[1,13*] Farren J. Isaacs[3,4*]

Data compression - repetition

- Degree of compression depends on the raw data and algorithm used
  - In general, repetitive data is more compressible

```
raw:          AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
compressed:   30A
```

```
raw:          AAAAACCCCCCGGGGGTTTTTAAAAAAAAA
compressed:   5A6C5G5T9A
```

```
raw:          ACGTGCTAGTACGTCTATGTGCAGTACAGT
compressed:   1A1C1G1T1G1C1T1A1G1T1A1C1G1T1C1T1A1T1G1T1G1C1A1G1T1A1C1A1G1T
```

Repetition Code

Repeat once:

Message	Code word
1010	11001100
1110	11111100

Detect 1-bit errors	Correct 1-bit errors	Detect 2-bit errors
✔	✘	✘

Repeat twice:

Message	Code word
1010	111000111000
1110	111111111000

Detect 1-bit errors	Correct 1-bit errors	Detect 2-bit errors
✔	✔	✘

Parity Check Linear Block Code (MHD2)

$$G = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Message	Code word
1010	01010
1110	11110

Min. distance	Detect 1-bit errors	Correct 1-bit errors	Detect 2-bit errors
2	✔	✘	✘

Hamming Distance

Hamming Distance:

the number of single-letter changes needed to mutate seq A into seq B

CAT **CAR**	**CAT** **BAR**	**CAT** **DOG**
d=1	**d=2**	**d=3**

Hamming Code

$$G = \begin{pmatrix} 1\,0\,0\,0\,1\,1\,0 \\ 0\,1\,0\,0\,0\,1\,1 \\ 0\,0\,1\,0\,1\,1\,1 \\ 0\,0\,0\,1\,1\,0\,1 \end{pmatrix}$$

Message	Code word
1010	0011010
1110	0101110

Min. distance	Detect 1-bit errors	Correct 1-bit errors	Detect 2-bit errors
3	✔	✔	✘

Modified Hamming Distance 4 Code (MHD4)

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Message	Code word
1010	11001010
1110	00101110

Min. distance	Detect 1-bit errors	Correct 1-bit errors	Detect 2-bit errors
4	✔	✔	✔

Modified Hamming Distance 4 Code (MHD4)

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Message	Code word
1010	11001010
1110	00101110

Min. distance	Detect 1-bit errors	Correct 1-bit errors	Detect 2-bit errors
4	✔	✔	✔

Codeword Validation and Decoding (MHD4)

$$\mathbf{H} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Example	Codeword	Parity check	Action
Valid	11001010	0000	Decode
1-bit error	11101010	0010	Correct error, then decode
2-bit error	10101010	0110	Error, mark invalid

Linear Block Code Summary

Message	MHD2	Hamming	MHD4
1010	01010	0011010	11001010
1110	11110	0101110	00101110

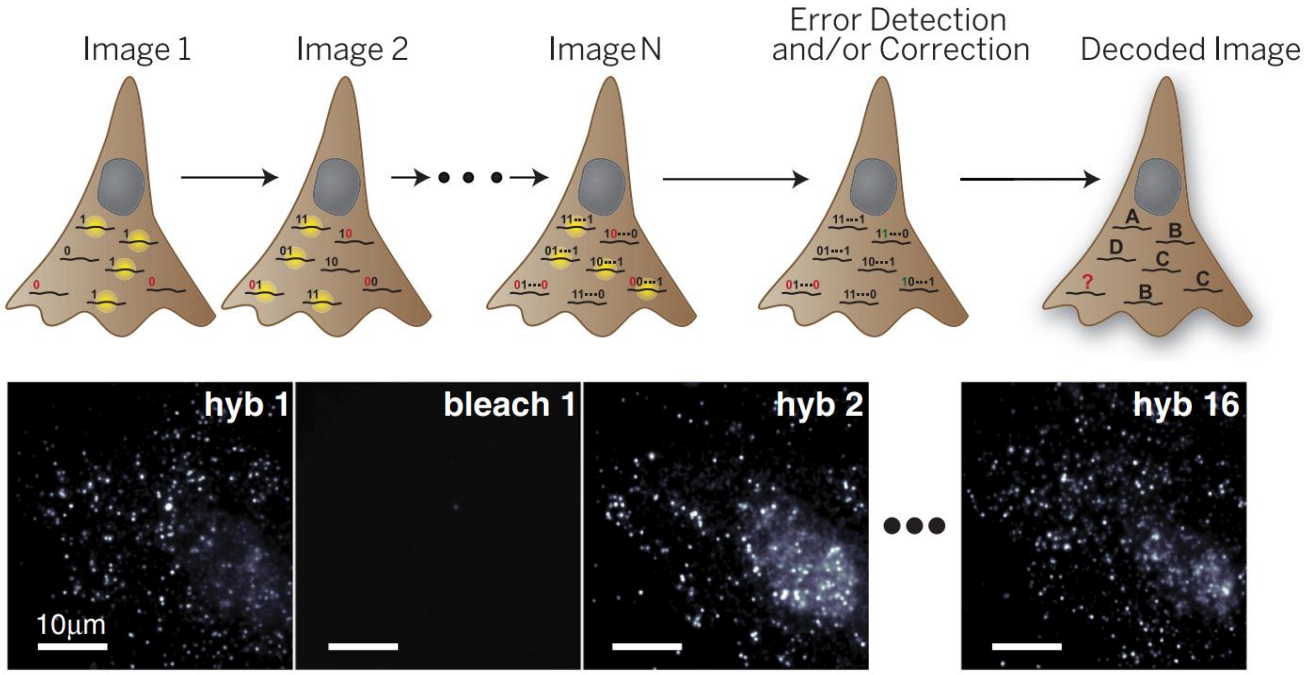
Code	Min. distance	Detect 1-bit errors	Correct 1-bit errors	Detect 2-bit errors
MHD4	4	✓	✓	✓
Hamming	3	✓	✓	✗
MHD2	2	✓	✗	✗

Error-correcting Codes and FISH

Multiplexed error-robust FISH (MERFISH)

Spatially resolved, highly multiplexed RNA profiling in single cells

Kok Hao Chen,[1*] Alistair N. Boettiger,[1*] Jeffrey R. Moffitt,[1*]
Siyuan Wang,[1] Xiaowei Zhuang[1,2†]



"In principle, combinatorial labeling allows the number of detectable RNA species to grow exponentially with the number of imaging rounds, but the detection errors also increase exponentially. To combat such accumulating errors, we exploited error-robust encoding schemes used in digital electronics, such as the extended Hamming code, in the design of our encoding probes but modified these schemes in order to account for the error properties in FISH measurements."

MHD4 Code

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

MERFISH theory critique

A Coding Theory Perspective on Multiplexed Molecular Profiling of Biological Tissues

Luca D'Alessio* — *Broad Institute, Cambridge, MA* — ldalessi@broadinstitute.org

Litian Liu* — *MIT, Cambridge, MA* — litianl@mit.edu

Ken Duffy — *Maynooth University, Ireland* — ken.duffy@nuim.ie

Yonina C. Eldar — *Weizmann Institute of Science, Israel* — yonina.eldar@weizmann.ac.il

Muriel Médard — *MIT, Cambridge, MA* — medard@mit.edu

Mehrtash Babadi — *Broad Institute, Cambridge, MA* — mehrtash@broadinstitute.org

We point out that the assumptions motivating the codebook construction and decoding, tacitly yet heavily, rely on source uniformity and to a certain extent on the binary symmetric channel paradigm, both of which are violated in the context molecular profiling. For channel coding in communication, source can be readily assumed as uniformly distributed thanks to compression in source coding and the separation theorem [6]. In molecular profiling, however, source compression is not applicable and the distribution of RNA molecules is...

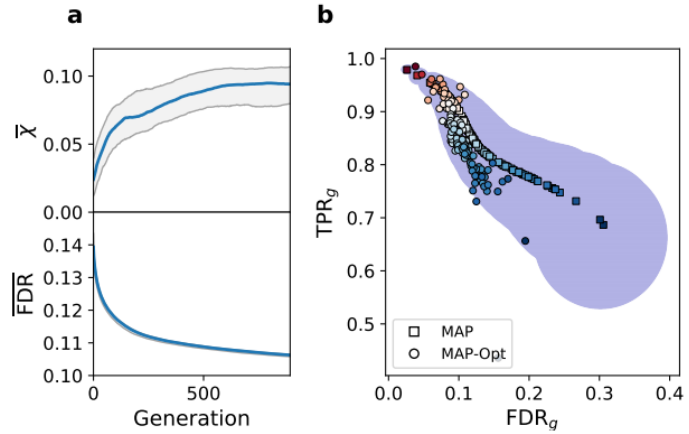
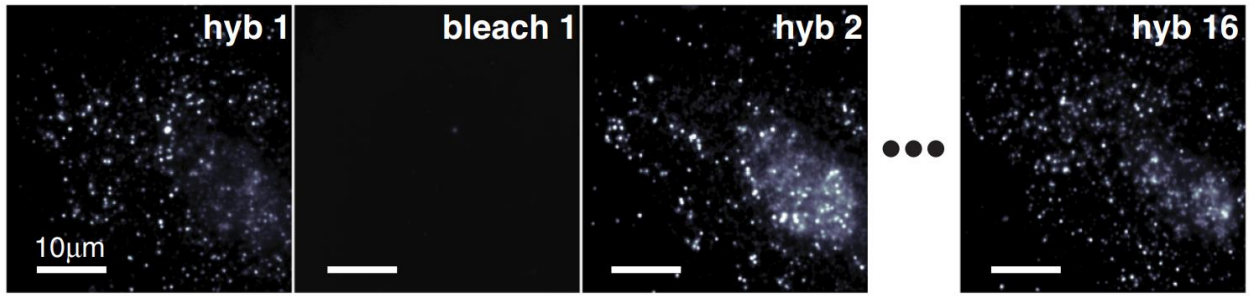


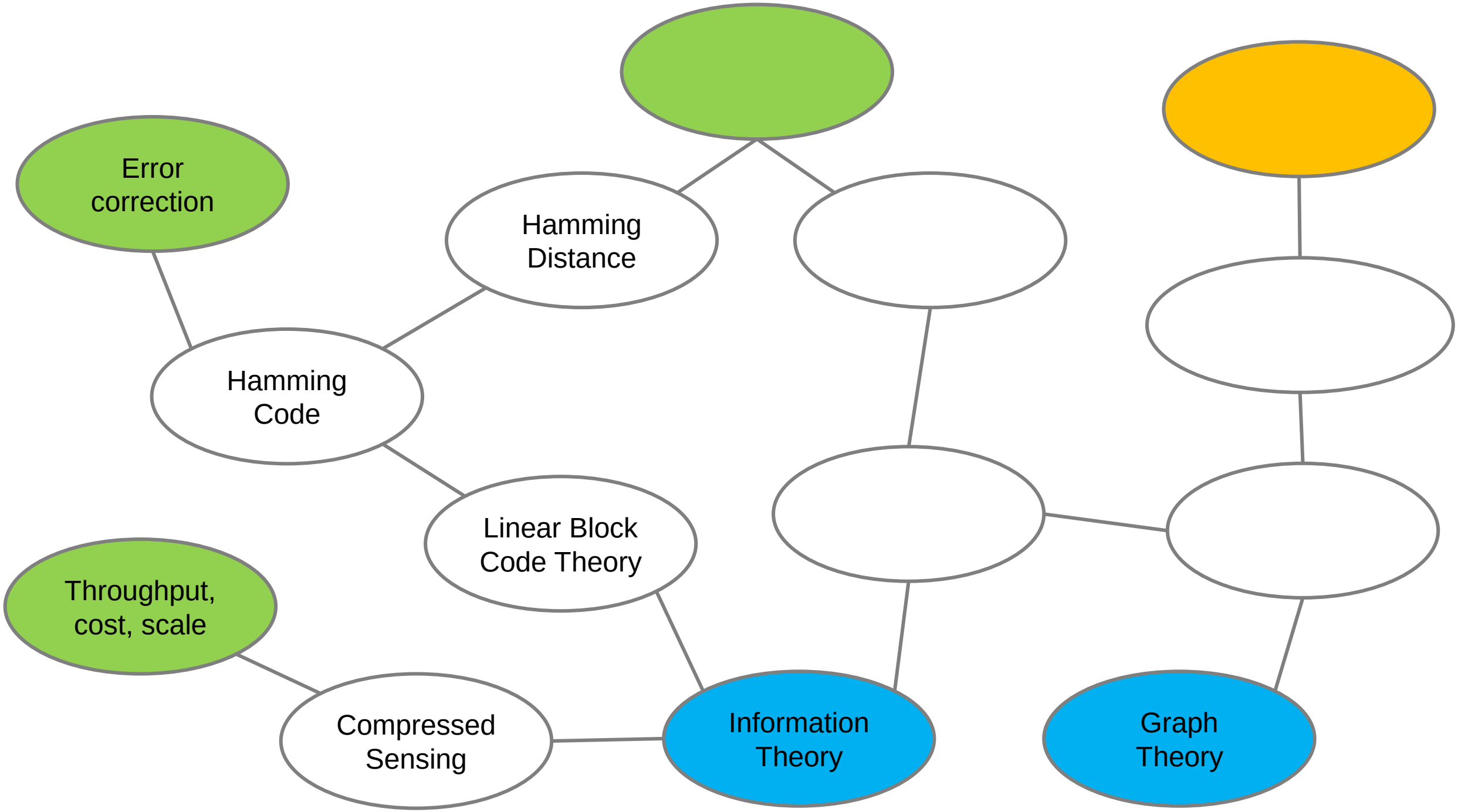
Fig. 5. Evolutionary optimization of code assignment for MHD4 codes (for channel model described in Fig. 3 and prior distribution from [3]). (a) bottom: mean FDR vs. generation; top: $d_{\mathrm{H}} - d_{\boldsymbol{\pi}}$ matching order parameter vs. generation (see Eq. 5); (b) the performance of MAP decoder for randomly assigned codes (squares) vs. optimized assignment (circles).

Hamming Weight

Code word	Hamming Weight
00001100	2
00110011	4
00111111	6
000000001100	2
000000110011	4

"each of the 140 possible barcodes has a constant Hamming weight (i.e., the number of 1 bits in each barcode) of 4 to avoid potential bias in the measurement of different barcodes due to a differential rate of 1 to 0 and 0 to 1 errors."




Error correction
Hamming Distance
Hamming Code
Linear Block Code Theory
Throughput, cost, scale
Compressed Sensing
Information Theory
Graph Theory

Outline

- Fluorescence in situ hybridization (FISH)
- Information theory & FISH
  - Compressed sensing
  - Linear block theory, error-correction
- de Bruijn sequences
- Orthogonal DNA sequence set design
  - k-mer symmetry minimization
  - Hamming distance approaches
- New algorithm idea…

Brute force attacks on digital locks


```
0000
0001
0002
…
2220
2221
2222
```

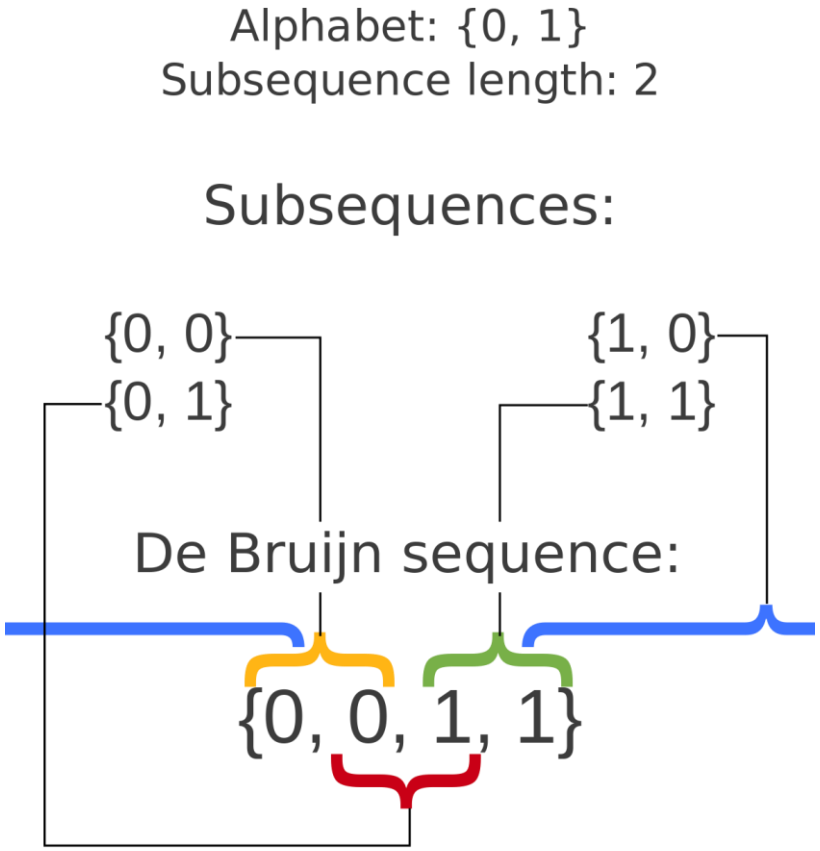
→

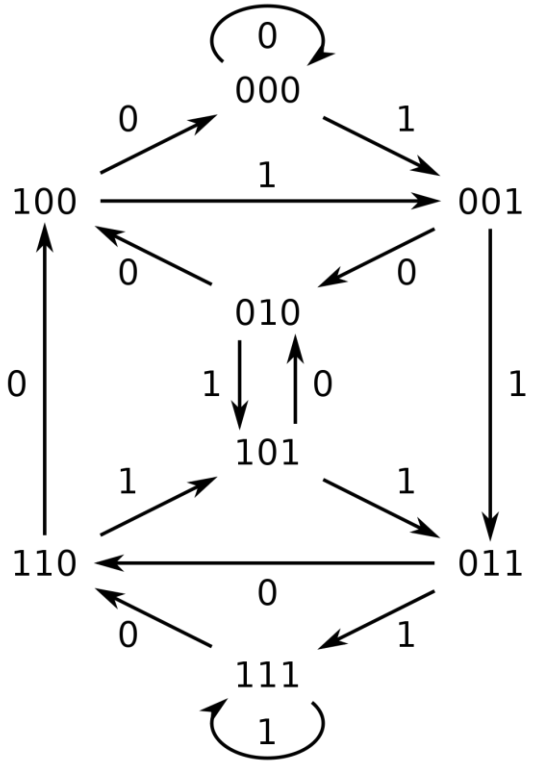
```
000000010002001000110012002000210022010001
010102011001110112012001210122020002010202
021002110212022002210222100010011002101010
111012102010211022110011011102111011111112
112011211122120012011202121012111212122012
211222200020012002201020112012202020212022
210021012102211021112112212021212122220022
012202221022112212222022212222
```

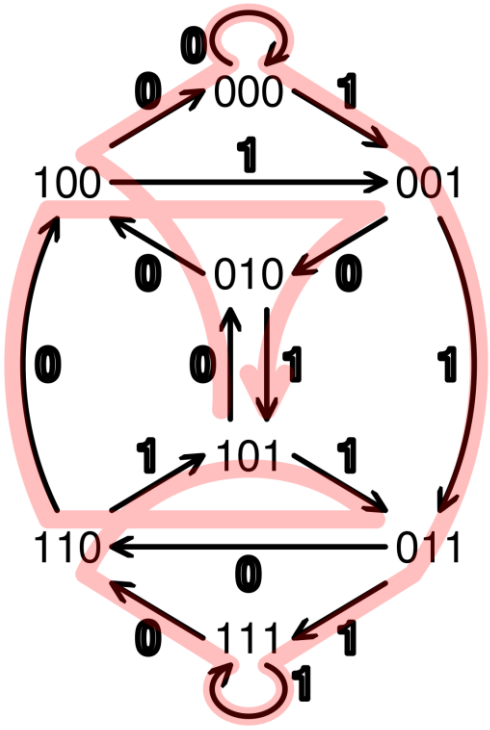
length: 324

https://www.johndcook.com/blog/2019/10/22/hacking-with-de-bruijn/

de Bruijn sequences







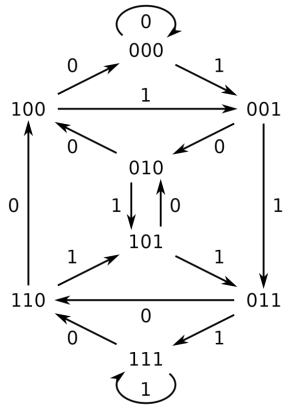
https://en.wikipedia.org/wiki/De_Bruijn_sequence

de Bruijn hacking digital locks

```
0000
0001
0002
…
2220
2221
2222
```

→

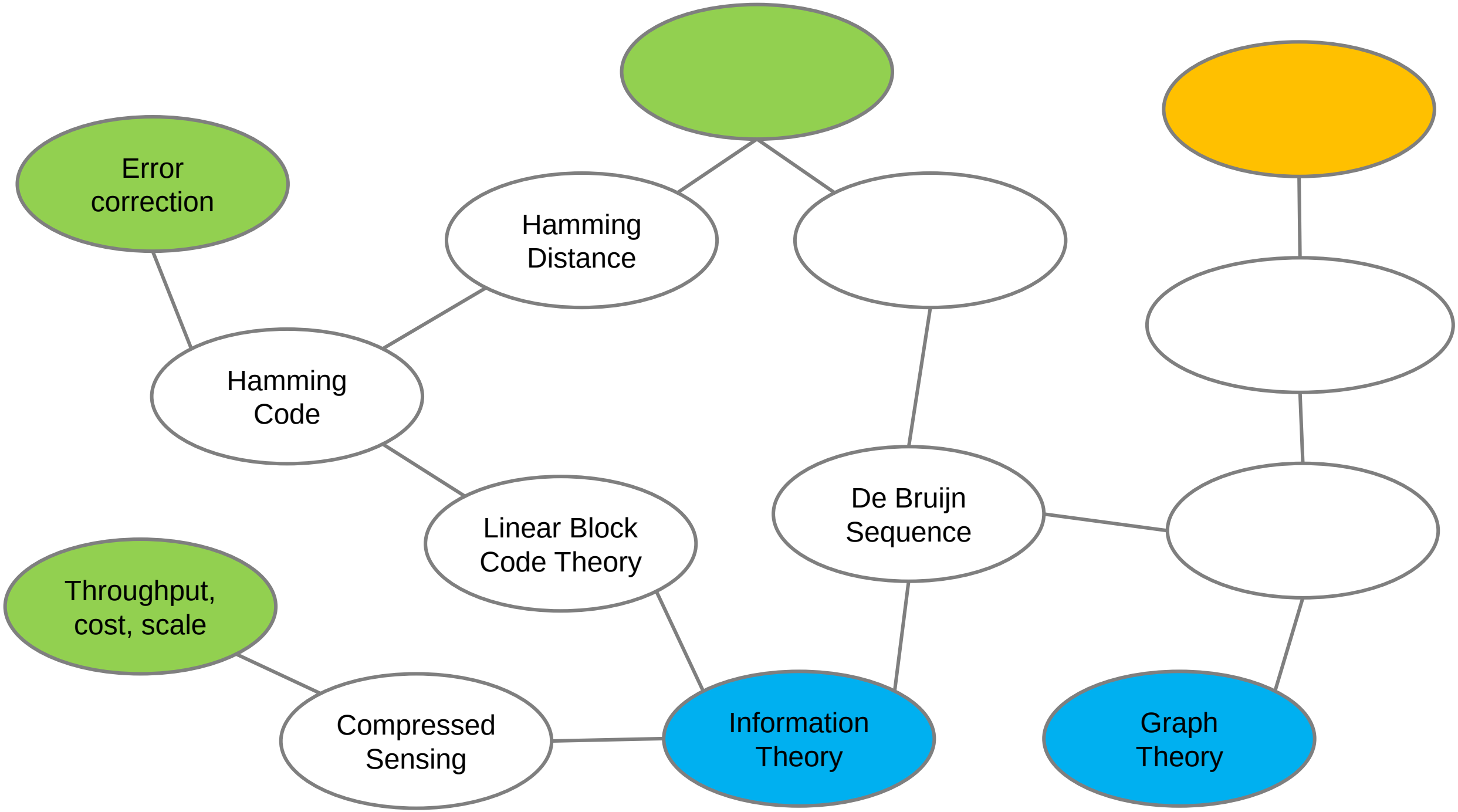
```
000000010002001000110012002000210022010001
010102011001110112012001210122020002010202
021002110212022002210222100010011002101010
111012102010211022110011011102111011111112
112011211122120012011202121012111212122012
211222200020012002201020112012202020212022
210021012102211021112112212021212122220022
012202221022112212222022212222
      length: 324
```



→

```
000010011012110021020122101011112220112120
002002202122001201021120202221110221 21
      length: 81
```

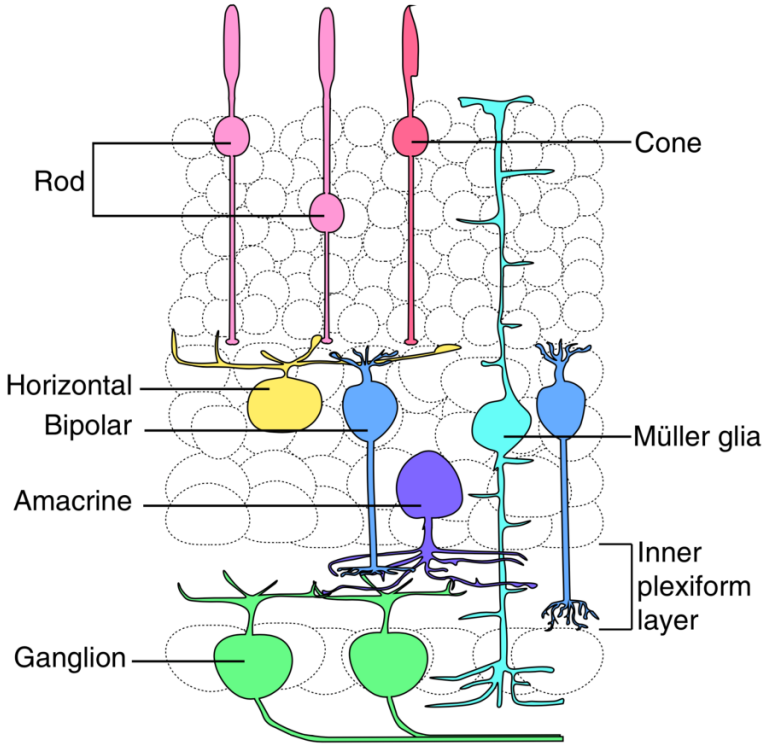
https://www.johndcook.com/blog/2019/10/22/hacking-with-de-bruijn/


Error correction
Hamming Distance
Hamming Code
Linear Block Code Theory
Throughput, cost, scale
Compressed Sensing
Information Theory
De Bruijn Sequence
Graph Theory

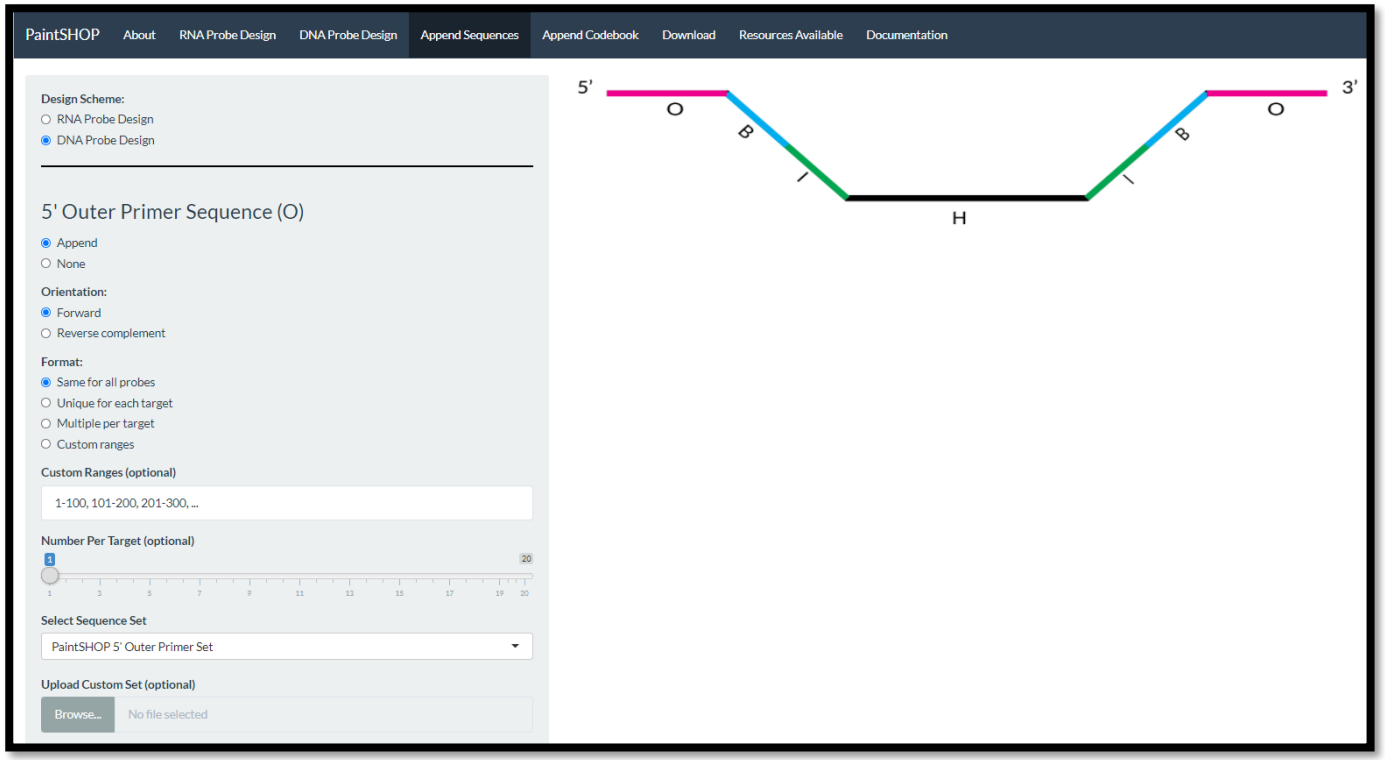
Outline

- Fluorescence in situ hybridization (FISH)
- Information theory & FISH
  - Compressed sensing
  - Linear block theory, error-correction
- de Bruijn sequences
- Orthogonal DNA sequence set design
  - k-mer symmetry minimization
  - Hamming distance approaches
- New algorithm idea…

PaintSHOP simplifies complex probe design



web server: paintshop.io

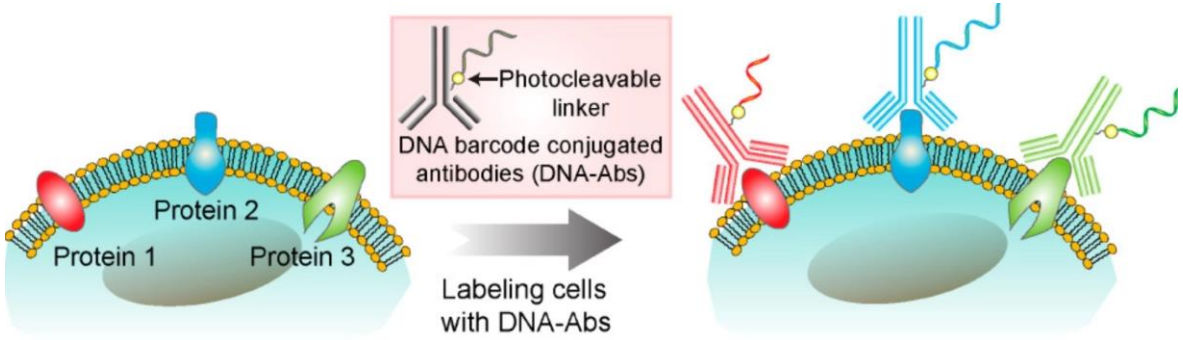


Features for multiplexing, barcoding probes

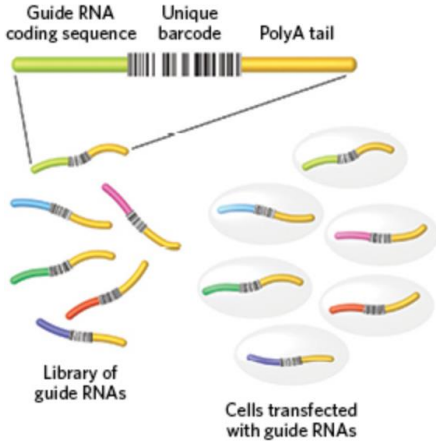
Kishi, J.Y., Lapan, S.W., Beliveau, B.J. et al. *Nat Methods* **16**, 533–544 (2019)

DNA Barcodes

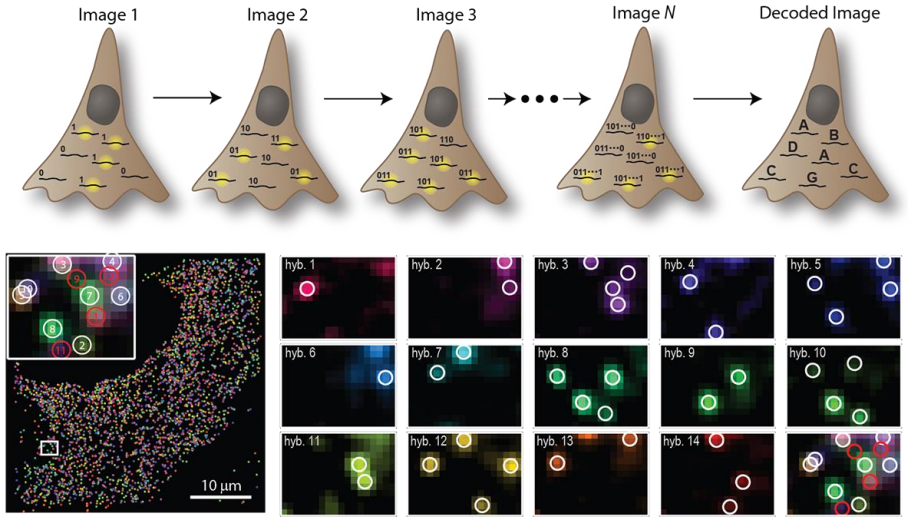
DNA-barcoded antibodies



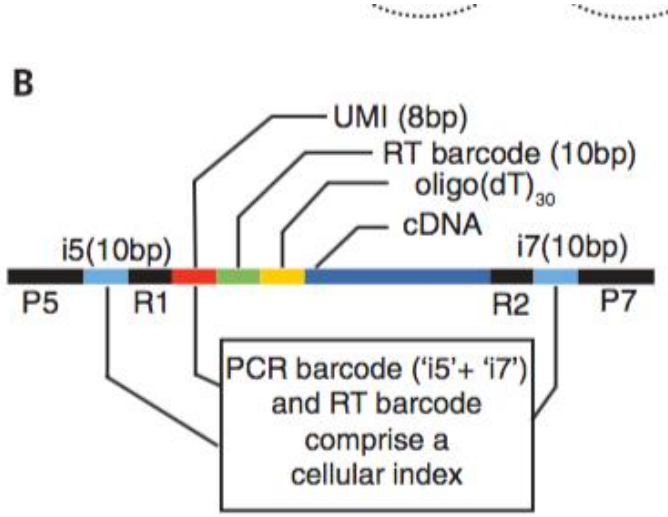
CRISPR screens



Multiplexed FISH

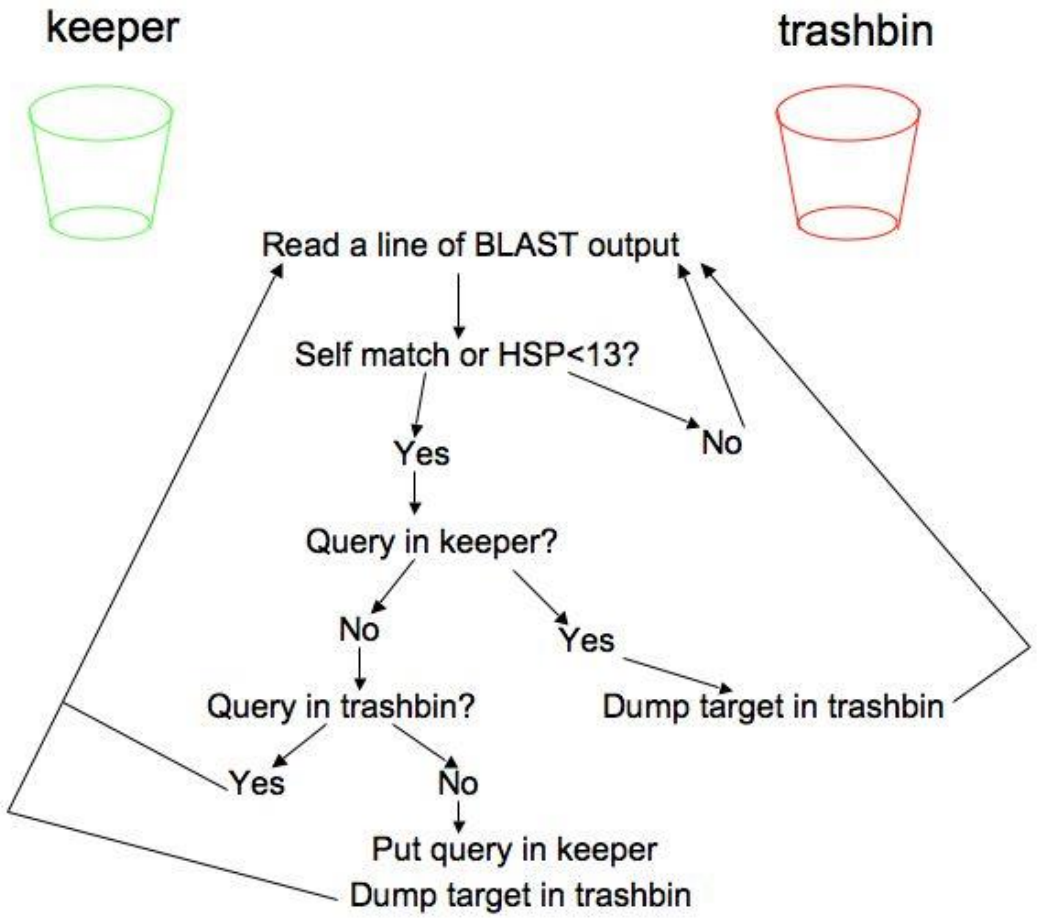


sci-RNA-seq



Design of 240,000 orthogonal 25mer DNA barcode probes

Qikai Xu[a], Michael R. Schlabach[a], Gregory J. Hannon[b], and Stephen J. Elledge[a,1]



Qikai Xu, et al., P.N.A.S., (2009)

Selected citations with bc25mer barcodes

SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues

Jocelyn Y. Kishi, Sylvain W. Lapan, Brian J. Beliveau, Emma R. West, Allen Zhu, Hiroshi M. Sasaki, Sinem K Saka, Yu Wang, Constance L. Cepko & Peng Yin

Nature Methods **16**, 533–544 (2019) | Cite this article

Article | Open Access | Published: 10 January 2022

Spatial transcriptomics using combinatorial fluorescence spectral and lifetime encoding, imaging and analysis

Tam Vu, Alexander Vallmitjana, Joshua Gu, Kieu La, Qi Xu, Jesus Flores, Jan Zimak, Jessica Shiu, Linzi Hosohama, Jie Wu, Christopher Douglas, Marian L. Waterman, Anand Ganesan, Per Niklas Hedde, Enrico Gratton & Weian Zhao

Open Access | Published: 12 May 2015

Single-molecule super-resolution imaging of chromosomes and *in situ* haplotype visualization using Oligopaint FISH probes

Brian J. Beliveau, Alistair N. Boettiger, Maier S. Avendaño, Ralf Jungmann, Ruth B. McCole, Eric F. Joyce, Caroline Kim-Kiselak, Frédéric Bantignies, Chamith Y. Fonseka, Jelena Erceg, Mohammed A. Hannan, Hien G. Hoang, David Colognori, Jeannie T. Lee, William M. Shih, Peng Yin, Xiaowei Zhuang & Chao-ting Wu

Protocol | Published: 24 October 2022

HT-smFISH: a cost-effective and flexible workflow for high-throughput single-molecule RNA imaging

Adham Safieddine, Emeline Coleno, Frederic Lionneton, Abdel-Meneem Traboulsi, Soha Salloum, Charles-Henri Lecellier, Thierry Gostan, Virginie Georget, Cédric Hassen-Khodja, Arthur Imbert, Florian Mueller, Thomas Walter, Marion Peter & Edouard Bertrand

Nature Protocols (2022) | Cite this article

Article | Open Access | Published: 22 May 2019

Multiplexed detection of RNA using MERFISH and branched DNA amplification

Chenglong Xia, Hazen P. Babcock, Jeffrey R. Moffitt & Xiaowei Zhuang

Scientific Reports **9**, Article number: 7721 (2019) | Cite this article

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

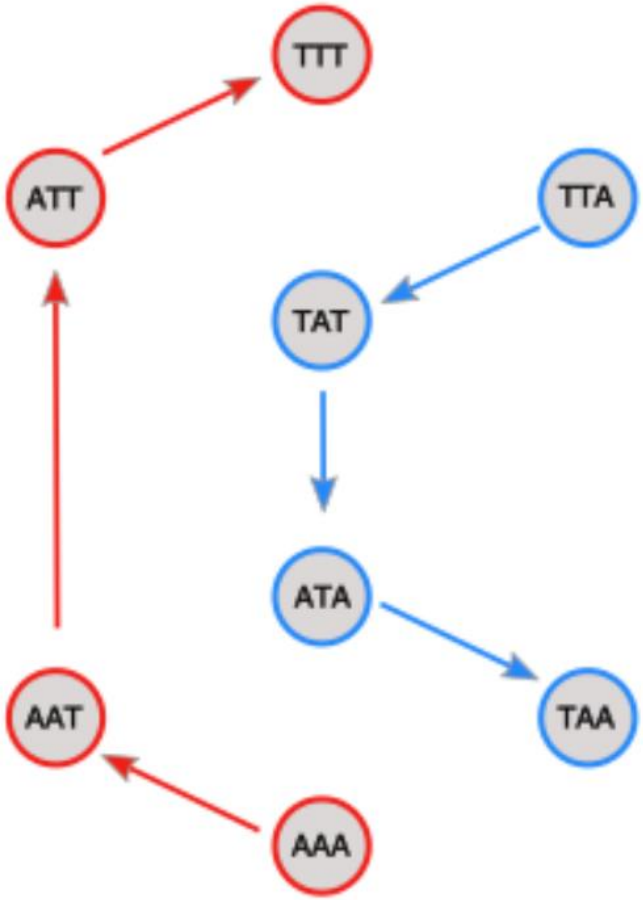
Efficient algorithms for designing maximally sized orthogonal DNA sequence libraries

Gokul Gowri, Kuanwei Sheng, Peng Yin

doi: https://doi.org/10.1101/2022.07.11.499592

Abstract

Orthogonal sequence library design is an essential task in bioengineering. Typical design approaches scale quadratically in the size of the candidate sequence space. As such, exhaustive searches of sequence space to maximize library size are computationally intractable with existing methods. Here, we present SeqWalk, a time and memory efficient method for designing maximally-sized orthogonal sequence libraries using the sequence symmetry minimization heuristic. SeqWalk encodes sequence design constraints in a de Bruijn graph representation of sequence space, enabling the application of efficient graph traversal techniques to the problem of orthogonal DNA sequence design. We demonstrate the scalability of SeqWalk by designing a provably maximal set of > 10^6 orthogonal 25nt sequences in less than 20 seconds on a single standard CPU core. We additionally derive fundamental bounds on orthogonal sequence library size under a variety of design constraints.



DNA k-mers

seq: ACGTAAACCCGGGTTT

k = 12

```
ACGTAAACCCGGGTTT
ACGTAAACCCGG
 CGTAAACCCGGG
  GTAAACCCGGGT
   TAAACCCGGGTT
    AAACCCGGGTTT
```

k = 8

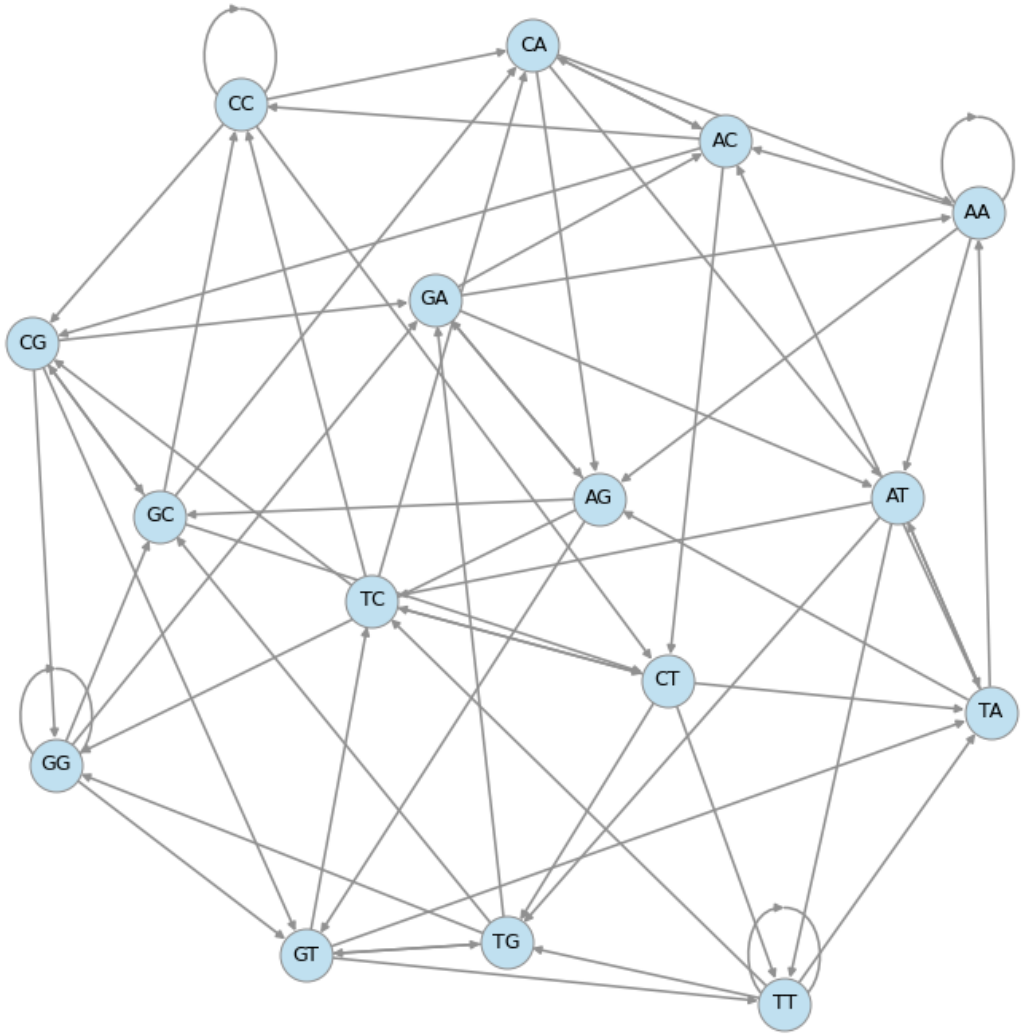
```
ACGTAAACCCGGGTTT
ACGTAAAC
 CGTAAACC
  GTAAACCC
   TAAACCCG
    AAACCCGG
     AACCCGGG
      ACCCGGGT
       CCCGGGTT
        CCGGGTTT
```

k = 3

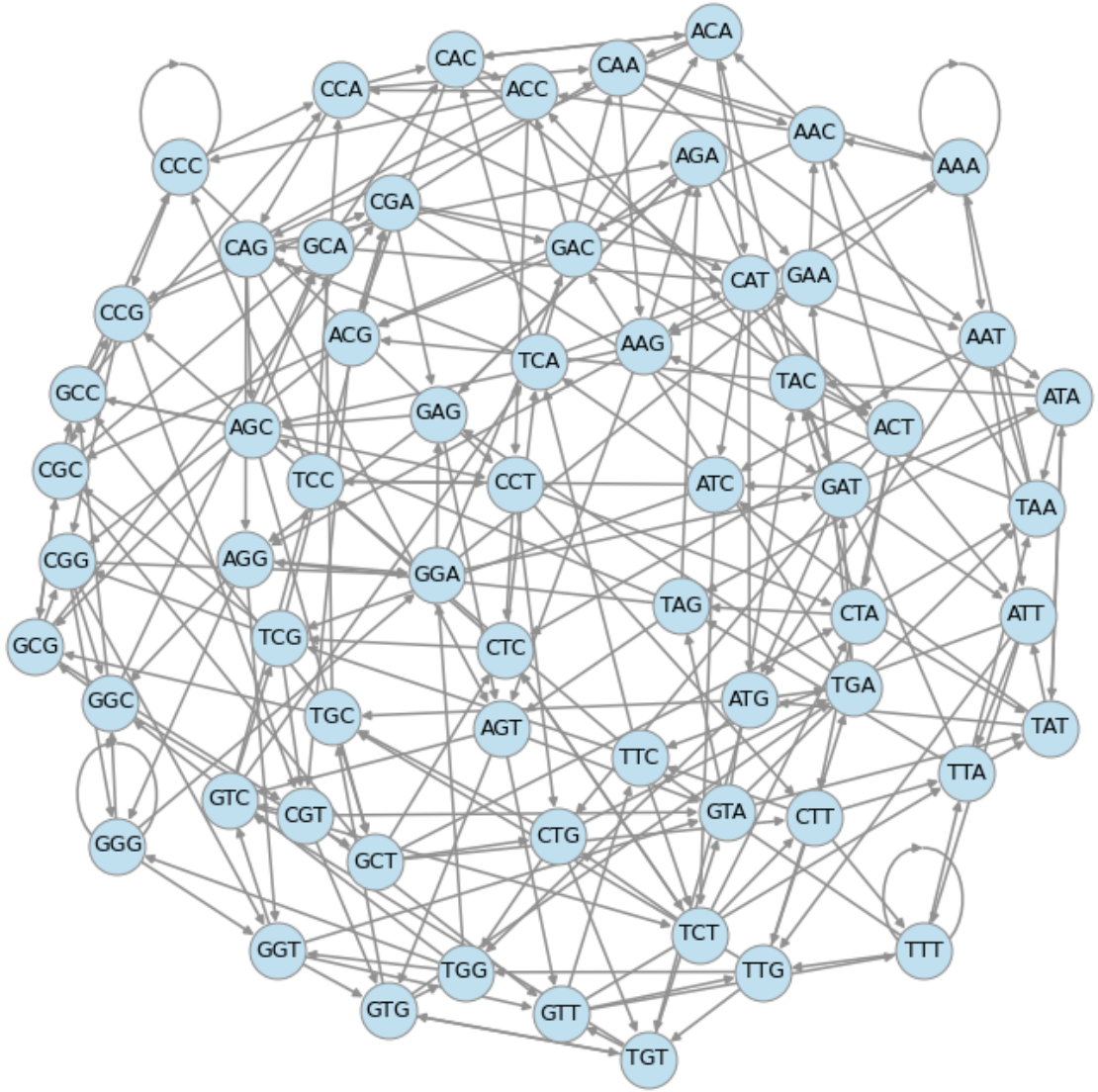
```
ACGTAAACCCGGGTTT
ACG
 CGT
  GTA
   TAA
    AAA
     AAC
      ACC
       CCC
        CCG
         CGG
          GGG
           GGT
            GTT
             TTT
```

DNA k-mer De Bruijn Graphs

k = 2



k = 3

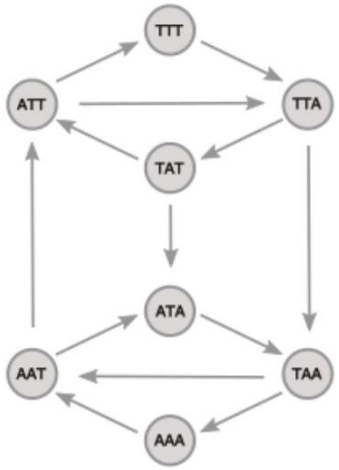


k-mer graph algorithm

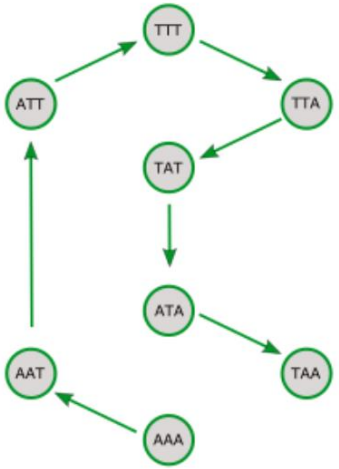
(1) select parameters

L = 6
k = 3
a = {A, T}

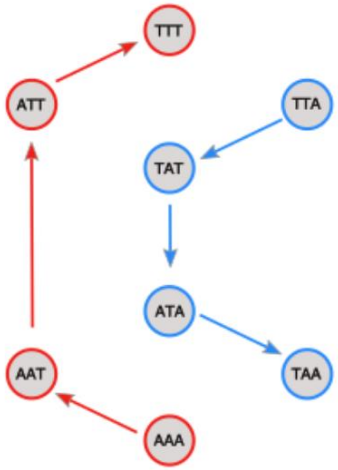
(2) make k-mer graph



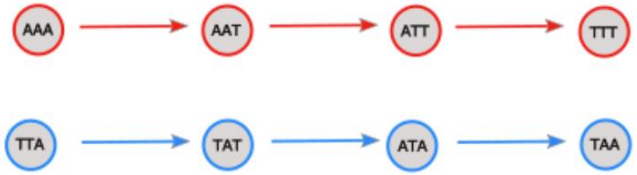
(3) find Hamiltonian path



(4) partition path



(5) map to sequences



seq 1: AAATTT
seq 2: TTATAA

SeqWalk on readthedocs.io

https://seqwalk.readthedocs.io/en/latest/index.html

seqwalk

latest

Search docs

Example usage

Changelog

Contributing

Code of Conduct

API Reference

» seqwalk

Edit on GitHub

seqwalk

`seqwalk` is a package for designing orthogonal DNA sequence libraries. If you want to design DNA barcodes (for sequencing, multiplexed imaging, molecular programming, etc.) `seqwalk` is for you! It can efficiently generate libraries of maximal size or maximal predicted orthogonality based on sequence symmetry. `seqwalk` additionally includes off-the-shelf orthogonal sequence libraries, as well as tools for analyzing orthogonal sequence libraries.

A code-free, interactive version of `seqwalk` can be found here

For more details, see our preprint (coming soon!).

Installation

```
$ pip install seqwalk
```

Usage

Designing a set of barcodes with maximum orthogonality

Hamming Graph Algorithm

Hamming Distance:

the number of single-letter changes needed to mutate seq A into seq B

CAT **CAR**	**CAT** **BAR**	**CAT** **DOG**
d=1	**d=2**	**d=3**

Goal:

given a set of k-length sequences, return a maximally large subset with min. Hamming distance = 2

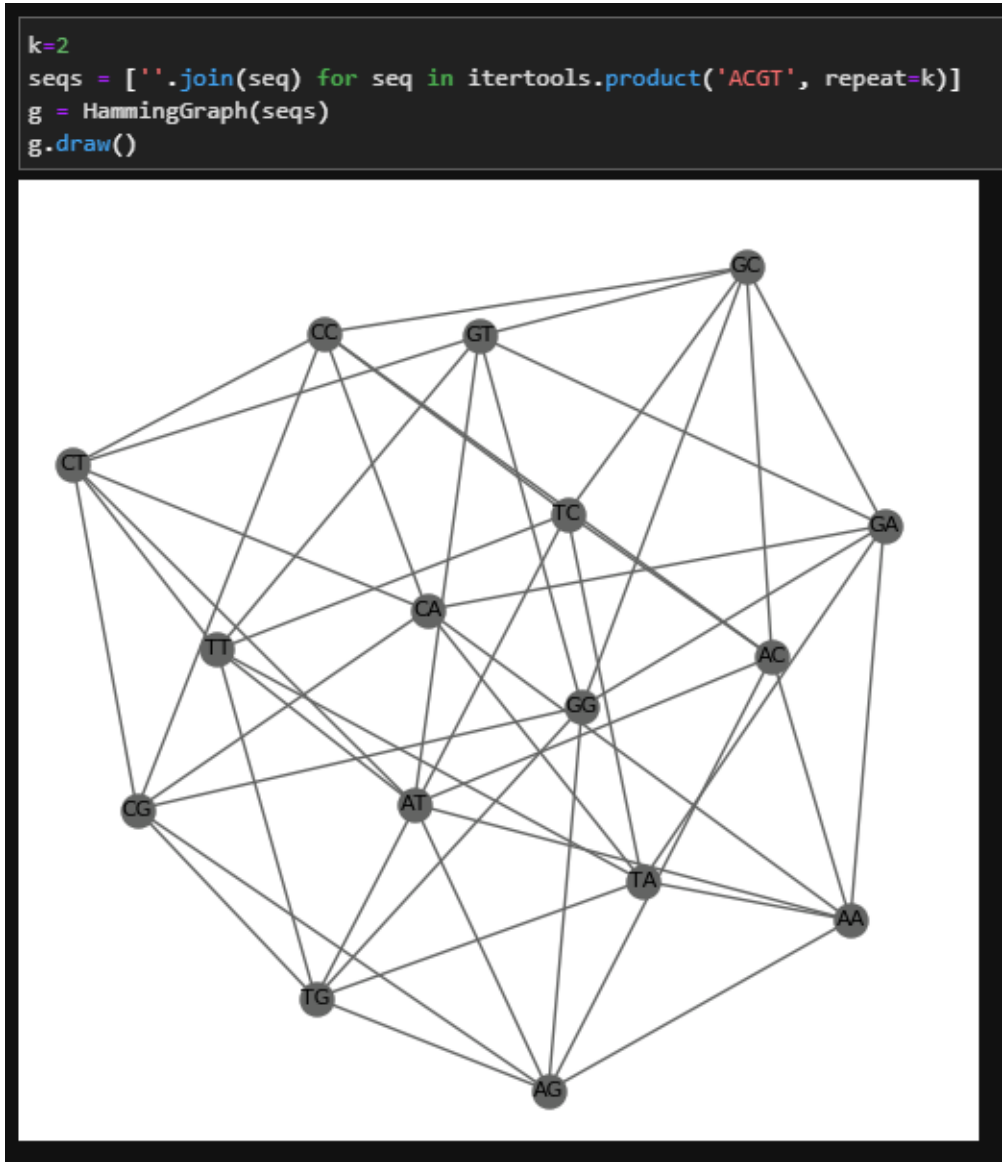
Hamming Graph Algorithm

Hamming Graph:

Create a node for each sequence

Create an edge connecting d=1 sequences

```
k=2
seqs = [''.join(seq) for seq in itertools.product('ACGT', repeat=k)]
g = HammingGraph(seqs)
g.draw()
```



Hamming Graph Algorithm

Hamming Graph:

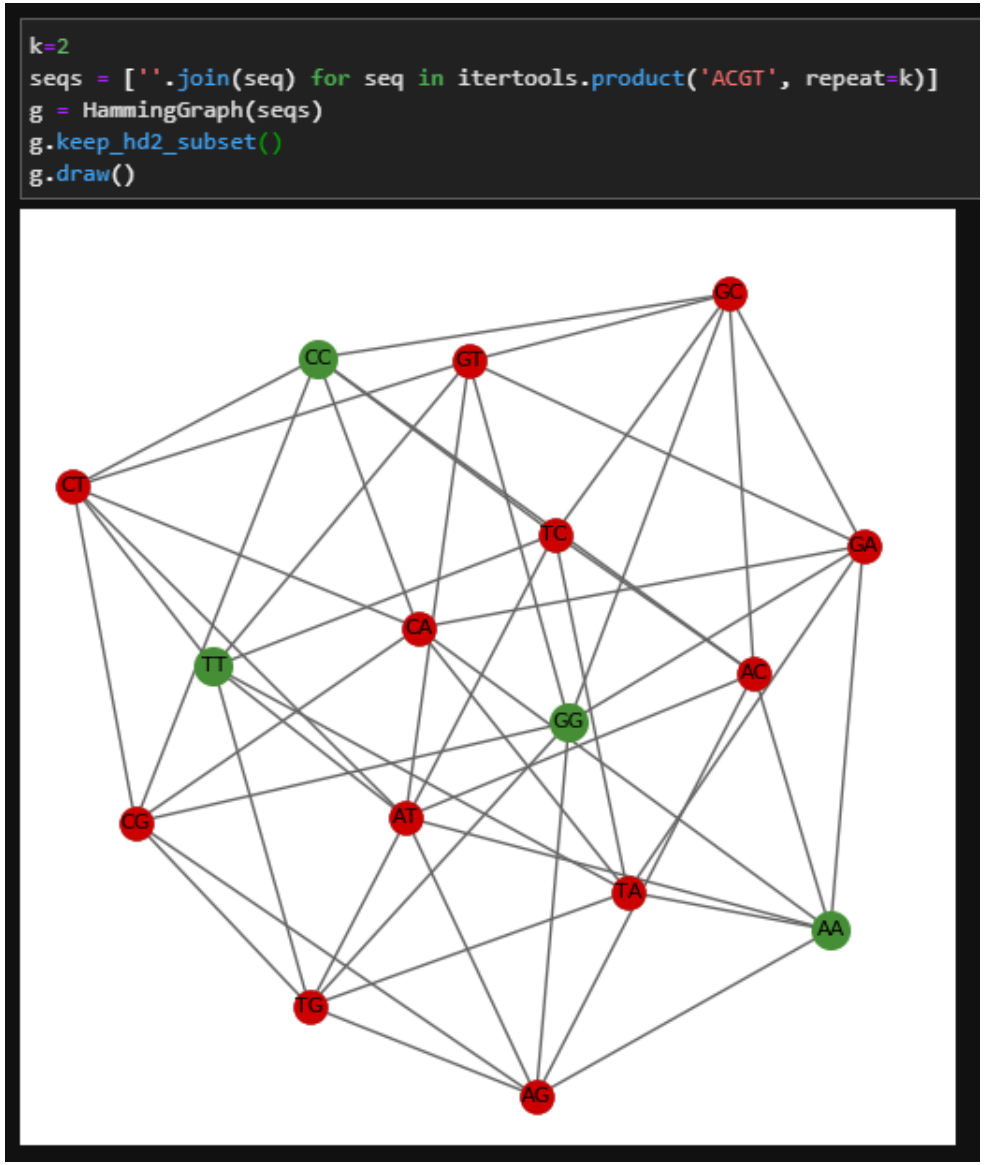
Create a node for each sequence

Create an edge connecting d=1 sequences

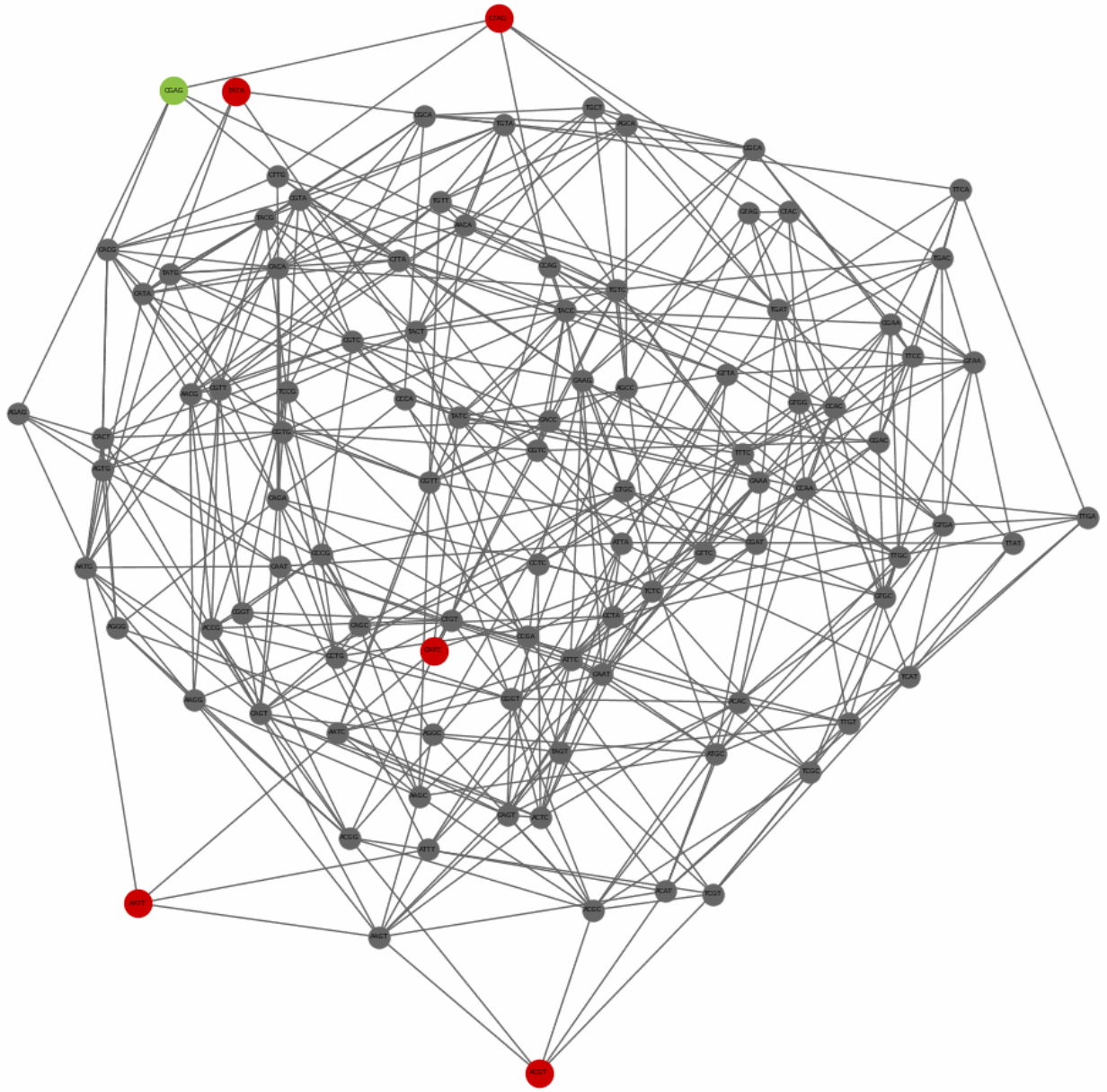
Goal:

given a set of k-length sequences, return a maximally large subset with min. Hamming distance = 2

```
k=2
seqs = [''.join(seq) for seq in itertools.product('ACGT', repeat=k)]
g = HammingGraph(seqs)
g.keep_hd2_subset()
g.draw()
```



Hamming Graph Algorithm



Hamming vs Levenshtein Distance

compare and contrast the utility of using the Hamming distance or Levenshtein distance between two oligonucleotide sequences to predict whether they will hybridize to the same target sequence

The Hamming distance and Levenshtein distance are two measures of the difference between two strings, including two oligonucleotide sequences. The Hamming distance is defined as the number of positions at which two sequences differ, while the Levenshtein distance considers not only the differences between two sequences but also allows for insertions and deletions.

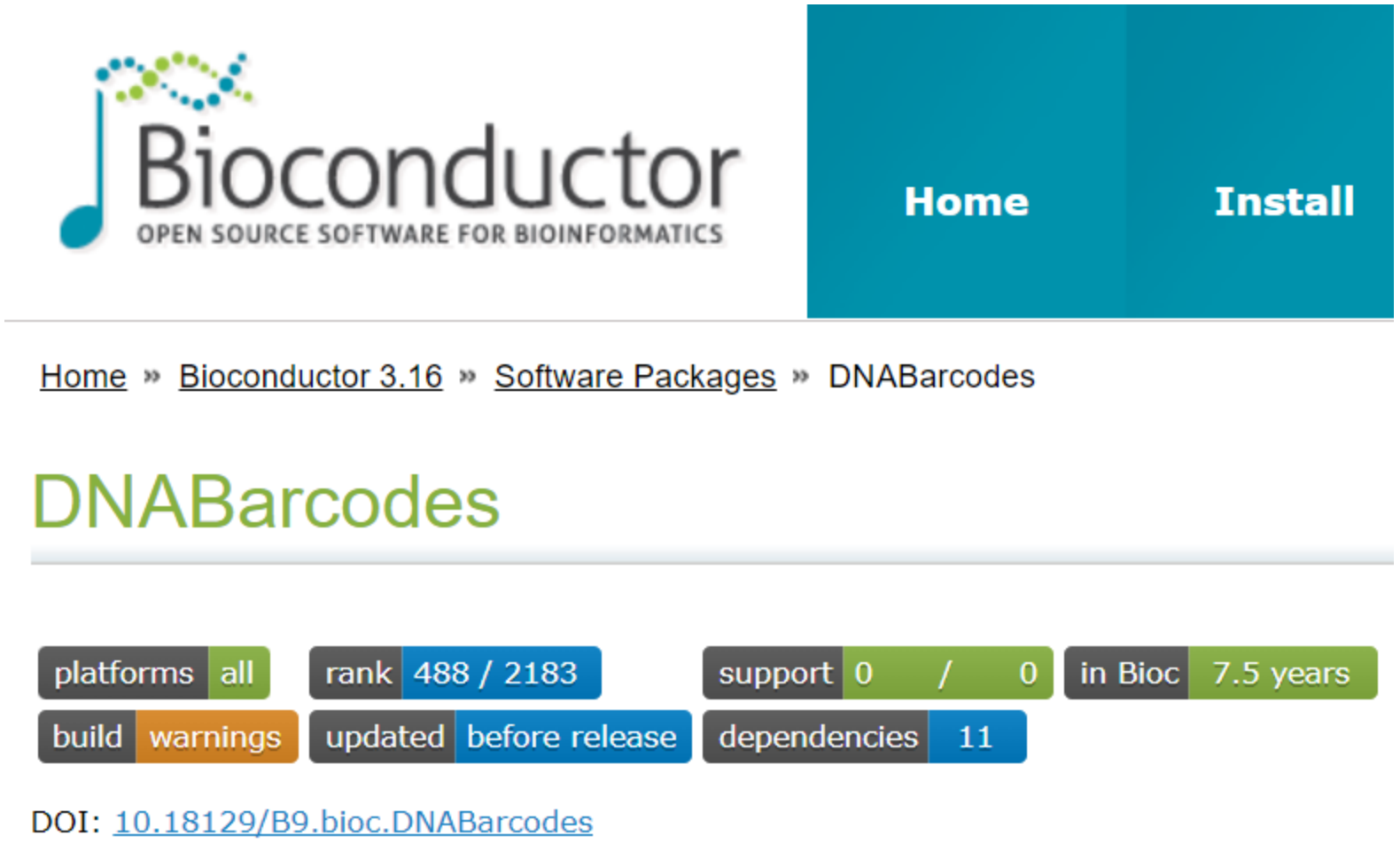
In terms of predicting whether two oligonucleotide sequences will hybridize to the same target sequence, the utility of using either the Hamming distance or the Levenshtein distance depends on the specific goals and requirements of the experiment.

The Hamming distance is a simple and straightforward measure that is useful for determining whether two sequences are similar or not, but it only takes into account the number of differences between two sequences, and not the actual location or type of those differences. This can be a disadvantage when it comes to hybridization, as a single difference in the right location may have a big impact on hybridization.

The Levenshtein distance is a more sophisticated measure that takes into account not only the differences between two sequences but also allows for the possibility of insertions and deletions. This can be a more accurate measure for predicting hybridization as it takes into account more information about the sequences and how they may interact.

chatGPT

dna.barcodes R Package



https://rdrr.io/bioc/DNABarcodes/man/create.dnabarcodes.html

dna.barcodes R Package

The heuristics for the generation of barcodes are as follows:

The Conway heuristic (`heuristic = "conway"`, named after John Conway) starts with an empty set of barcodes, goes through the list of candidate barcodes (the `pool`) in lexicographical order and adds each candidate barcode to the initial set if the distance if the candidate barcode to each barcode in the intial set is at least `d >= dist`.

The Clique heuristic (`heuristic = "conway"`) first generates a graph representation of the pool. Each barcode in the pool is a node of the graph and two barcodes/nodes are connected undirectionally if their distance is at least `d >= dist`. The barcode set problem is now reduced to finding the maximal clique in this graph. Because that problem is also computationally infeasible, we use the heuristic clique algorithm of Pattabiraman et al.

The sampling heuristic (`heuristic = "sampling"`) extends the principle of the Conway heuristic. Instead of starting with an empty initial set, we generate small random sets of barcodes as initial sets (the so called seeds). Those seeds are then "closed" using the Conway method. The size of the seed is fixed to three barcodes. The number of random seeds is given by the parameter `iterations`, hence that often a Conway closure is calculated.

Finally, the Ashlock heuristic (`heuristic = "ashlock"`, named after Daniel Ashlock) extends the sampling heuristic by adding an evolutionary algorithm. A population of random seeds is generated only for the first iteration. Each seed is then closed using the Conway method. The size of the barcode set after closure defines the *fitness* of that seed. For the next iteration, succesful seeds (with a higher fitness) are cloned and slightly mutated (some barcodes in the seed are replaced with a random new barcode). Those changed seeds are now closed again and their respective fitness calculated. In the first iteration, as many instances of the Conway closure are calculated as there are seeds. In the second round, only half of the seeds (the changed ones) are calculated. Therefore, the total number of calculated Conway closures is `iterations + population/2 * (iterations - 1)`.

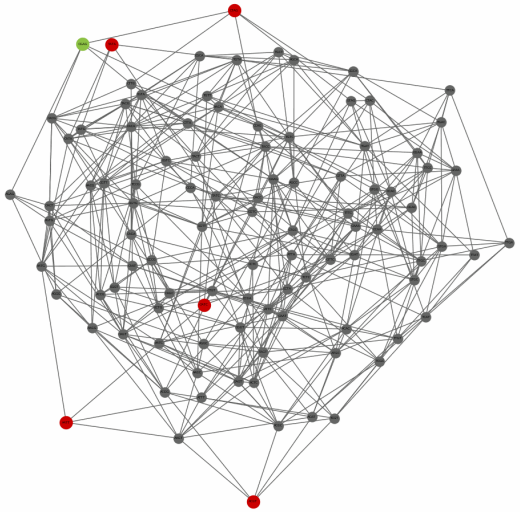
https://rdrr.io/bioc/DNABarcodes/man/create.dnabarcodes.html

Orthogonal DNA Set Design

Hamming distance

Min. k-mer symmetry

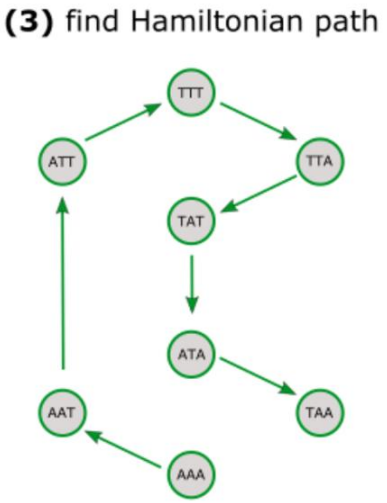
Heuristic approaches:

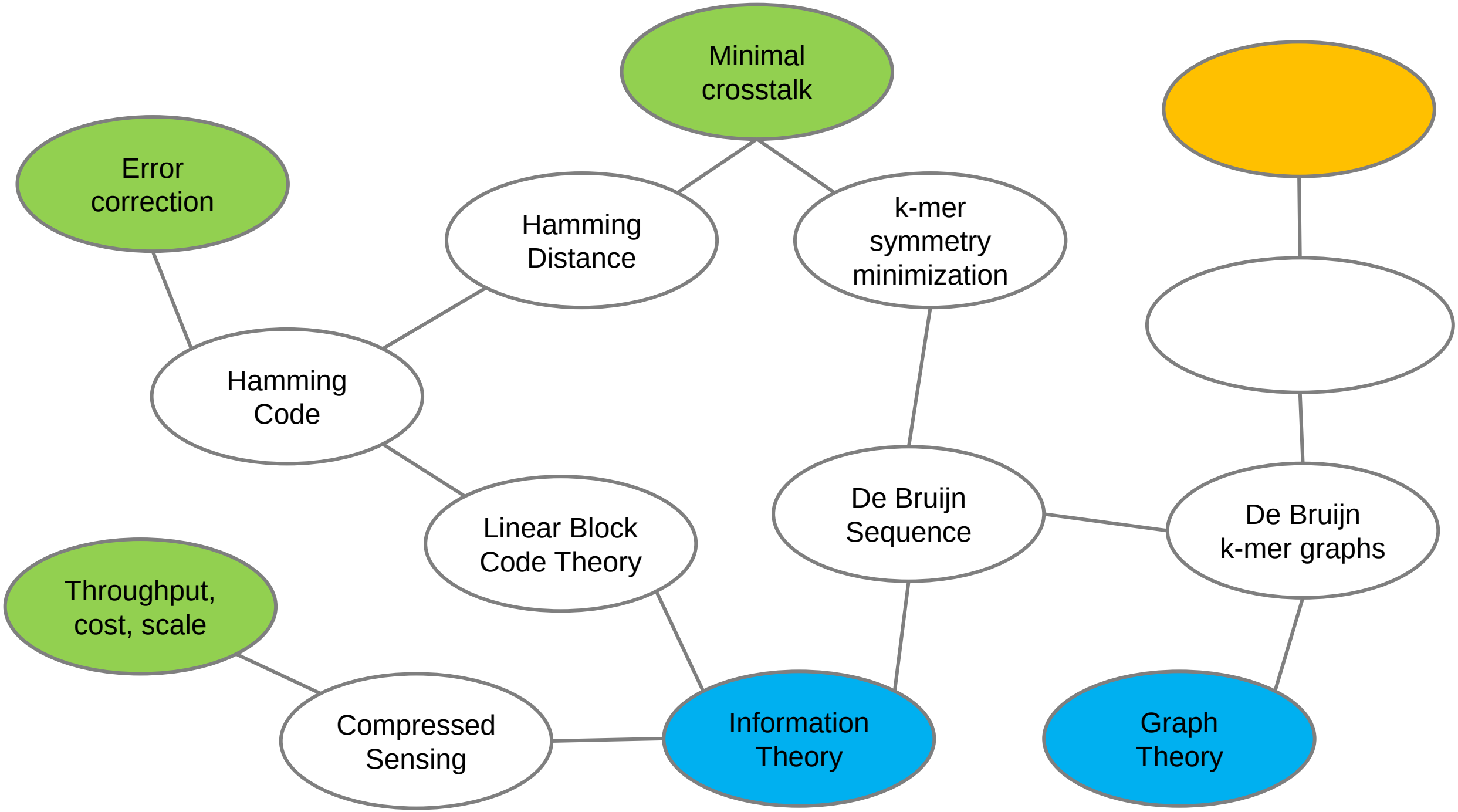


set() arithmetic
in python, e.g.

Exhaustive algos:

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$




Minimal crosstalk
Error correction
Hamming Distance
k-mer symmetry minimization
Hamming Code
Linear Block Code Theory
De Bruijn Sequence
De Bruijn k-mer graphs
Throughput, cost, scale
Compressed Sensing
Information Theory
Graph Theory

Outline

- Fluorescence in situ hybridization (FISH)
- Information theory & FISH
  - Compressed sensing
  - Linear block theory, error-correction
- de Bruijn sequences
- Orthogonal DNA sequence set design
  - k-mer symmetry minimization
  - Hamming distance approaches
- New algorithm idea…

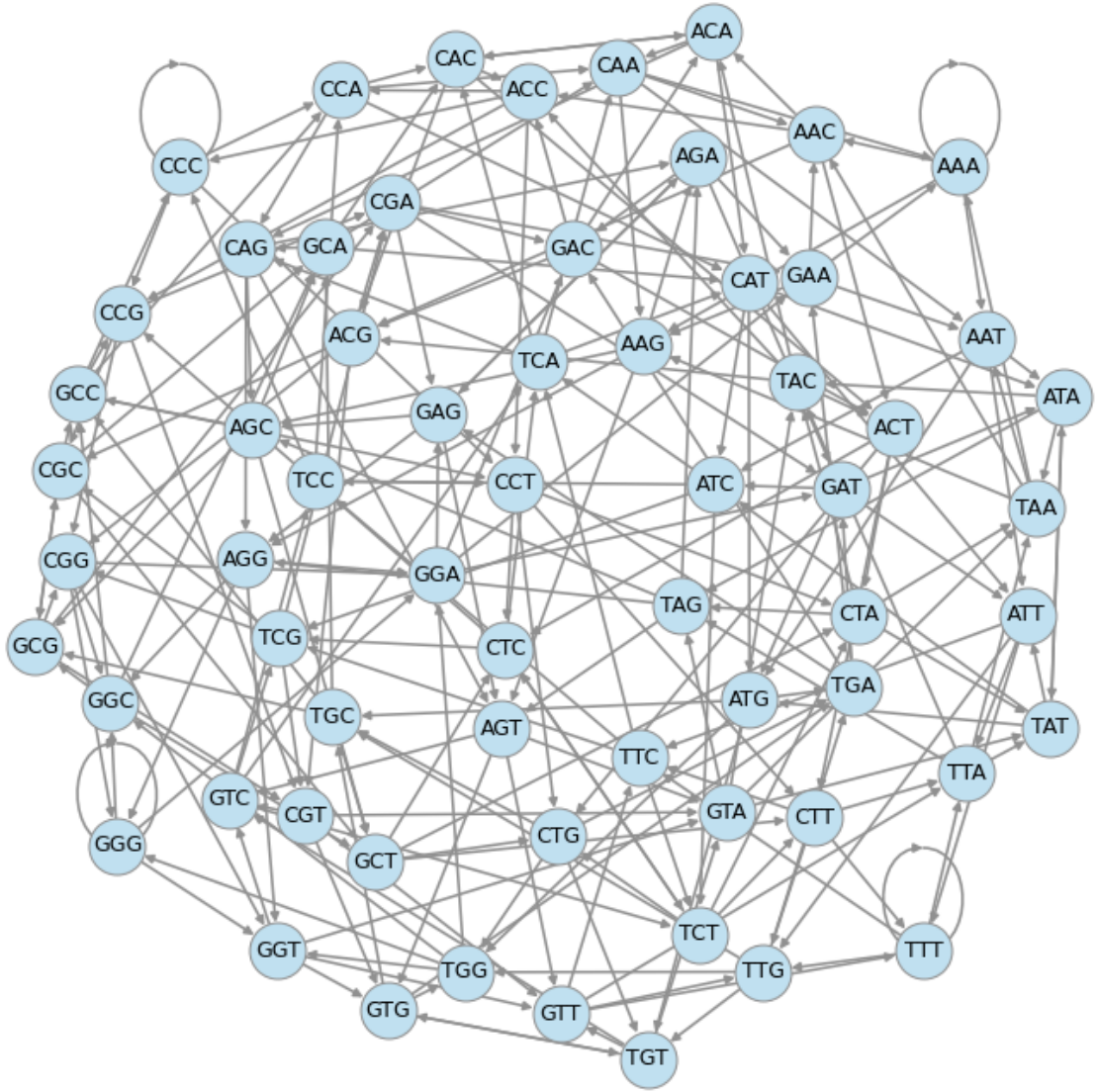
de Bruijn k-mer graph as K-order HMM

k = 3

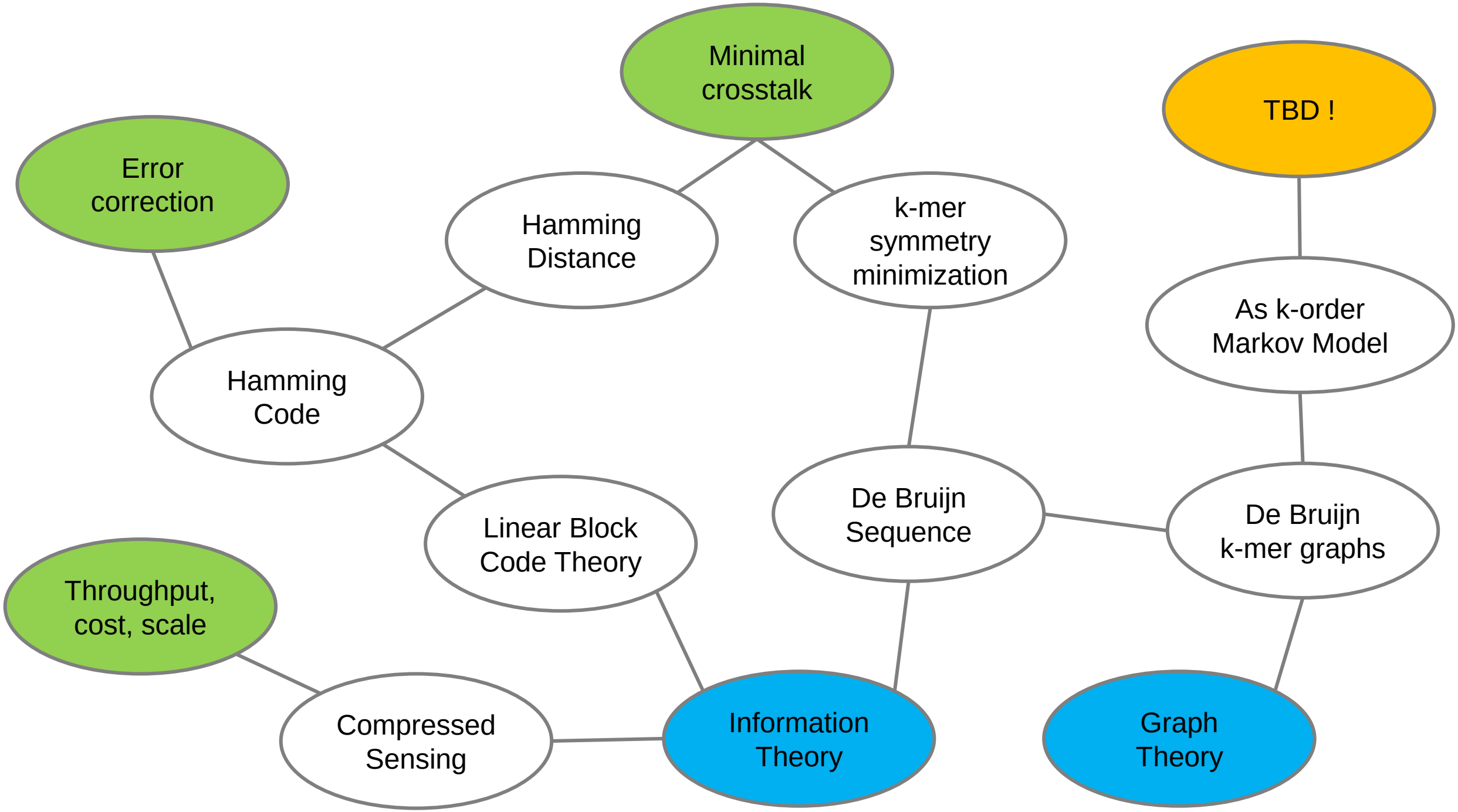
HMM examples: Markov models

- Ordinary Markov chain model:
  - states = observed symbols
  - emission probs = 1 or 0
  - transition probs = prob of observing a symbol, given the preceding one.
- Order k Markov model
  - states = length k words (e.g. $b_1 b_2 \ldots b_k$)
  - (unique) symbol emitted by $b_1 b_2 \ldots b_k$ is b_k
  - transition prob from $b_1 b_2 \ldots b_k$ to $c_1 c_2 \ldots c_k$ is non-zero only if
    - $c_1 c_2 \ldots c_{k-1} = b_2 b_3 \ldots b_k$, in which case it is $P(b_{k+1} | b_1 b_2 \ldots b_k)$ where $b_{k+1} = c_k$

22



http://bozeman.mbt.washington.edu/compbio/mbt599_2022/Lec13.pdf


Minimal crosstalk
TBD !
Error correction
Hamming Distance
k-mer symmetry minimization
As k-order Markov Model
Hamming Code
Linear Block Code Theory
De Bruijn Sequence
De Bruijn k-mer graphs
Throughput, cost, scale
Compressed Sensing
Information Theory
Graph Theory

. . .