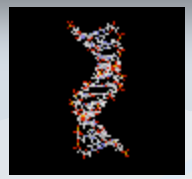


Engineering DNA Nanostructures and Computations with Bioinformatic Tools

Russell Deaton
Professor
Comp. Science & Engineering
The University of Arkansas
Fayetteville, AR 72701
rdeaton@uark.edu

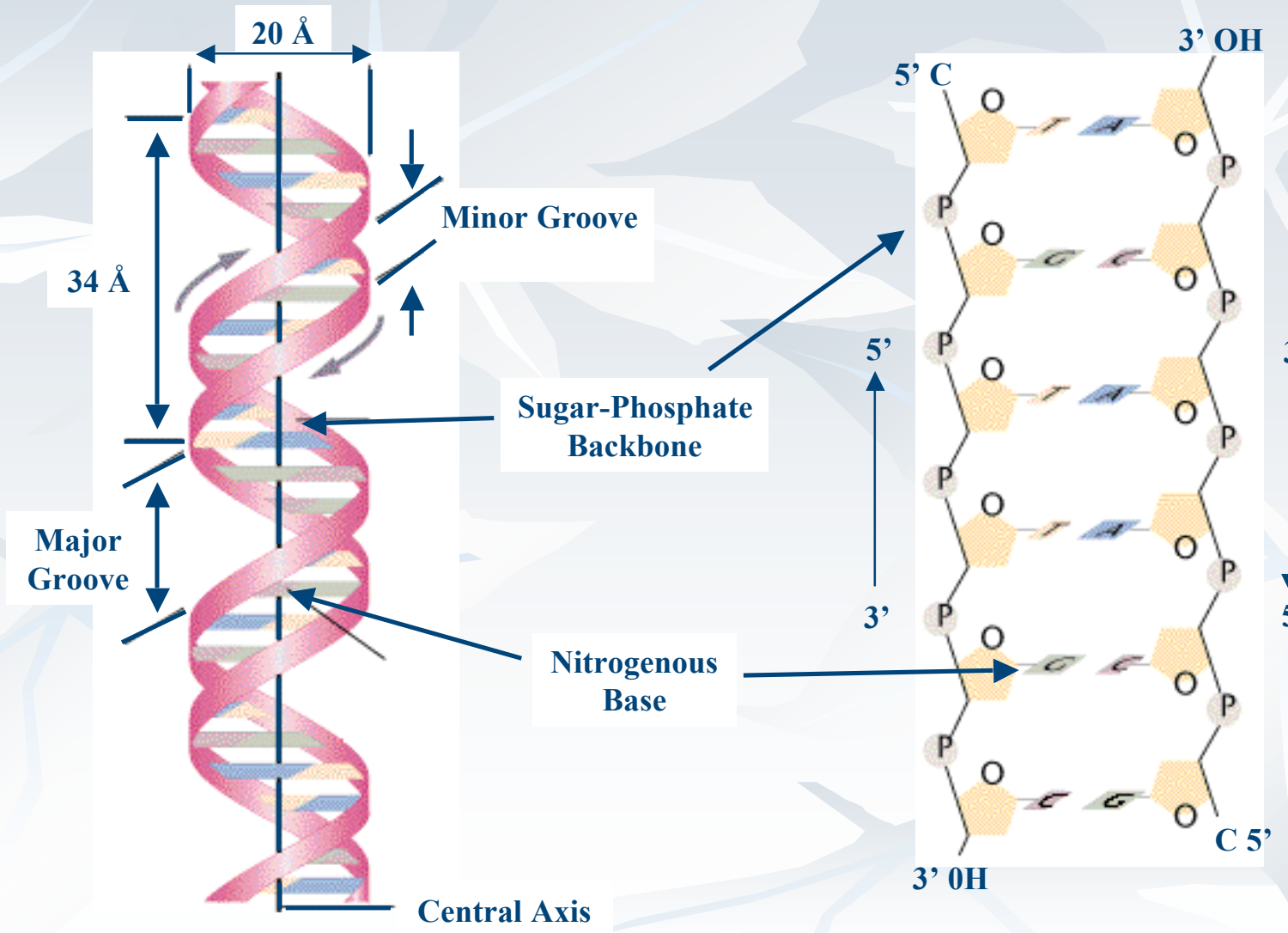
DNA for Non-Biological Purposes

- Density:
  - DNA: 1 bit per nm^3, 10^{20} molecules, exabytes
  - Video: 1 bit per 10^{12} nm^3
- Efficiency (Adleman)
  - DNA: 10^{19} ops / J
  - Supercomputer: 10^9 ops / J
- Speed (Adleman):
  - DNA: 10^{14} ops per s
  - Supercomputer: 10^{12} ops per s

What makes this possible?

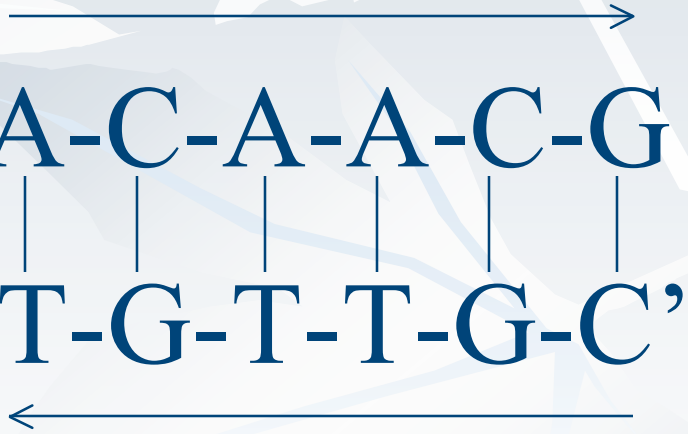
- Great advances in molecular biology
  - PCR (Polymerase Chain Reaction)
  - DNA Microarrays
  - New enzymes and proteins
  - Better understanding of biological molecules
- Ability to produce massive numbers of DNA molecules with specified sequence and size
- DNA molecules interact through template matching reactions

PHYSICAL STRUCTURE OF DNA



Template Matching Hybridization Reaction

` A-C-A-A-C-G



T-G-T-T-G-C’

Hybridization Allows:

- Massively Parallel Search based on Watson-Crick Complements
- Directed Self-Assembly of Nanostructures
- Search Stored Information for Similar Sequence Content

Mismatches

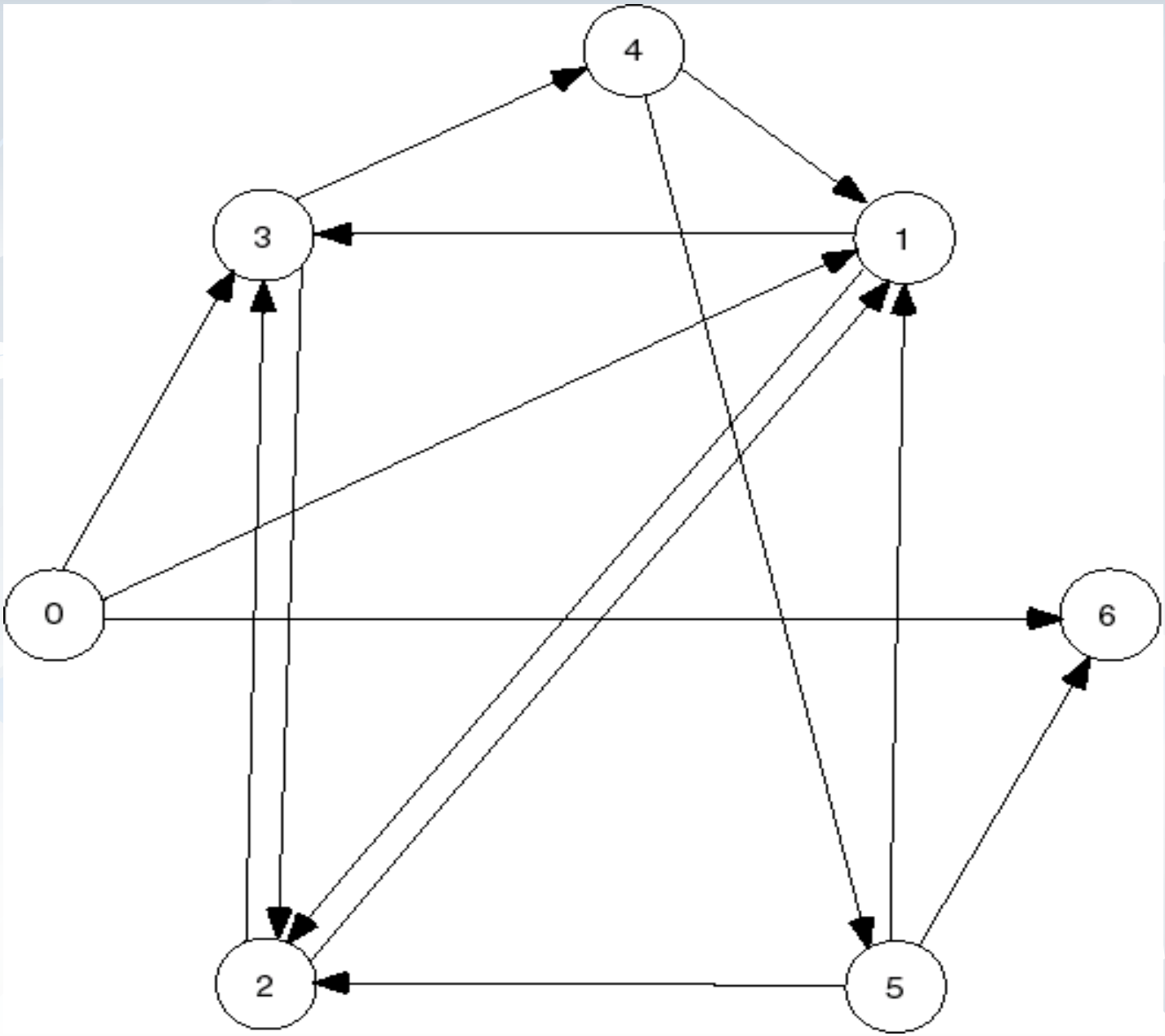
Mismatched Hybridization

Hairpin Hybridization

Shifted Hybridization

What is an example?

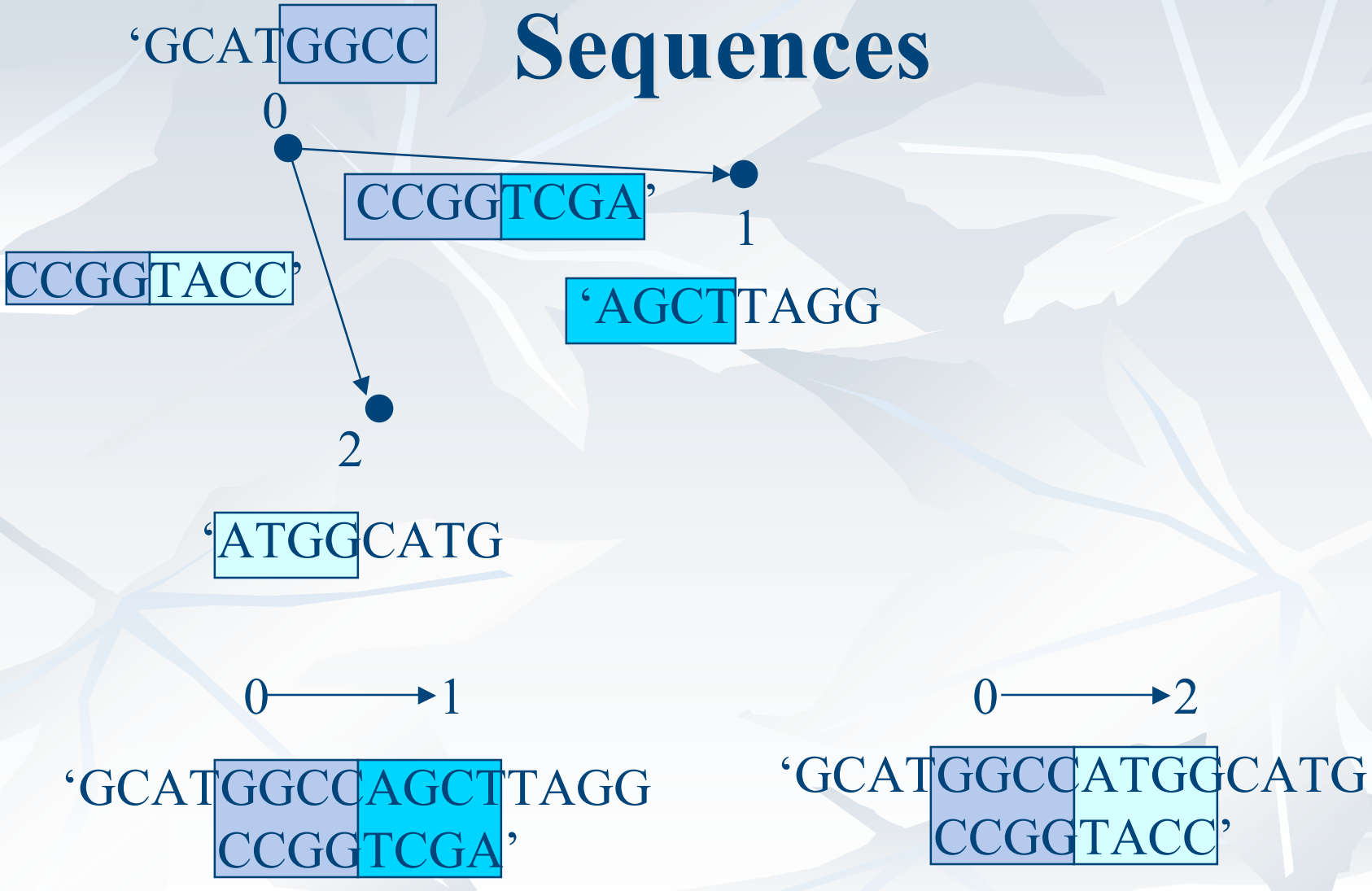
- "Molecular Computation of Solutions to Combinatorial Problems"
- Adleman, *Science*, v. 266, p. 1021.

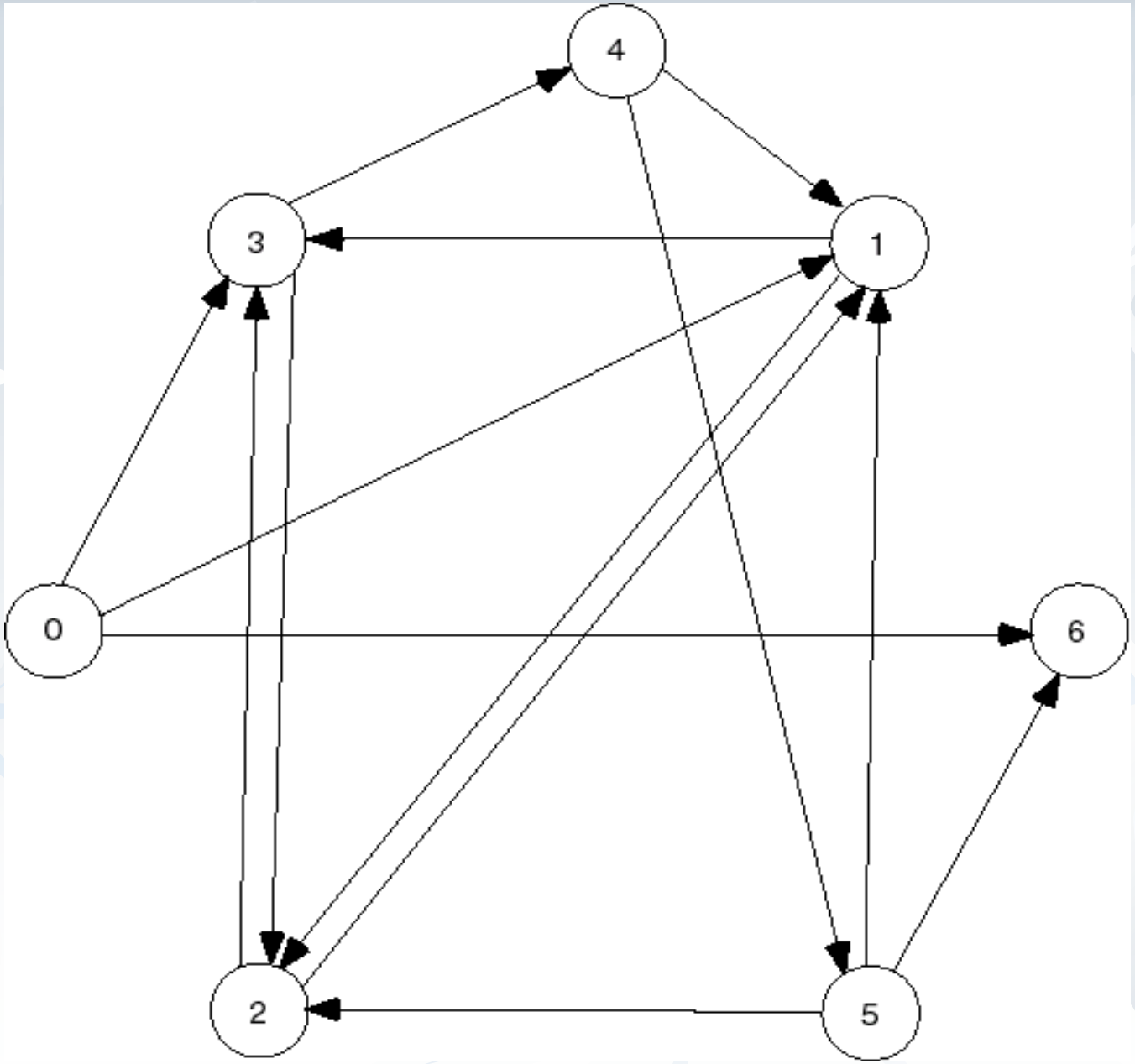

4
3
1
0
6
2
5

Algorithm

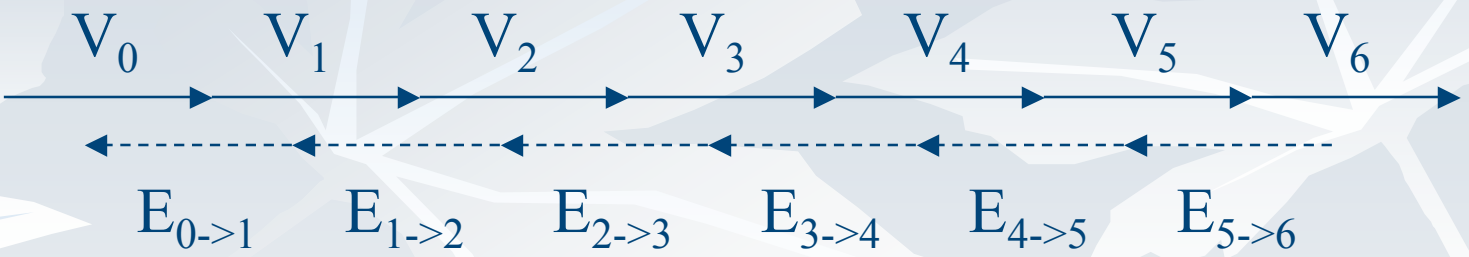
- Generate Random Paths through the graph.
- Keep only those paths that begin with v_{in} and end with v_{out}.
- If graph has n vertices, then keep only those paths that enter exactly n vertices.
- Keep only those paths that enter all the vertices at least once.
- In any paths remain, say “Yes”; otherwise, say “No”

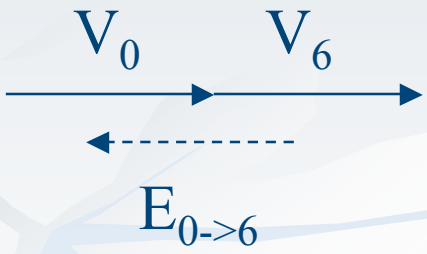
Representing a Graph with Sequences

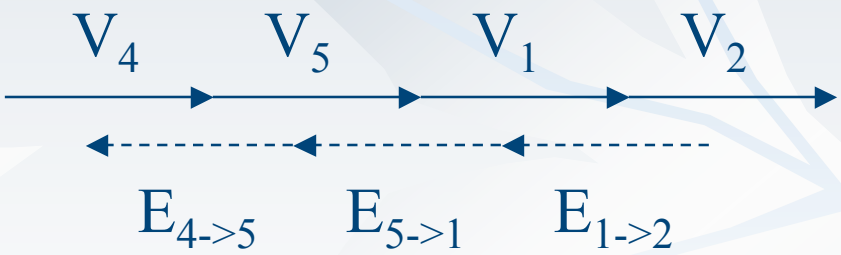


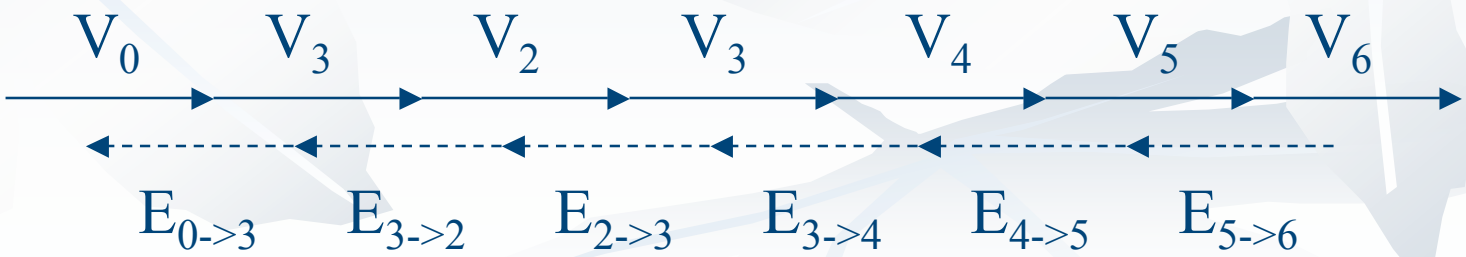

4
3
1
0
6
2
5

Massively Parallel Search

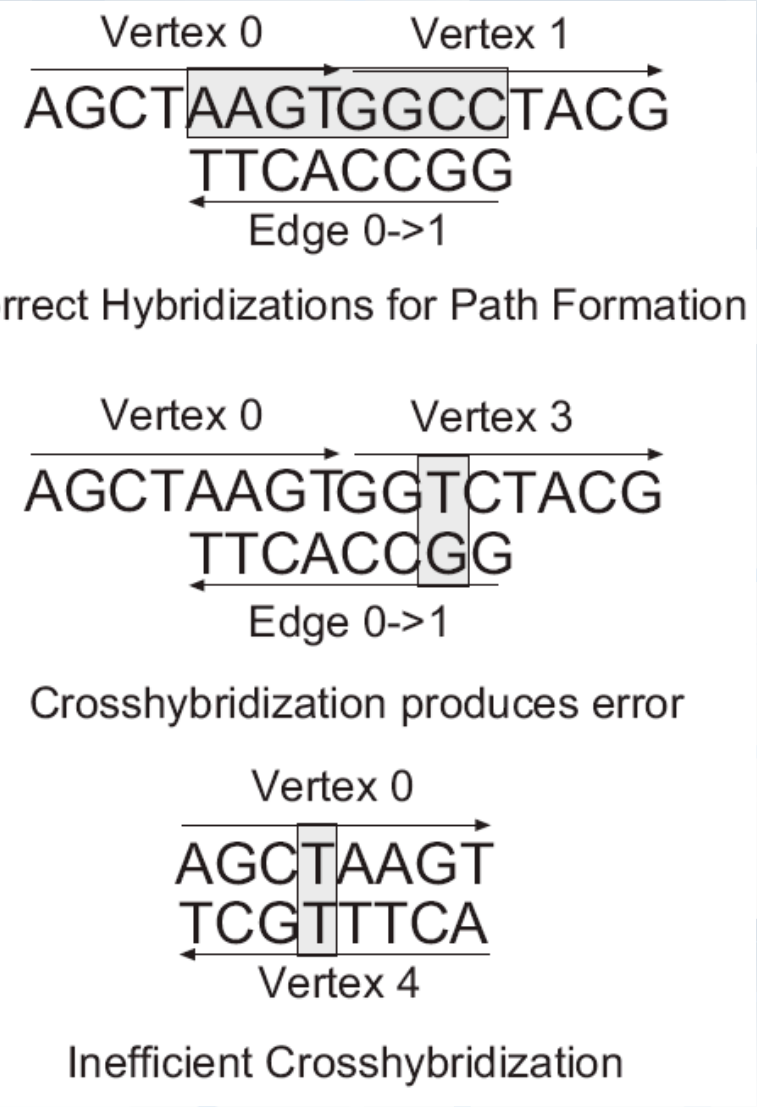








Mismatches

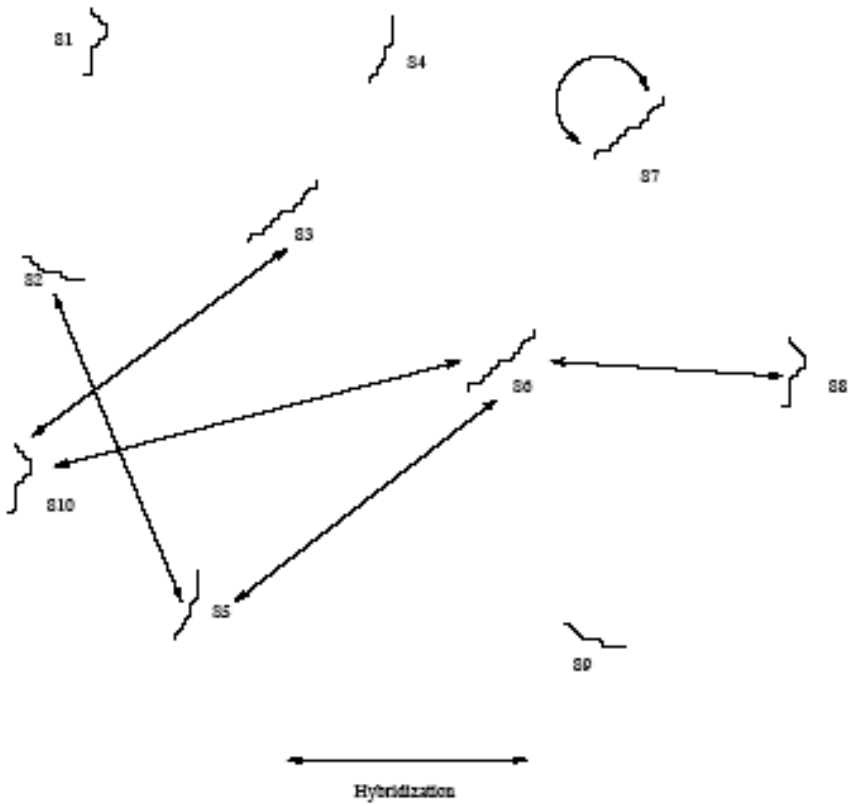

Vertex 0
Vertex 1
AGCTAAGTGGCCTACG
TTCACCGG
Edge 0->1
Correct Hybridizations for Path Formation
Vertex 0
Vertex 3
AGCTAAGTGGTCTACG
TTCACCGG
Edge 0->1
Crosshybridization produces error
Vertex 0
AGCTAAGT
TCGTTTCA
Vertex 4
Inefficient Crosshybridization

DNA Word Design Constraints

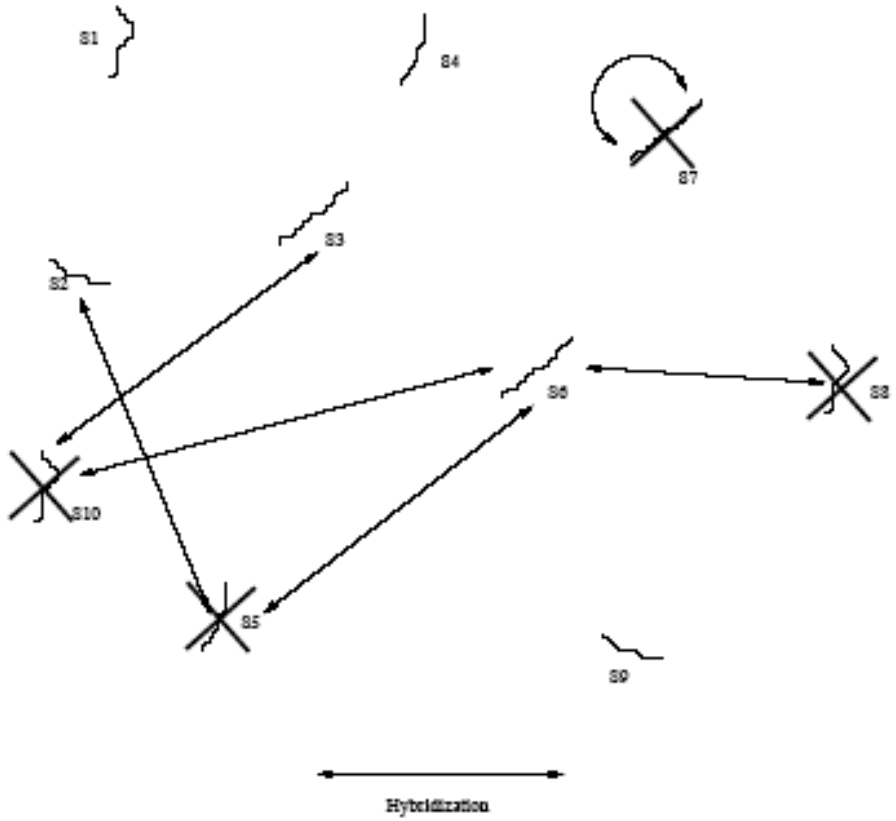
- Sequence design should implement the architecture.
  - Planned Hybridizations
  - Problem Size
  - Subsequent Processing Reactions
- Designed sequences should minimize unplanned “cross-hybridizations.”
- Consequences of Bad Designs: Errors and Poor Efficiency

DNA Word Design

- Design problem is hard (NP-Complete).
- As number of sequences required to represent the problem increases, this constraints increasingly conflicts with the requirement of non-crosshybridization.
- How much of DNA sequence space is available for computation?



DNA word design problem can be represented as a graph. Oligonucleotides (or Watson-Crick pairs) are the vertices, and all-or-nothing hybridizations are edges.



To find a non-crosshybridizing set of oligos, a greedy approach to eliminating one of a hybridized pair can be taken (See Suyama *et al.*, DNA 5).

Implementation Details

1. Large random sets of oligonucleotides and their Watson-Crick complements are generated.

2. Oligonucleotide are chosen in order,and added to the library if they are still available.

3. All oligonucleotides that have a minimum free energy of hybridization with the added sequence, or its complement, that are less than some threshold are eliminated from further consideration.

4. By repeating this process, a non-crosshybridizing library can be selected from the original random population.

Bioinformatics Tools

- Sequence Comparison Important
- Smith-Waterman Dynamic Programming algorithm to computer minimum free energy of hybridization
- Nearest-neighbor model of DNA duplex thermal stability

Align $v = AAAC$ and $w = AGC$. Match = +1, Mismatch = -1, Gap = -2.

		0	A	G	C
0	0	0	-2	-4	-6
A	1	-2	1	-1	-3
A	2	-4	-1	0	-2
A	3	-6	-3	-2	-1
C	4	-8	-5	-4	-1

		0	A	G	C
0	0		←	←	←
A	1	↑	↖	←	←
A	2	↑	↑, ↖	↖	↖, ←
A	3	↑	↑, ↖	↖	↖
C	4	↑	↑	↑, ↖	↖

Experimental Validation

TABLE I: Non-crosshybridizing library of 40 Watson-Crick pairs. Only one sequence of the pair is shown. The simulation conditions were 23°C, 1 M NaCl, and 1 μM DNA concentration.

#	Sequence	#	Sequence
1	aacaatcttttcaagctaac	2	tgtttctatctaggcgtgat
3	gttagagagtaaatgttagg	4	taccgtagtaaactgtctac
5	tgtctcaacgattacccccg	6	tacatgacghaagccaaggg
7	tcaaatgaagctattttgta	8	ctttggatttatcttcgaca
9	atgaccatattaggtagtag	10	gatcctatatcttaatgcac
11	gttccgaaataacagaatcg	12	aacgaaccttctagagtatg
13	cagcgtcttcccttaagtac	14	gcacaattaggcactaaccc
15	ctttcccagtagaattacaa	16	ggaccctgtataacatacaa
17	gttggaatcacctctatgat	18	cataaaaagttaataagtta
19	gtttttgatattttagtcg	20	atcagttgttgttaaattac
21	agactttatggataccattc	22	attttaagactatctcttag
23	ccttttttcgtatcgctcc	24	catactttgtaagtaattat
25	acatttttctacatccacat	26	agtaacttcaaccataggcc
27	tttatcattattacactatc	28	gtattaatttccatctaaaa
29	actagaccaagaaatttaga	20	ggtctctgtacttttctgact
31	tttcctaatactgcttatat	32	aggtttaattagtcaaatag
33	tatgctaggtaaaaaataag	34	cttctctatataatatttca
35	cctaaagaactcttattatt	36	agacataatttttatatactc
37	aggagaatcttacttctacg	38	tcttatagatcccgtactga
39	aatgtatgagtttattctaa	40	tcattcatatacaagttatc

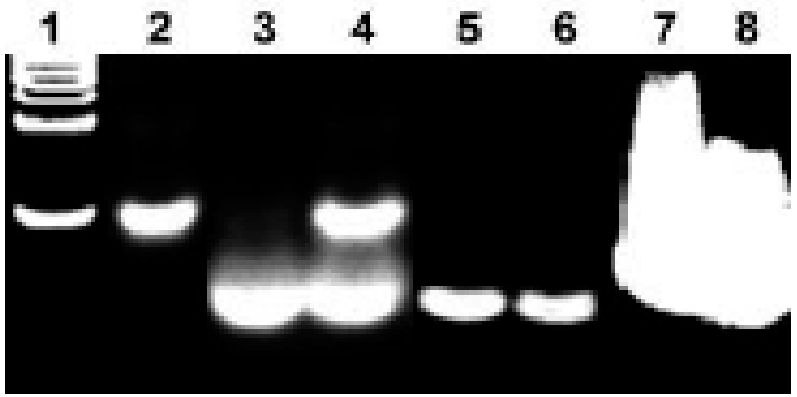


FIG. 1: Experimental results after annealing of oligonucleotides. Lane 1 is a molecular size marker for duplexes of 20 to 120 bp in 20 bp increments. Lane 2 contains sequence 5 and its Watson-Crick complement. Lane 3 contains sequences 1-40 after annealing. Lane 4 is sequences 1-40 plus the complement of sequence 5. Lanes 5 and 6 contain sequences 13 and 23, respectively. Lanes 7 and 8 contain oligonucleotides *ggggggggaaaaccccccccc* and *atgcatgcaaaagcatgcat*, respectively, with known secondary structure. There are no detectable duplexes in lane 3, supporting the non-crosshybridizing properties of the designed sequences. By comparing smears in lanes 7 and 8 to well-defined bands in lanes 5 and 6, it is confirmed that designed sequences have little to no secondary structure.

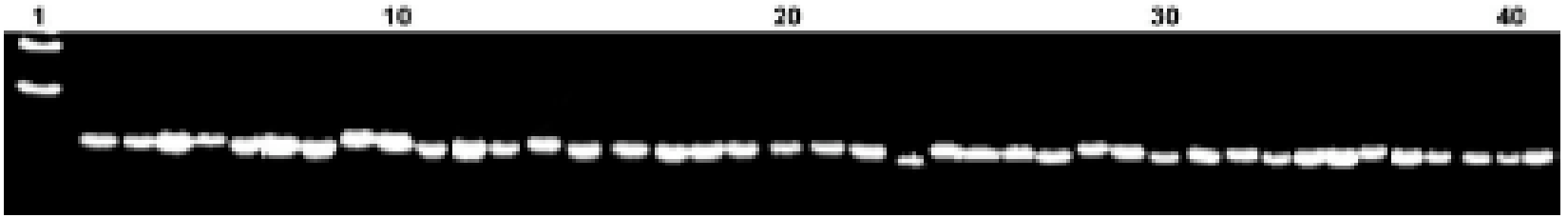


FIG. 2: Lane 1 is a molecular size marker for duplexes of 20 and 40 bp. Lanes 2-41 contain sequences 1-40 by themselves. These results indicate no secondary structure.

1. Synthesize initial population.

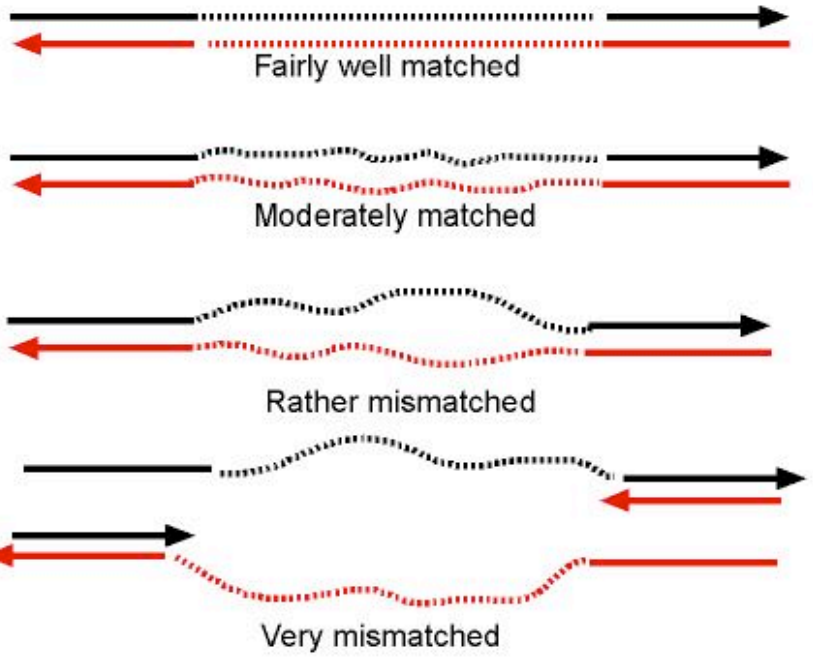
2. PCR Amplify with mutation using P1, P2 Primers.

3. More PCR, but with lower maximum temperature. Monitoring shows no amplification (double strands do not separate).

Repeat, using slightly higher temperature.

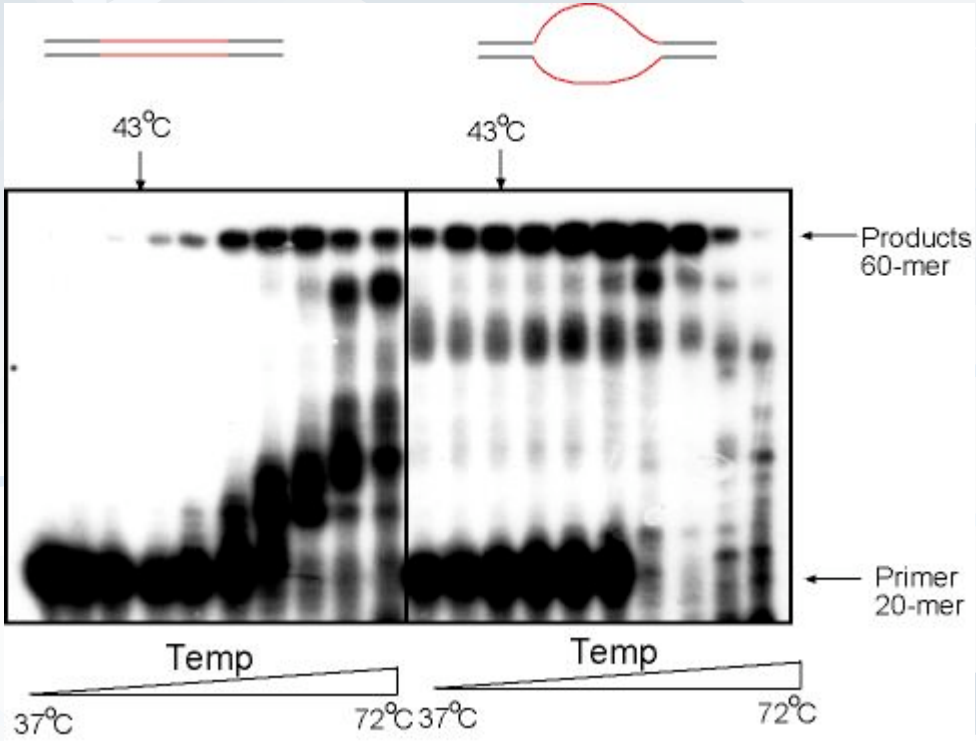
Repeat again, until amplification is detected.

4. What happened? We had double strands with various degree of mismatches.

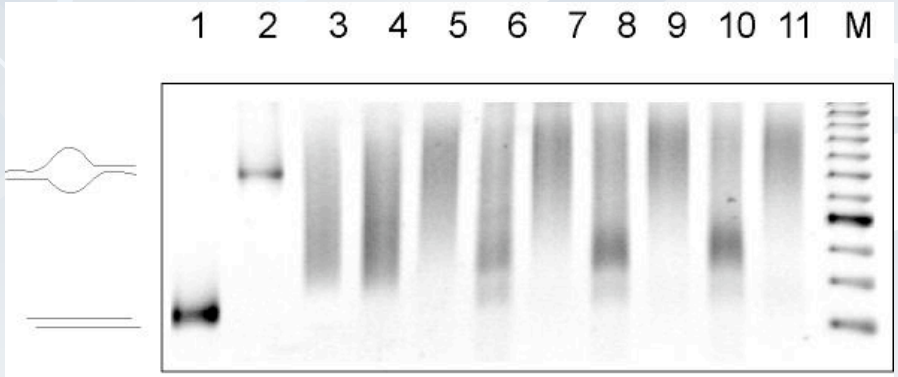


Only the very mismatched will melt apart, allowing primers to enter and extend. Only the very mismatched are amplified.

Experimental Results



Protocol Iterated



1: Perfect Matched Duplex
2: Total Unmatched Duplex
3: Original Pool
4: Extension
5: Purify 60-mer, Reannealing
6: Extension
7: Purify 60-mer, Reannealing
8: Extension
9: Purify 60-mer, Reannealing
10: Extension
11: Purify 60-mer, Reannealing

Extension: 42 oC for 25 mins.
Annealing: 95 oC 5 mins, room

Team

- Russell Deaton, University of Arkansas, Computer Science and Engineering
- Junghuei Chen, University of Delaware, Chemistry and Biochemistry
- Jin-Woo Kim, University of Arkansas, Biological Engineering
- Hong Bi, University of Delaware, Chemistry and Biochemistry
- Max Garzon, University of Memphis, Computer Science
- Harvey Rubin, University of Pennsyvania, School of Medicine
- David Wood, University of Delaware, Computer and Information Science

Acknowledgement

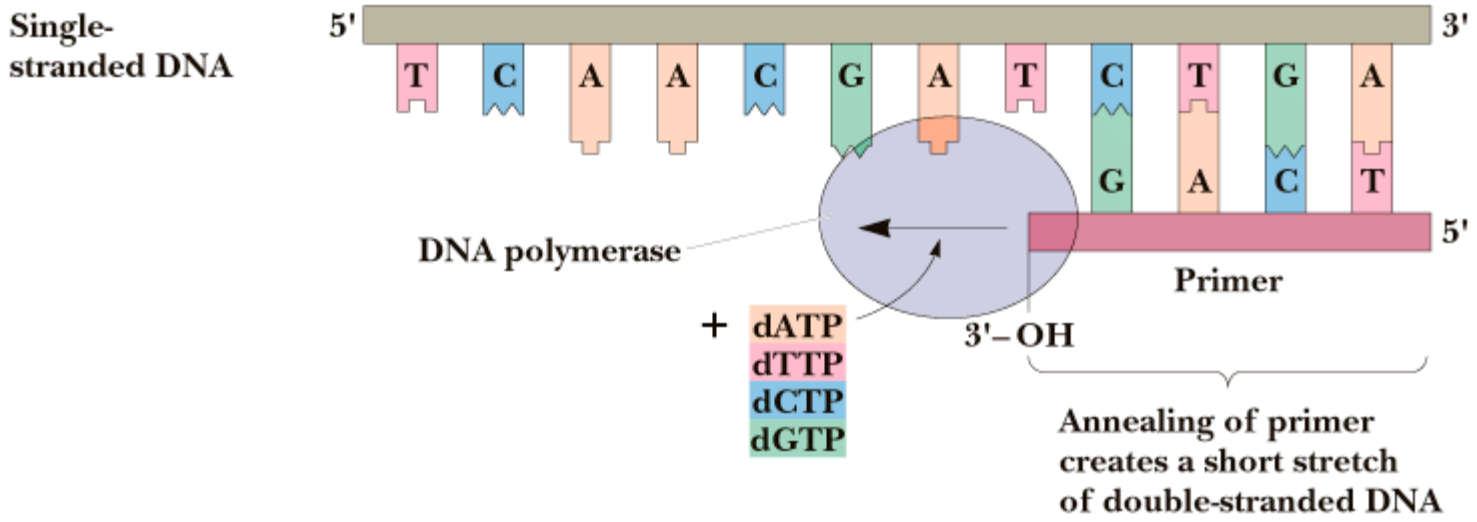
- This work was supported by the NSF QuBIC program, award number EIA-0130385

Algorithm

- Generate Random Paths through the graph.
- Keep only those paths that begin with v_{in} and end with v_{out}.
- If graph has n vertices, then keep only those paths that enter exactly n vertices.
- Keep only those paths that enter all the vertices at least once.
- In any paths remain, say “Yes”; otherwise, say “No”

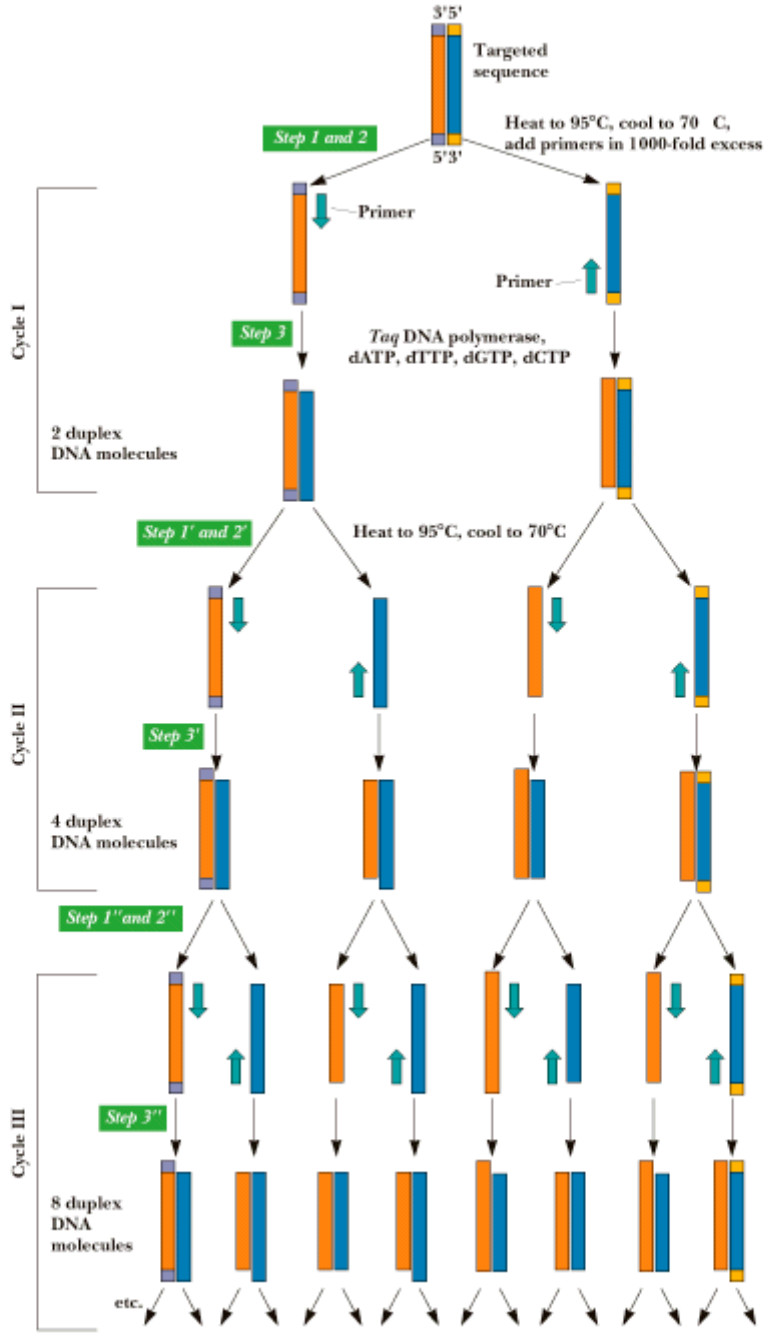
DNA Polymerase

Garrett & Grisham: Biochemistry, 2/e
Figure 12.2

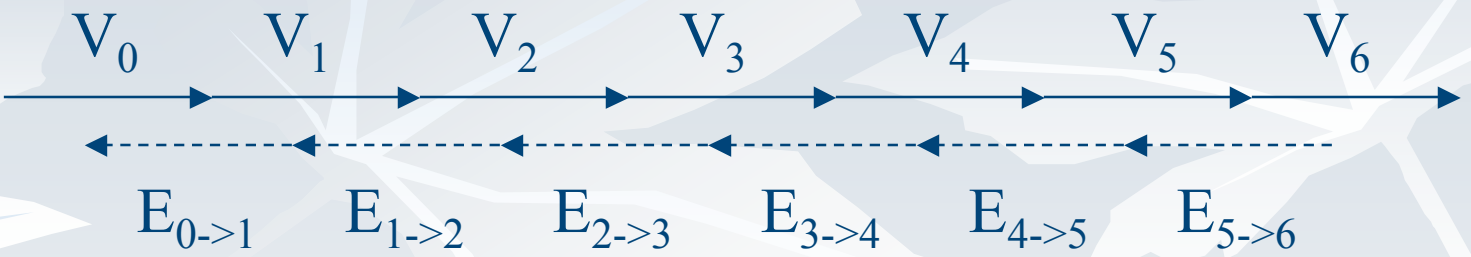


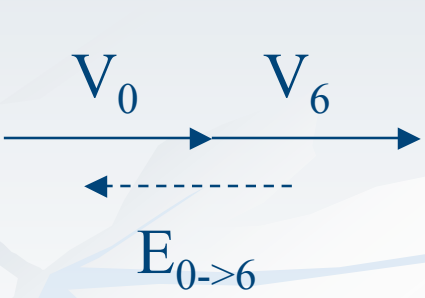
Saunders College Publishing

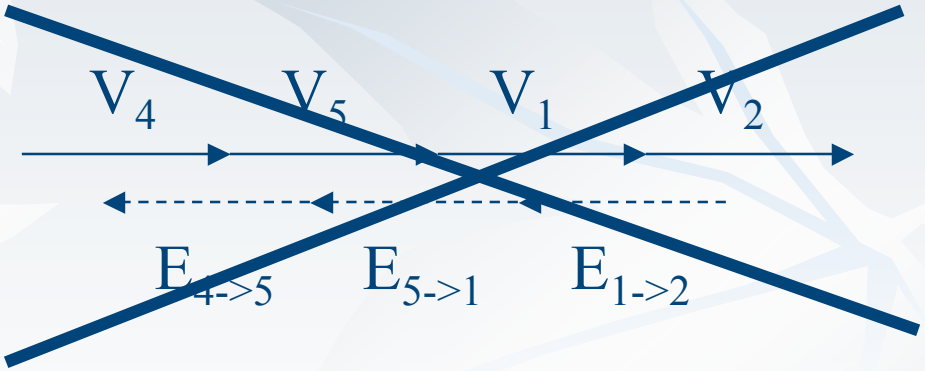
POLYMERASE CHAIN REACTION

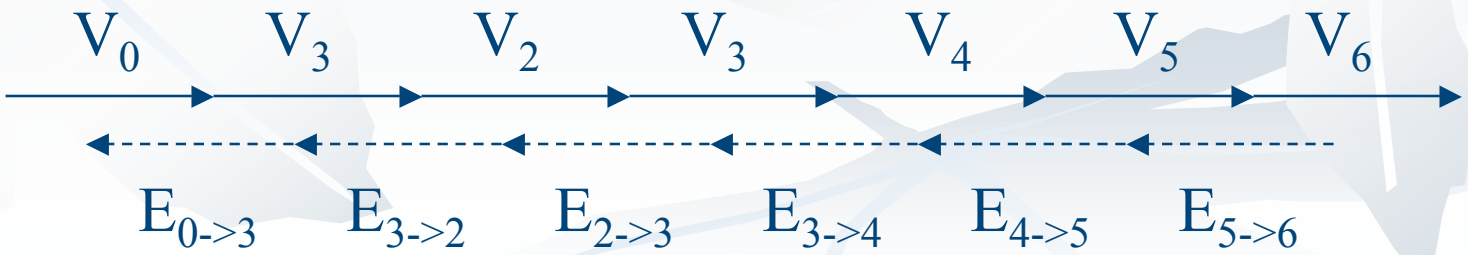


Start = V0, Stop = V6





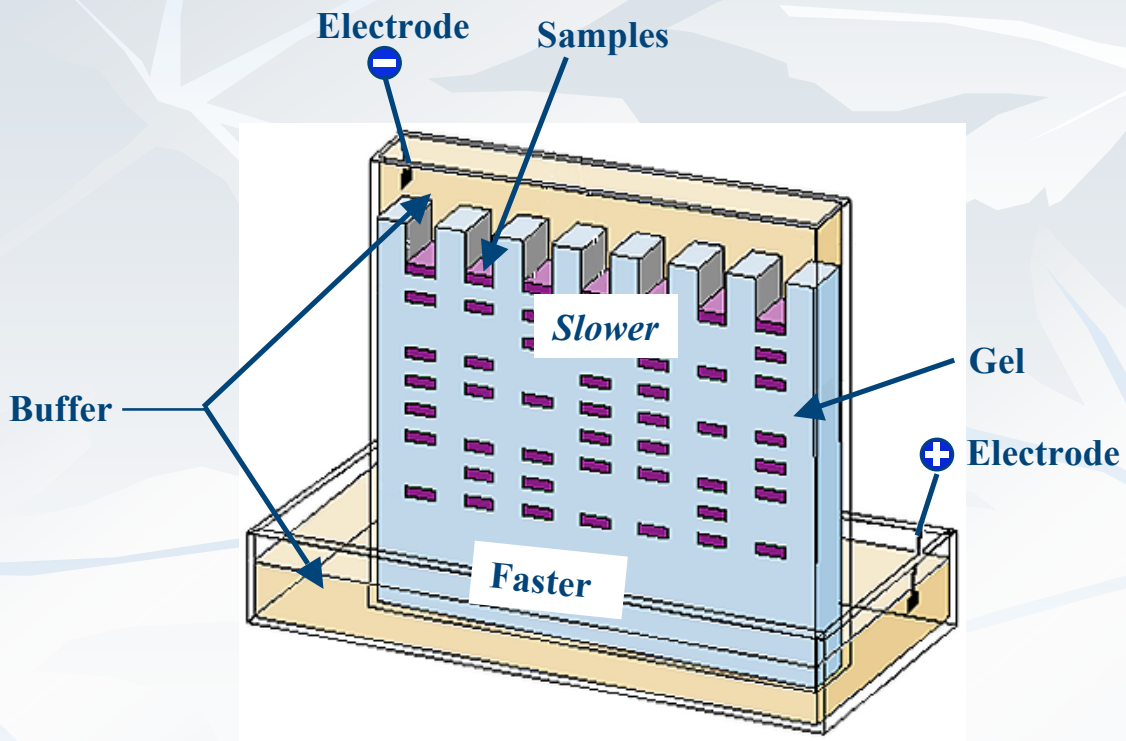




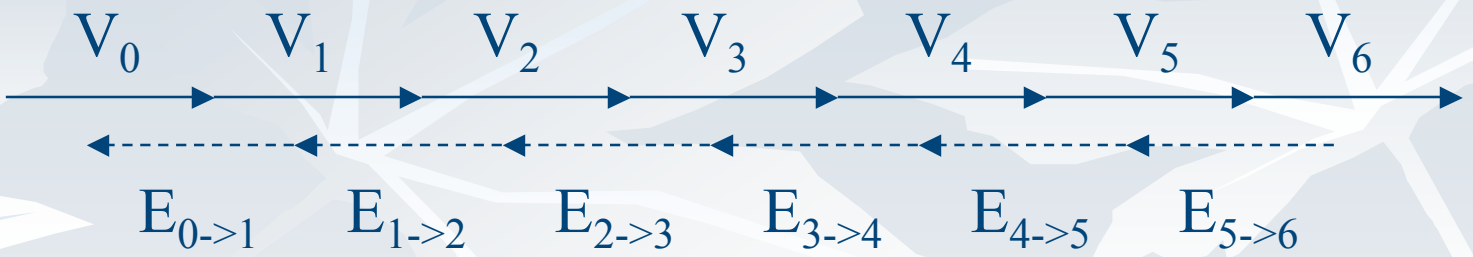
Algorithm

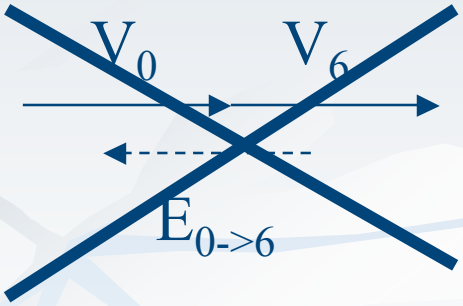
- Generate Random Paths through the graph.
- Keep only those paths that begin with v_{in} and end with v_{out}.
- If graph has n vertices, then keep only those paths that enter exactly n vertices.
- Keep only those paths that enter all the vertices at least once.
- In any paths remain, say “Yes”; otherwise, say “No”

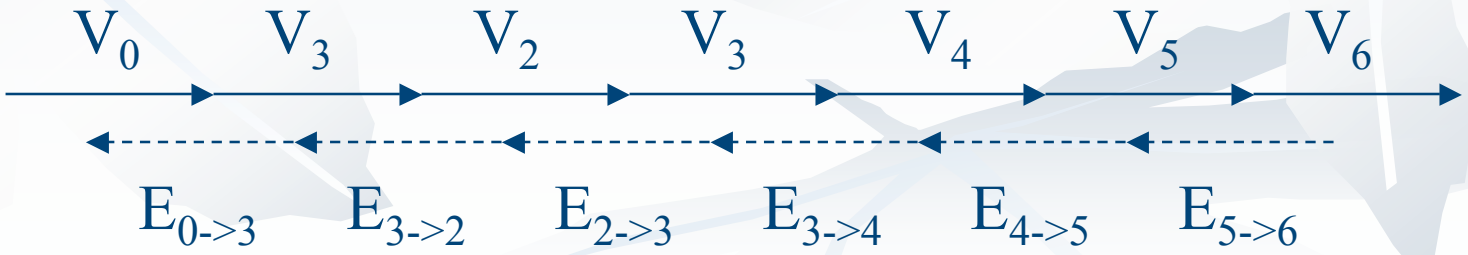
GEL ELECTROPHORESIS - SIZE SORTING



Right Length



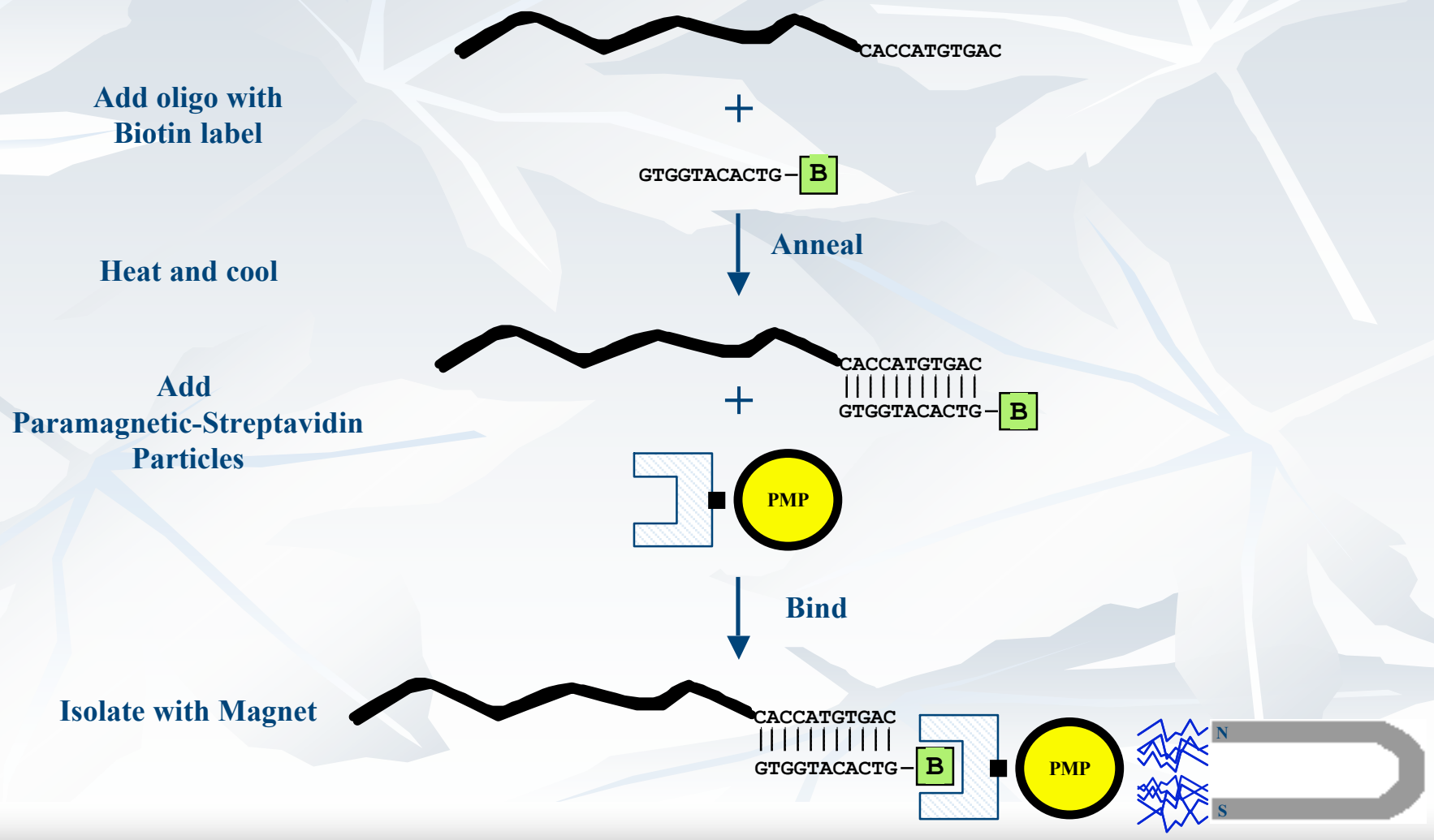




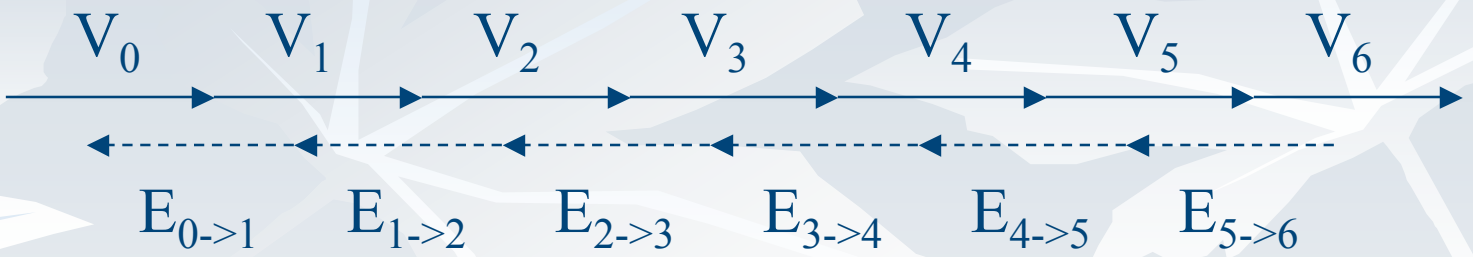
Algorithm

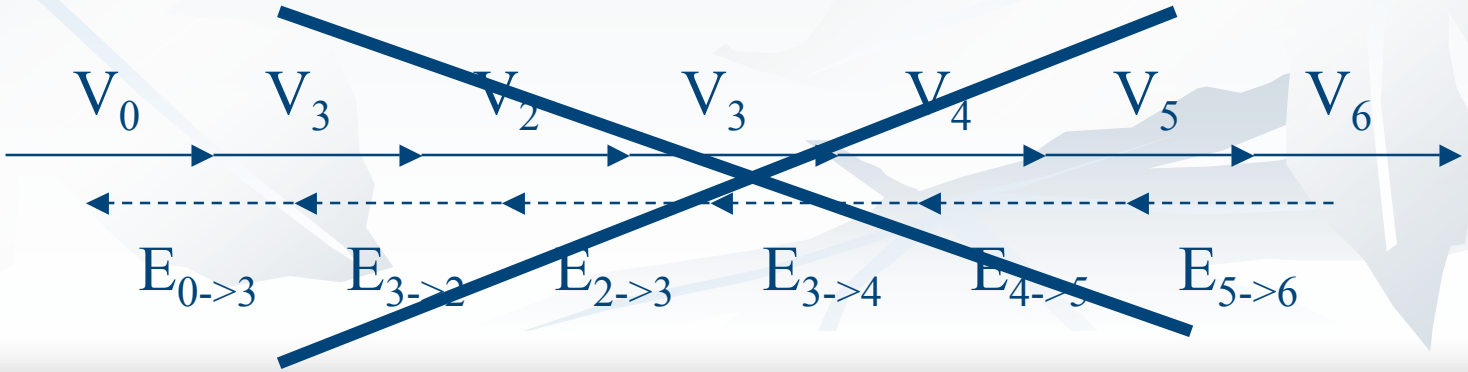
- Generate Random Paths through the graph.
- Keep only those paths that begin with v_{in} and end with v_{out}.
- If graph has n vertices, then keep only those paths that enter exactly n vertices.
- Keep only those paths that enter all the vertices at least once.
- In any paths remain, say "Yes"; otherwise, say "No"

Antibody Affinity



Every Vertex





Algorithm

- Generate Random Paths through the graph.
- Keep only those paths that begin with v_{in} and end with v_{out}.
- If graph has n vertices, then keep only those paths that enter exactly n vertices.
- Keep only those paths that enter all the vertices at least once.
- In any paths remain, say "Yes"; otherwise, say "No"

Hamiltonian Path