Lecture: Sequence Alignment

- Topics: genomics basics, sequence alignment, seed selection, Shifted Hamming Distance, Smith-Waterman algorithm

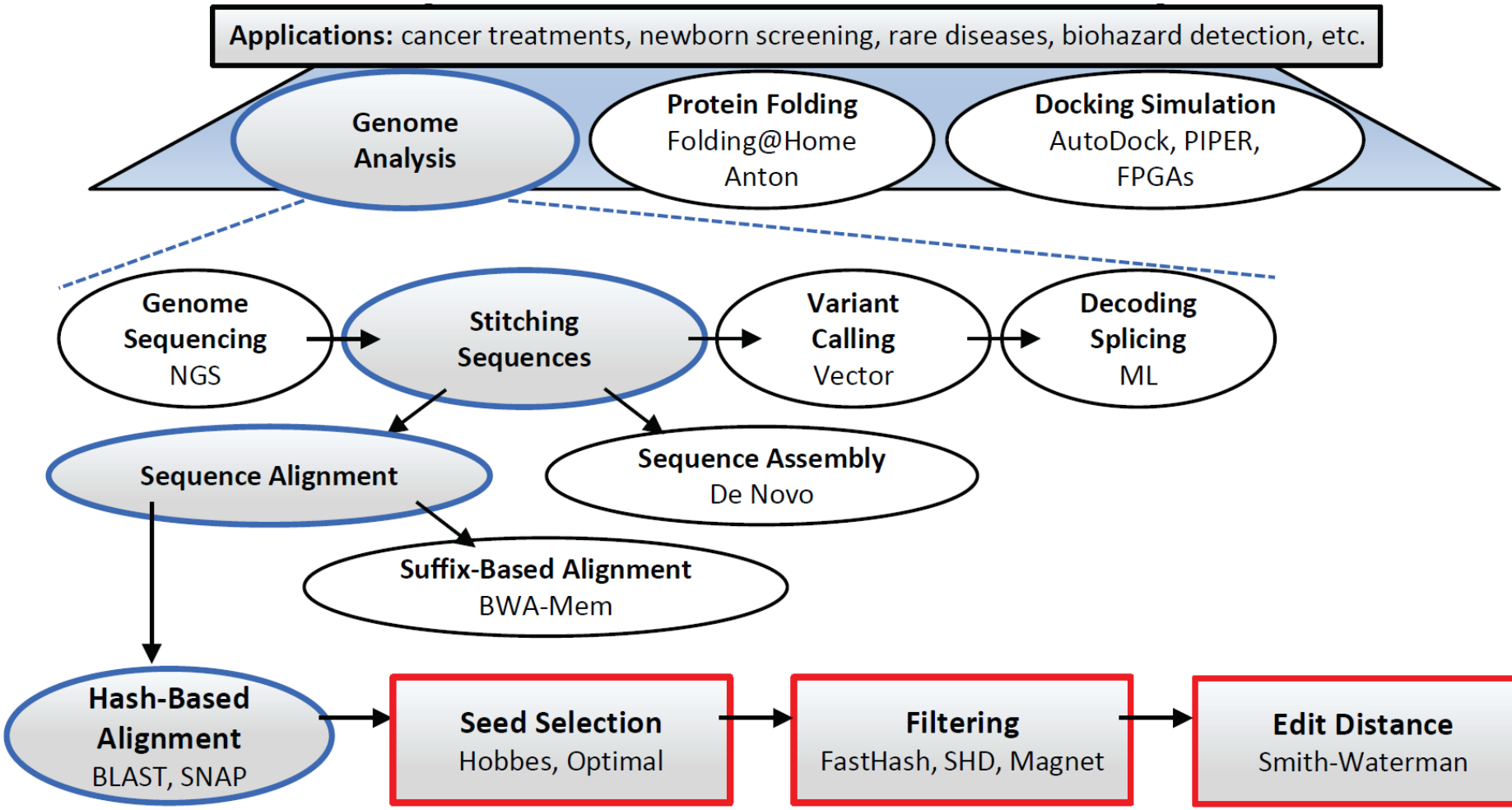
Precision Medicine

- Three broad classes of problem domains
- Information discovery from vast corpuses – NLP and ML
- Looking for genome matches and mismatches – sequence alignment
- Modeling protein-protein and protein-drug interactions – molecular dynamics

Genomics 101

- Genome – Chromosomes – Genes
- Chromosomes come in pairs; one is inherited from dad, the other from mom
- Genes are made up of exons (protein-making instructions), introns, and regulatory sequences that determine which genes are turned on/off
- A strand of DNA is a sequence of base pairs A, T, C, G
- A group of 3 bases (a triplet or codon) specifies one of 20 amino acids that must be produced; a protein is a sequence of amino acids

Precision Medicine

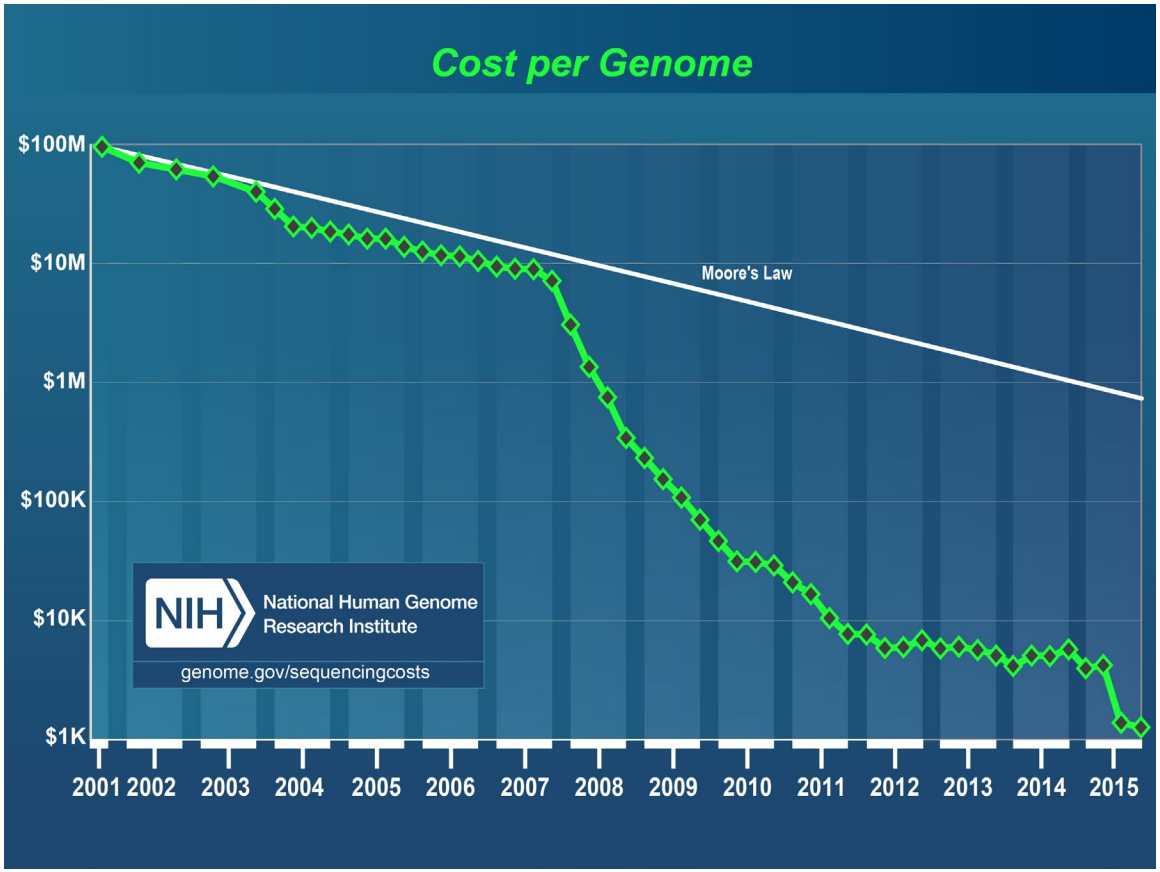


Sequencing

- High-throughput sequencing (HTS) devices fragment a DNA into short segments (at random locations) and produce their values (short reads) at high rates
- Matching these short reads (a few hundred base-pairs) to a reference genome is the bottleneck – sequence alignment, to be discussed next
- 2nd generation sequencing has read lengths of about 100; 3rd generation has read lengths of thousands; 3rd gen is also more error-prone and more compute-intensive, but it can deal with larger mutations

Sequencing Cost

- Sequencing cost has shrunk dramatically
- More compute demands as it is deployed more widely



Source: National Human Genome Research Institute

Seed Selection

- Seed selection with pigeon-hole principle
- Naïve, Hobbes, Optimal seed selectors
- Selected seeds look up a hash table to identify possible matching locations

Filtering with SHD and SRS

- Takes 2 strings a and b, and computes a distance score for those 2 strings; only similar locations then perform SWA

- A Hamming vector is computed for a and b; this accounts for substitutions; to account for insertions and deletions, a and b are also shifted left and right to produce new Hamming vectors; the vectors are bit-wise AND-ed to produce the final vector; the number of 1s in this vector gives us the Shifted Hamming Distance

SHD Example

Read:

TCCATTGACATTCGTGACTCTCCTTCTCTCCCACCCCTTTGCCCCC

Reference:

TCCATTGACATTCGTGAGCTGCTCCTTCTCTCCCACCCCTTTGCCC

Hamming mask 0:

0000000000000000111101111000010101001101110000

Read>>2:

TCCATTGACATTCGTGACTCTCCTTCTCTCCCACCCCTTTGCCCCC

Reference:

TCCATTGACATTCGTGAGCTGCTCCTTCTCTCCCACCCCTTTGCCC

Hamming mask 4:

00111111110111111010111000000000000000000000000

Final bit-vector:

00000000000000000001000000000000000000000000000

SRS

- If we are only tolerating 5% errors, we should have large exactly matching regions, i.e., nearly 20 consecutive 0s in the Hamming vector

- Therefore, a few consecutive 0s (<3) can be viewed as noise; these 0s are converted into 1s; without this technique, it is highly likely for one of the shifted vectors to have a zero in every position

- Can prove that this technique will not miss a matching pair; filters out 90+% of all locations; efficient SIMD implementation on a general-purpose processor; can get an additional 17x with an FPGA implementation

Smith Waterman Algorithm

A precise, but slow algorithm; used for local alignment, i.e., where does a string a best fit in a larger string b

O(mn), where *m* and *n* are the sizes of a and b

Most sequence alignment algorithms will use SWA as a last verification step

$$H(i,j) = max \begin{cases} 0 \\ H(i-1,j-1) + s(a_i,b_j) & (match/mismatch) \\ max_k \; H(i-k,j) + W_k & (deletion) \\ max_l \; H(i,j-l) + W_l & (insertion) \end{cases}$$

s() is the similarity score, W is the gap score.

Example

$$H(i,j) = max \begin{cases} 0 \\ H(i-1,j-1) + s(a_i,b_j) & \textit{(match/mismatch)} \\ max_k \; H(i-k,j) + W_k & \textit{(deletion)} \\ max_l \; H(i,j-l) + W_l & \textit{(insertion)} \end{cases}$$

Sequence a : ACACACTA

Sequence b: AGCACACA

s = +2 (match) or -1 (mismatch); W = -1

		A	C	A	C	A	C	T	A
	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

a: A– CACACTA

b: AGCACAC–A

Analysis

- O(mn) time
- A systolic array would take m+n time
- Memory bandwidth requirement is small
- Some papers have implemented this on FPGAs to get 2-3 orders of magnitude speedups

The Sequence Alignment Pipeline

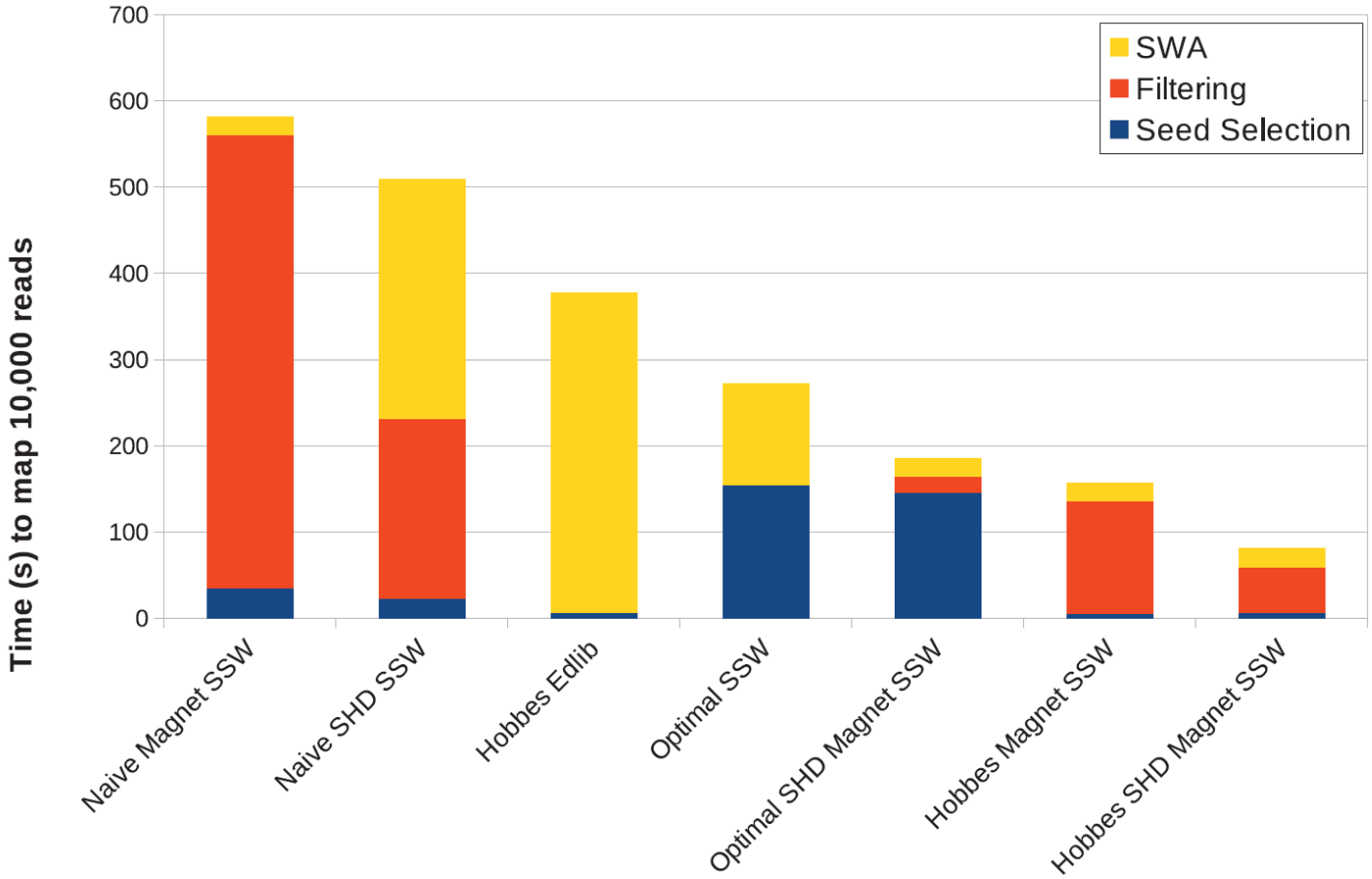


Figure 3: A design space exploration with various algorithms for the different stages of hash-based sequence alignment. Seed selection can use Naive/Hobbes/Optimal, Filtering can use nothing/SHD/Magnet, and SWA can use SSW/Edlib.

References

- Genomics basics: http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp1_1_1.shtml#genome1
- SWA overview: Wikipedia
- FPGA acceleration of SWA: “160-fold acceleration of …”, Li, Shum, Truong, BMC Bioinformatics, 2007
- Shifted Hamming Distance: Xin et al., Bioinformatics, 2015
- SHD on FPGA: Gatekeeper, Alser et al., Arxiv 2016
- BWT overview: Wikipedia
- BWA: “Fast and Accurate …”, Li and Durbin, Bioinformatics, 2009
- BWA on FPGA: “Hardware-Acceleration of …”, Waidyasooriya and Hariyama, IEEE TPDS, 2016
- Sequence alignment survey: Li and Homer, Briefings in Bioinformatics, 2010